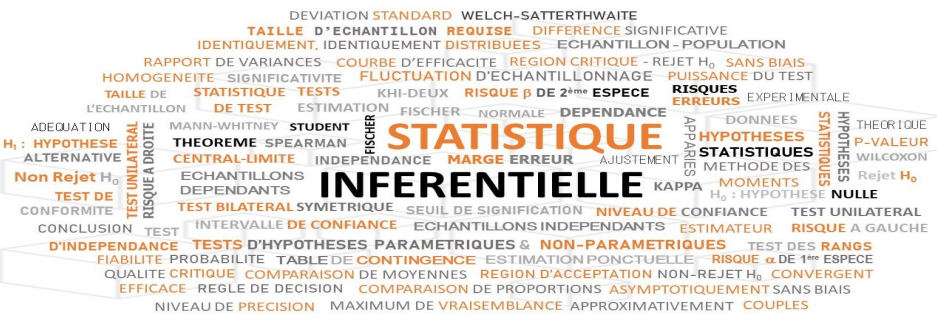
UFR de Sciences économiques et de Gestion

Université Lumière Lyon 2

UNIVERSITÉ DE LYON



UFR SEG L3-S6

Statistique Inférentielle

Support de cours (1/3)

Estimation ponctuelle & Intervalle de confiance

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/

Plan du cours

UFR de Sciences économiques et de Gestion

Université de Lyon

L3-S6 : Sciences Economiques & Gestion

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 de paramètres et les tests d'hypothèses.
- Une bonne base de statistique et de probabiltés 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.

Objectif du cours

- Pré-requis indispensable : Cours L3-S5 - Statistique & Probabilités.
- Approche pédagogique : Sept séances de Cours Magistraux (durée 3h - Amphi) et sept séances de Travaux Dirigés (durée 2h - Salle de TD).
- Matériel pédagogique : Polycopiés de support de cours et de Travaux Dirigés avec indications de correction. Polycopié supplémentaire de TD, Problèmes de révision et Tests dauto-évaluation avec corrigés. Aide mémoire, tables statistiques et synthèse des principales statistiques de test.
- Modalités de Contrôle des Connaissances : 2 Contrôles Continus (50% - durée 1h30).
  Contrôle Continu n°1 : Vendredi 21 février 2025, 17h - 18h 30, Amphis SAY & AUBRAC
  Contrôle Continu n°2 : Vendredi 28 mars 2025, 17h - 18h 30, Amphis SAY & AUBRAC
- Quelques références bibliographiques :

[1] R. Abdesselam, ***"Statistique Inférentielle. Exercices d'application et problèmes corrigés avec rappels de cours"***. La collection Références sciences, Editions Ellipses, 2020.

[2] P. Roger ***"Probabilités, statistique et processus stochastiques"*** Cours et exercices. Collection synthex, Pearson Education.

[3] B. Grais ***"Méthodes statistiques"*** Modules Économiques, Dunod.

[4] Y. Herbert ***"Mathématiques probabilités et statistique"*** Vuibert.

[5] Sheldon Y. Ross ***"Initiation aux probabilités"*** Traduction de la 4ème Etidion américaine Presses Polytechniques et Universitaires Romandes.

[6] G.R. Grimmett and D.R. Stirzaker ***"Probability and Random Processes"*** Oxford Science Publications.

Plan détaillé

- Introduction
  - Echantillonnage - Estimation de paramètres - Population, échantillon.
  - Théorème de la limite centrale - Convergence en loi et en probabilité.

- Chapitre 1 : Estimation ponctuelle & Intervalle de confiance
  - Lois construites à partir de la loi normale : Khi-deux de Pearson, Student et Fisher-Snédécor, lecture des tables.
  - Estimateurs - Construction : Méthode du Maximum de Vraisemblance, Méthode des Moments - Propriétés.
  - Estimation ponctuelle - intervalle de confiance : moyenne, proportion et variance. Comparaisons : moyennes (échantillons indépendants - appariés), proportions et rapport de variances.

Contrôle Continu n°1

- Chapitre 2 : Tests d'hypothèses paramétriques
  - Concept et formulation des hypothèses et conditions d'application. Démarche d'un test statistique.
  - Risques de première et deuxième espèce.
  - Tests de conformité d'une moyenne, d'une proportion et d'une variance. comparaisons : moyennes, proportions et variances.

Plan détaillé

- Chapitre 3 : Tests d'hypothèses non-paramétriques
  - Principe général et formulation des hypothèses.
  - Application des tests du Khi-deux de Pearson :
    - Test d'indépendance entre 2 caractères - Tableau de contingence.
    - Test d'homogénéité de plusieurs populations.
    - Test d'ajustement - Conformité entre deux distributions.
  - Echantillons indépendants
    - Test de la somme des rangs (Wilcoxon & Mann-Whitney) ou Test U de Mann-Whitney.
  - Echantillons appariés
    - Test de la somme des rangs des différences positives (Wilcoxon).
    - Test de corrélation de rangs de Spearman.

Contrôle Continu n°2

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)
   - démarche inductive "du particulier au général". On connaît l'échantillon, on s'intéresse à la population.
   - Elle vise à étudier, à prédire les paramètres d'une population inconnue à partir des résultats obtenus grâce à des échantillons.
   - Estimation ponctuelle et par intervalle de confiance,
   - Tests d'hypothèses statistiques.

ESTIMATION (Echantillon $\mapsto$ Population)

- Schématiquement

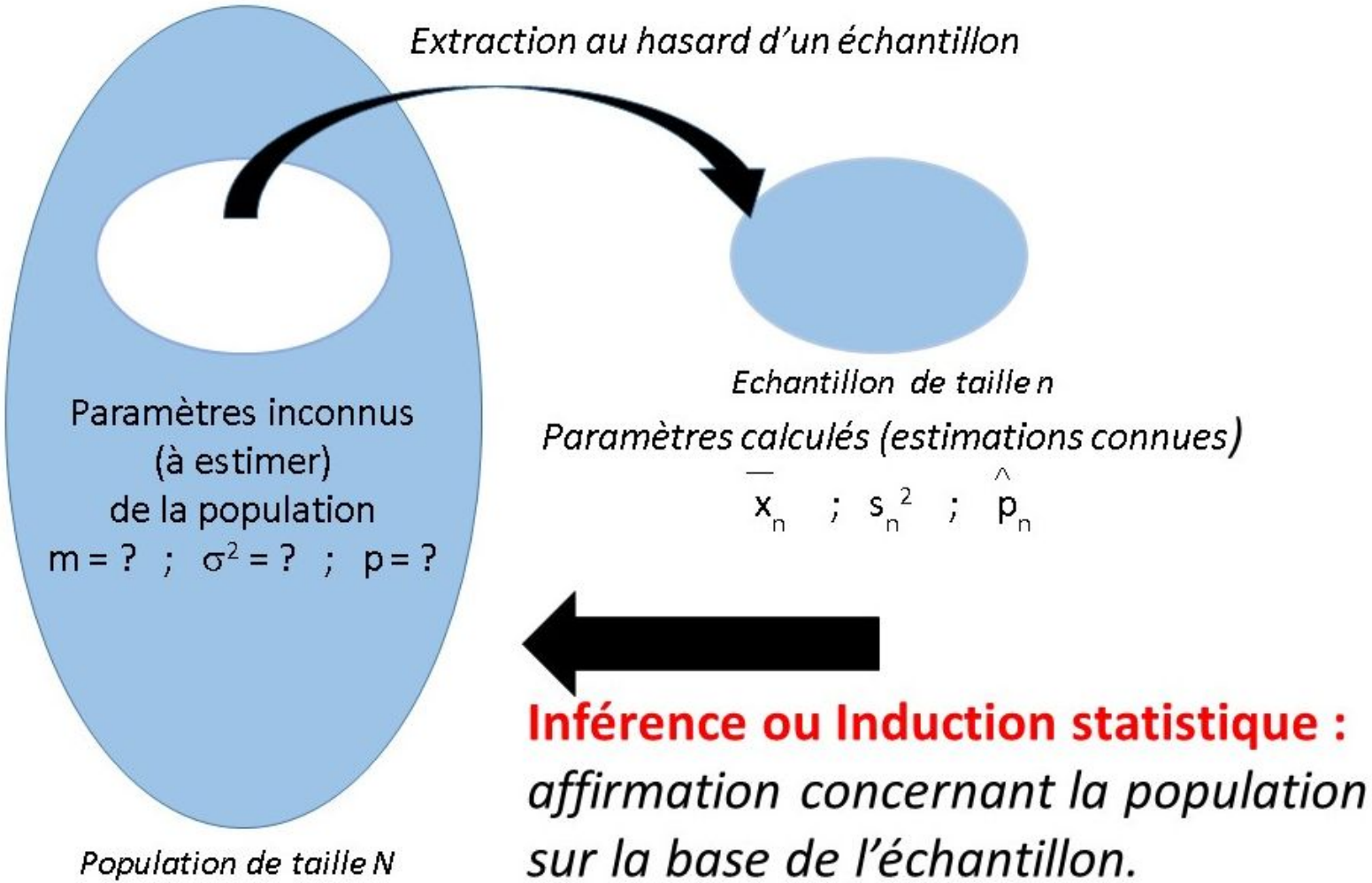


Fig. Processus de la statistique inférentielle

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.
- Exemple 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 :
    - la moyenne m (rendement moyen du titre)
    - la variance σ^2 (volatilité du titre).

Estimation de paramètres

- Exemple 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 probabilité 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 indépendantes et identiquement distribuées (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 $S_n = \sum_{i=1}^{n} X_i \Rightarrow E(S_n) = nm$ et $V(S_n) = n\sigma^2 = (\sigma\sqrt{n})^2$ alors :

  $$S_n \to_{n\to+\infty} N(nm; (\sigma\sqrt{n})^2) \quad \Leftrightarrow \quad \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - nm}{\sigma\sqrt{n}} \to_{n\to+\infty} N(0;1)$$

  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 : $\overline{X}_n = \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow E(\overline{X}_n) = m$ et $V(\overline{X}_n) = \frac{\sigma^2}{n}$ alors :

  $$\frac{\overline{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \to_{n\to+\infty} N(0;1)$$

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 :

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

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

Propriétés d'un estimateur

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

1. Estimateur sans biais : $E(\overline{\theta}_n) = \theta$.
2. Estimateur asymptotiquement sans biais : $E(\overline{\theta}_n) \rightarrow_{n\rightarrow\infty} \theta$
   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 : $Var(\overline{\theta}_n) \rightarrow_{n\rightarrow\infty} 0$.
4. Estimateur efficace :
   soient $\overline{\theta_n}$ et $\overline{\theta'_n}$ deux estimateurs sans biais de θ,
   $\overline{\theta}_n$ est dit plus efficace que $\overline{\theta'}_n$ si $Var(\overline{\theta}_n) \leq Var(\overline{\theta'}_n)$.

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 $\overline{\theta}_n$ 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 $\overline{\theta}_n$ d'après l'équation suivante :

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

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

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

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 $\overline{\theta}_n$ et du niveau de confiance $(1-\alpha)$ ou du risque d'erreur α choisi à priori.

Lois de probabilités utiles - Lecture des tables

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

$$\chi^2_n = \sum_{i=1}^{n} U_i^2 \qquad \begin{cases} \forall (i,j)\, i \neq j \\ U_i \rightarrow N(0;1) \\ U_i \text{ et } U_j \text{ indépendantes} \end{cases}$$

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^2_n divisée par son nombre de degrés de liberté (d.d.l.) n :

$$T_n = \frac{U}{\sqrt{\frac{X}{n}}} \qquad \begin{cases} U \rightarrow N(0;1) \\ X \rightarrow \chi^2_n \\ U \text{ et} X \text{ indépendantes} \end{cases}$$

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^2_m et χ^2_n indépendantes, divisées par leur nombre de d.d.l. m et n :

$$F_{(m,n)} = \frac{\frac{X}{m}}{\frac{Y}{n}} \qquad \begin{cases} X \rightarrow \chi^2_m \\ Y \rightarrow \chi^2_n \\ X \text{ et} Y \text{ indépendantes} \end{cases}$$

Hypothèses et conditions d'application

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

1. Si la loi de probabilité de X est normale, quelle que soit la taille n 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 de petite taille $(n < 30)$: 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).

1 Estimation ponctuelle de la moyenne : m

- $(X_1, X_2, ..., X_n)$ 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 :

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

1. Estimateur **sans biais** :

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

2. Estimateur **convergent** en probabilité :

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

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

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

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

$$\boxed{\frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \hookrightarrow N(0;1)}$$

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

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

  $P\left(-u_{\frac{\alpha}{2}} \leq U \leq u_{\frac{\alpha}{2}}\right) = P\left(|U| \leq u_{\frac{\alpha}{2}}\right) = 1 - \alpha$
- On en déduit l'intervalle de confiance de niveau $(1-\alpha)$ de m :

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

- *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 $\overline{x}_n$.

1.2 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 :

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

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

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

$$\boxed{\frac{\overline{X}_n - m}{S_n^*/\sqrt{n}} \rightarrow T_{\nu=n-1 \ d.d.l.}}$$

- On peut alors écrire : $P\left(-t_{\frac{\alpha}{2}} \leq \frac{\overline{X}_n - m}{s_n^*/\sqrt{n}} \leq t_{\frac{\alpha}{2}}\right) = 1 - \alpha$

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

  $P\left(-t_{\frac{\alpha}{2}} \leq T_\nu \leq t_{\frac{\alpha}{2}}\right) = P\left(|T_\nu| \leq t_{\frac{\alpha}{2}}\right) = 1 - \alpha$
- On en déduit l'intervalle de confiance de niveau $(1-\alpha)$ de m :

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

Approximation - Exemple d'application 1

- 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 $\nu = (n-1)$ d.d.l. par celle du fractile $u_{\frac{\alpha}{2}}$ de la loi normale centrée-réduite $N(0,1)$ (cf. Tables statistiques $N(0,1)$ et T_ν).

- 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 de cette entreprise.

Exemple d'application 1 - Solution

Conditions d'application : n = 16 petit, mais C.A. mensuel est supposé normalement distribué :
$C.A. \to N(m, \sigma^2)$ de variance σ^2 inconnue.

- Statistique de test : $\frac{\overline{X}_n - m}{S_n^*/\sqrt{n}} \to T_{n-1\ d.d.l.}$

  Taille de l'échantillon et degrés de liberté : $n = 16 \Rightarrow \nu = n - 1 = 15\ d.d.l.$
  $\overline{x}_{16} = 250$ K.€. : moyenne calculée sur l'échantillon de taille $n = 16$ (estimation ponctuelle de m.)

  Estimation de l'écart-type σ : $s_{16} = 52$ K.€ $\Rightarrow s*_{16} = s_{16}\sqrt{\frac{n}{n-1}} = 52\sqrt{\frac{16}{15}} = 53.705$ K.€.

  Seuil de signification : $\alpha = 2\%$

  Fractiles de la loi de Student : $t_{\frac{\alpha}{2}} = t_{1\%} = \pm 2.6025$ cf. Table de Student à $\nu = n - 1 = 15\ d.d.l.$

- Marge d'erreur dans l'estimation de m :
  $$E = t_{\frac{\alpha}{2}} \frac{s*_{16}}{\sqrt{n}} = 2.6025 \frac{53.705}{\sqrt{16}} = 34.94 \quad \text{K.€.}$$

- Intervalle de confiance de niveau $1 - \alpha = 98\%$ de m (variance σ^2 inconnue) :
  $$250 - 34.94 = 215.06 \leq m \leq 250 + 34.94 = 284.94$$
  $$m \in [215.06 \text{ K.€. }, 284.94 \text{ K.€.}]$$

- Conclusion : Il y a 98 chances sur 100 que le C.A. mensuel moyen de cette entreprise se trouve dans cet intervalle.

2 Estimation ponctuelle d'une proportion : p

- Soient $A_1, \ldots, A_i, \ldots, A_n$ 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 : $\widehat{P}_n = \frac{1}{n}\sum_{i=1}^{n} 1_{A_i}$
  Estimation : $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} P(A_i)$

- L'estimateur ainsi défini,

  1. Estimateur **sans biais** :

  $$E(\widehat{P}_n) = E(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} 1_{A_i}) = \tfrac{1}{n}\sum_{i=1}^{n} E(1_{A_i}) = \tfrac{1}{n}\sum_{i=1}^{n} P(A_i) = \tfrac{np}{n} = p$$

  2. Estimateur **convergent** en probabilité :

  $$V(\widehat{P}_n) = Var(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} 1_{A_i}) = \tfrac{1}{n^2}\sum_{i=1}^{n} Var(1_{A_i})$$

  $$= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} P(A_i)(1 - P(A_i)) = \tfrac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \tfrac{npq}{n^2} \to 0$$

- *La proportion $\widehat{p}_n$ 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,

$$\begin{cases} (X_1,...,X_i,...,X_n)\ i.i.d. \\ X_i \to B(p)\ \textit{Bernouilli} \\ \forall i\ E(X_i)=p \text{ et } V(X_i)=pq \end{cases} \Rightarrow \begin{cases} Y=\sum_i^n X_i \to B(n;p)\ \textit{Binomiale} \\ E(Y)=np \\ V(Y)=npq\ \textit{avec}\ q=1-p \end{cases}$$

$$\Rightarrow \begin{cases} \widehat{P}_n = Y/n = 1/n \sum_i^n X_i \\ E(\widehat{P}_n) = \frac{1}{n}E(Y) = p \\ V(\widehat{P}_n) = \frac{1}{n^2}V(Y) = \frac{pq}{n} \end{cases}$$

$$\Rightarrow \text{T.C.L. n grand : n} \geq 30 \begin{cases} \widehat{P}_n \to N(p;(\sqrt{\frac{pq}{n}})^2) \\ E(\widehat{P}_n)=p \text{ et } V(\widehat{P})=\frac{pq}{n} \\ \frac{(\widehat{P}_n-p)}{\sqrt{\frac{pq}{n}}} \hookrightarrow N(0;1)\ \ \textit{Approximativement Normale} \end{cases}$$

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 :

$$\widehat{P}_n = \frac{1}{n}\sum_i^n X_i \qquad (X_1, ..., X_i, ..., X_n) \textbf{ i.i.d. } X_i \to B(p)$$

- La statistique de test :

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

- La distribution d'échantillonnage de $\widehat{P}_n$ est approximativement normale de moyenne $E(\widehat{P}_n) = p$ et de variance en fonction de p (inconnue) $Var(\widehat{P}_n) = \frac{pq}{n} \leftarrow \frac{\widehat{p}_n \widehat{q}_n}{n}$ (estimation de la variance).

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

- On peut alors écrire : $P\left(-u_{\frac{\alpha}{2}} \leq \frac{\widehat{P}_n - p}{\sqrt{\frac{\widehat{p}_n \widehat{q}_n}{n}}} \leq u_{\frac{\alpha}{2}}\right) = 1 - \alpha$
- On en déduit l'intervalle de confiance de niveau $(1-\alpha)$ de p :

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

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

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 de la proportion d'employés de cette entreprise dont le salaire est supérieur ou égal à 2 M.€., avec un risque d'erreur de $\alpha = 10\%$, de $\alpha = 5\%$ et de $\alpha = 1\%$.

Exemple d'application 2 - Solution

Conditions d'application : n = 50, approximation par une distribution normale.

- Statistique de test : $\frac{\widehat{P}_n - p}{\sqrt{\frac{p_n q_n}{n}}} \hookrightarrow N(0;1)$
- Estimation ponctuelle de la proportion vraie d'employés de cette entreprise dont le salaire est supérieur ou égal à 2 M.€. :
$$\widehat{p}_{50} = \frac{20+12}{50} = \frac{32}{50} = 64\%.$$
- Intervalle de confiance la proportion d'employés de cette entreprise dont le salaire est supérieur ou égal à 2 M.€. :

  Niveau de confiance : $1-\alpha = 95\%$; risque d'erreur : $\alpha = 5\%$
  Fractiles de la loi normale : $u_{\frac{\alpha}{2}} = u_{2.5\%} = \pm 1.96$ cf. Table $N(0\,,1)$
  Marge d'erreur dans l'estimation de p : $E = u_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\widehat{q}}{n}} = 1.96\sqrt{\frac{0.64\times 0.36}{50}} = 13.30\%$. Avec $\widehat{p}_{50} = 64\%$ et $\widehat{q}_{50} = 1-\widehat{p}_{50} = 36\%$.
  Intervalle de confiance de niveau 95% de p :
  $$0.64 - 0.1330 = 0.5070 \leq p \leq 0.64 + 0.1330 = 0.7730 \Rightarrow p \in [50.70\%\,,\,77.30\%]$$
- Conclusion : dans cette entreprise, Il y a 95 chances sur 100 que la proportion d'employés dont le salaire est supérieur ou égal à 2 M.€., soit comprise entre 50.70% et 77.30%.

α	$1-\alpha$	$u_{\frac{\alpha}{2}}$	Marge	Intervalle de confiance
10%	90%	1.645	0.1117	[52.83% , 75.17%]
5%	95%	1.960	0.1330	[50.70% , 77.30%]
1%	99%	2.580	0.1751	[46.49% , 81.51%]

Précision - Taille d'échantillon - Risque d'erreur ou Niveau de confiance

1. La marge d'erreur ou niveau de précision recherché dans l'estimation par intervalle de confiance, lorsqu'on utilise l'estimation $\overline{\theta}_n$ de l'échantillon pour estimer la vraie valeur θ de la population, est l'écart (en valeur absolue), noté $E = |\overline{\theta}_n - \theta|$.
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	$E = u_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{pq}}{n}}$	$n = (\frac{u_{\frac{\alpha}{2}}}{E})^2 \widehat{pq}$	$u_{\frac{\alpha}{2}} = \sqrt{\frac{n}{\widehat{pq}}} E$

- Propriété : Dans les cas d'une moyenne (variance connue) ou d'une proportion, réduire la marge d'erreur (précision) de moitié (k = 2), nécessite une taille d'échantillon $k^2 = 2^2 = 4$ fois plus grande.

Moyenne (variance connue)		Proportion
$E = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$		$E = u_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{pq}}{n}}$
$E' = \frac{E}{k} = \frac{E}{2}$		$E' = \frac{E}{k} = \frac{E}{2}$
$u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n^*}} = \frac{1}{2} u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$		$u_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{pq}}{n^*}} = \frac{1}{2} \sqrt{\frac{\widehat{pq}}{n}}$
$\frac{1}{\sqrt{n^*}} = \frac{1}{2} \frac{1}{\sqrt{n}}$		$\frac{1}{\sqrt{n^*}} = \frac{1}{2} \frac{1}{\sqrt{n}}$
$\sqrt{n^*} = 2\sqrt{n}$		$\sqrt{n^*} = 2\sqrt{n}$
$n^* = 2^2 n = 4n$		$n^* = 2^2 n = 4n$

D'une façon générale : $n^* = k^2 n$.

- Plus le niveau de confiance est élevé plus la marge d'erreur est grande.
- Pour le même niveau de confiance et le même écart-type, plus la marge d'erreur requise est faible, plus la taille de l'échantillon sera élevée.

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 ?

Exemple d'application 3.1 - Solution

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

  Conditions d'application : C.A. mensuel m est supposé normalement distribué : $C.A. \rightarrow N(m, \sigma^2)$ de variance σ^2 inconnue.

- Statistique de test : $\frac{\overline{X}_n - m}{S_n^*/\sqrt{n}} \rightarrow T_{n-1\ d.d.l.}$

  Remarque : pour n = 16 mois et avec un risque d'erreur $\alpha = 2\%$, la marge d'erreur dans l'estimation du C.A. mensuel moyen était de $E = 34.94$ K.€.. On réduit cette marge à $E' \leq 15$ K.€. donc la taille d'échantillon requise n' va augmenter ($n' > 16$ mois) : la condition d'approximation de la loi de Student par celle de la loi normale (n' grand). On pourra donc remplacer le fratile de la loi de Student $t_{\frac{\alpha}{2}}$ (le nombre de degrés de liberté dépend de la taille d'échantillon n' recherchée) par celui de la loi normale $u_{\frac{\alpha}{2}}$.

  Risque d'erreur : $\alpha = 2\%$
  Fractiles de la loi normale : $u_{\frac{\alpha}{2}} = u_{1\%} = \pm 2.325$ cf. Table $N(0\,,1)$

  Marge d'erreur dans l'estimation de m :

$$E' = t_{\frac{\alpha}{2}} \frac{s^*}{\sqrt{n'}} \approx u_{\frac{\alpha}{2}} \frac{s^*}{\sqrt{n'}} \leq 15 \Rightarrow n' \geq \left(\frac{u_{\frac{\alpha}{2}} s^*}{E'}\right)^2 = \left(\frac{2.325 \times 53.705}{15}\right)^2 = 69.29$$

- Conclusion : Pour avoir un intervalle de confiance de niveau 98% avec une marge d'erreur au plus égale à 15 K.€., la taille d'échantillon requise doit être supérieure ou égale à 70 relevés mensuels.

Exemple d'application 3.2 - Solution

- Taille de l'échantillon requise 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 $\alpha = 5\%$:

  Remarque : pour n = 50 employés et avec un risque d'erreur $\alpha = 5\%$, la marge d'erreur était de $E = 13.33\%$. On réduit cette marge (meilleure précision) à $E' \leq 4\%$ donc la taille d'échantillon requise n' va augmenter ($n' > 50$ employés) : la condition d'application (n' grand) sera toujours respectée.

- Statistique de test : $\frac{\widehat{P}_n - p}{\sqrt{\frac{pq}{n}}} \hookrightarrow N(0;1)$.
  Niveau de confiance : $1 - \alpha = 95\%$.
  Risque d'erreur : $\alpha = 5\%$
  Fractiles de la loi normale : $u_{\frac{\alpha}{2}} = u_{2.5\%} = \pm 1.96$ cf. Table $N(0\,,1)$
  Marge d'erreur dans l'estimation de p :
  $$E' = u_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{pq}}{n'}} \leq 4\% \Rightarrow n' \geq (\frac{u_{\frac{\alpha}{2}}}{E'})^2 \widehat{pq} = (\frac{1.96}{0.04})^2 0.64 \times 0.36 = 553.19$$

- Conclusion : Pour avoir un intervalle de confiance de niveau 95% avec une marge d'erreur au plus égale à 4%, il faudrait interroger 554 employés et plus.

Exemple d'application 3.3 - Solution

- Taille minimale n de l'échantillon :
  Niveau de confiance : $1-\alpha = 95\%$, Risque d'erreur : $\alpha = 5\%$
  Fractile : $u_{\frac{\alpha}{2}} = u_{2.5\%} = \pm 1.96$ cf. table N(0 , 1).
- Statistique de test : $\frac{\widehat{P}_n - p}{\sqrt{\frac{pq}{n}}} \hookrightarrow N(0;1)$.
  Sans aucune information préalable sur la proportion p, on l'estime par $\widehat{p} = \widehat{q} = 50\%$, estimation ponctuelle de la proportion des votants.

  Marge d'erreur : $E = u_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{pq}}{n}} \leq 0.03 \Rightarrow n \geq (\frac{u_{\frac{\alpha}{2}}}{E})^2 \widehat{pq} = (\frac{1.96}{0.03})^2 \times 0.50^2 = 1067.11$

  Il faudrait interroger $n \geq 1068$ personnes.

- Taille minimale n^* de l'échantillon :

  Pour réduire de moitié la marge de l'estimation de la proportion p des votants, il faudrait interroger 4 fois plus de personnes : $n^* = 2^2 n = 4268$. cf. propriété

- Niveau de confiance de l'intervalle de la proportion $p \in [48.71\% ; 51.29\%]$ de votants :

  Estimation ponctuelle : $\widehat{p} = \widehat{q} = 50\%$., Marge d'erreur : $E = \frac{(51.29-48.71)}{2} = 1.29\%$.

  $$E = u_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{pq}}{n}} = 0.0129 \Rightarrow u_{\frac{\alpha}{2}} = E\sqrt{\frac{n}{pq}} = 0.0129\sqrt{\frac{100^2}{0.50^2}} = 2.58$$
  $$u_{\frac{\alpha}{2}} = 2.58 \Rightarrow \Phi(2.58) = 0.995 = 1 - \frac{\alpha}{2}$$
  $$\Rightarrow \frac{\alpha}{2} = 0.5\% \Rightarrow \alpha = 1\% \text{ cf. table N(0 , 1).}$$

  D'où le niveau de confiance : $1-\alpha = 99\%$.

3.1 Estimation de la variance σ^2 (m connue)

- $x_1, x_2, \ldots, x_n$ 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 :

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

1. Estimateur **sans biais** :

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

2. Estimateur **convergent** :

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

- *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.*

Intervalle de confiance de σ^2 (m connue)

- Lorsque la moyenne m est connue, on peut montrer que :
  $\sum_{i=1}^{n} \frac{(X_i-m)^2}{\sigma^2} = \sum_{i=1}^{n} (\frac{X_i-m}{\sigma})^2 = \sum_{i=1}^{n} U_i^2 \to \chi^2_{n\ d.d.l.}$ avec $U_i \to N(0,1)$.
  (cf. définition d'une variable aléatoire du khi-deux comme somme de carrés de variables aléatoires normales centrées réduites indépendantes).
- La statistique de test :

$$\boxed{\sum_{i=1}^{n} \frac{(X_i-m)^2}{\sigma^2} = n\frac{S_n^2}{\sigma^2} \to \chi^2_{\nu=n\ d.d.l.}}$$

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

  où, $k_1 = \chi^2_{\frac{\alpha}{2}}$ et $k_2 = \chi^2_{1-\frac{\alpha}{2}}$ sont les fractiles de la du khi-deux à $\nu = n$ degrés de liberté (cf. table du khi-deux).
  c'est-à-dire : $P(\chi^2_\nu \leq k_1) = \frac{\alpha}{2}$ et $P(\chi^2_\nu \leq k_2) = 1-\frac{\alpha}{2}$
- On en déduit l'intervalle de confiance de niveau $(1-\alpha)$ de σ^2 :

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

3.2 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 :

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

  Cependant, l'estimateur S_n^2 est biaisé : $E(S_n^2) = \frac{n-1}{n}\sigma^2$, on préfère alors utiliser l'estimateur :

  $$\boxed{S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n}{n-1}S_n^2}$$

  appelé : carré de la déviation standard empirique.

  1. Estimateur **sans biais** :

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

  2. Estimateur **convergent** : $Var(S_n^{*2}) =$

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

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

Intervalle de confiance de σ^2 (m inconnue)

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

$$\boxed{\sum_{i=1}^{n} \frac{(X_i - \overline{X}_n)^2}{\sigma^2} = (n-1)\frac{S_n^{*2}}{\sigma^2} \rightarrow \chi^2_{\nu=n-1 \ \ d.d.l.}}$$

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

  où, $k_1 = \chi^2_{\frac{\alpha}{2}}$ et $k_2 = \chi^2_{1-\frac{\alpha}{2}}$ sont les fractiles de la du khi-deux à $\nu = n-1$ degrés de liberté (cf. table du khi-deux).
  c'est-à-dire : $P(\chi^2_\nu \leq k_1) = \frac{\alpha}{2}$ et $P(\chi^2_\nu \leq k_2) = 1 - \frac{\alpha}{2}$
- 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. €. ?
  4. A quel niveau de confiance correspond l'intervalle [37.61 ; 90.96] de l'écart-type σ du C.A. mensuel de cette entreprise ?

Exemple d'application 4 - Solution (1/2)

Conditions d'application : échantillon de petite taille n = 16, le C.A. mensuel est normalement distribué de moyenne m et de variance σ^2 inconnues.

- Estimation ponctuelle de l'écart-type σ du chiffre d'affaires mensuel de cette entreprise (moyenne m inconnue) :

  $s_{16}^{*2} = \frac{n}{n-1} s_{16}^2 = \frac{16}{15} 52^2 \Rightarrow s_{16}^* = 53.70$ K. €., écart-type corrigé - Déviation standard.

- Statistique de test : $(n-1)\frac{S_n^{*2}}{\sigma^2} \rightarrow \chi^2_{(n-1)=15\ d.d.l.}$

  Risque d'erreur : $\alpha = 5\%$
  Fractiles de la loi du khi-deux à 15 d.d.l. : $k_1 = 6.262$ et $k_2 = 27.488$ cf. table du khi-deux.

- Intervalle de confiance de niveau $(1-\alpha) = 95\%$ de σ^2 :

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

$$1573.61 = 15\frac{53.70^2}{27.488} \leq \sigma^2 \leq 15\frac{53.70^2}{6.262} = 6907.59$$

  Ou encore pour l'écart-type σ : $39.67 = \sqrt{1573.61} \leq \sigma \leq \sqrt{6907.59} = 83.11$

  Intervalle de confiance de niveau $1-\alpha = 95\%$: $\sigma \in$ [39.67 K.Euros , 83.11 K.Euros]

- Comme $80 \in [39.67\ ,\ 83.11]$, on peut donc conclure avec un risque d'erreur $\alpha = 5\%$, que l'écart-type σ du C.A. mensuel de cette entreprise n'est pas significativement différent de 80 K.€..

Exemple d'application 4 - Solution (2/2)

- Niveau de confiance de $\sigma \in [37.61\ ;\ 90.96]$: Ecart-type du C.A. mensuel de l'entreprise.

On a : $n = 61$, $s_{16} = 52$ K. €.

La statistique de test : $\frac{(n-1)S^{*2}}{\sigma^2} \rightarrow \chi^2_{n-1=15\,d.d.l.}$

Intervalle de confiance : $\frac{(n-1)s^{*2}}{k_2} \leq \sigma^2 \leq \frac{(n-1)s^{*2}}{k_1}$

On en déduit :

$$\begin{cases} \frac{(n-1)s^{*2}}{k_2} = \frac{ns^2}{k_2} = 90,96^2 \Rightarrow k_2 = \frac{16 \times 52^2}{37.61^2} = 30.58 \\ \frac{(n-1)s^{*2}}{k_1} = \frac{ns^2}{k_1} = 37,61^2 \Rightarrow k_1 = \frac{16 \times 52^2}{90.96^2} = 5.229 \end{cases}$$

Pour $n = 16$; $\nu = n - 1 = 15$ d.d.l. cf. table du Khi-deux à $\nu = 15$ d.d.l. :

$k_1 = 5.229 : P(\chi^2_{15} \leq 5.229) = \frac{\alpha}{2} = 1\%$ et $k_2 = 30.58 : P(\chi^2_{15} \leq 30.58) = 1 - \frac{\alpha}{2} = 99\%$.

En en déduit que : $\alpha = 2\%$ et $1 - \alpha = 98\%$.

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 $E(\overline{X}_n) = m$:

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

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

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

$$V(\widehat{P}_n) = \frac{\widehat{pq}}{n}\frac{N-n}{N-1}$$

- 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 $\left(\frac{N-n}{N-1} \approx 1\right)$ 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

1. Donner une estimation de la proportion de l'ensemble des employés dont le salaire mensuel est de 2 M.€. et plus.
2. Quel est le taux de sondage ?
3. 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.€.
4. 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%.

Exemple d'application 5 - Solution

Conditions d'application : n = 50, approximation par une distribution normale.

- Estimation ponctuelle de la proportion vraie d'employés de cette entreprise dont le salaire est supérieur ou égal à 2 M.€. : $\widehat{p} = \frac{20+12}{50} = \frac{32}{50} = 64\%$.
- Taux de sondage : Taille de la population : $N = 200$ employés.
  $t = \frac{n}{N} = \frac{50}{200} = 25\%$ ($> 10\%$: facteur de correction : $\frac{N-n}{N-1}$)
- Statistique de test : $\frac{\widehat{P}-p}{\sqrt{\frac{pq}{n}(\frac{N-n}{N-1})}} \hookrightarrow N(0;1)$
- Intervalle de confiance la proportion d'employés dont le salaire est supérieur ou égal à 2 M.€. :

  Niveau de confiance : $1-\alpha = 95\%$; risque d'erreur : $\alpha = 5\%$
  Fractiles de la loi normale : $u_{\frac{\alpha}{2}} = u_{2.5\%} = \pm 1.96$ cf. Table $N(0\,,\,1)$

  Marge d'erreur dans l'estimation de p : $E = u_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{pq}}{n}(\frac{N-n}{N-1})} = 1.96\sqrt{\frac{0.64\times 0.36}{50}(\frac{150}{199})} = 11.55\%$.
  Avec $\widehat{p} = 64\%$ et $\widehat{q} = 1-\widehat{p} = 36\%$.

  Intervalle de confiance de niveau 95% de p :
  $0.64 - 0.1155 = 0.5245 \leq p \leq 0.64 + 0.1155 = 0.7555 \Rightarrow p \in [52.45\%\,,\,75.55\%]$
- 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% :

  Ecart-type de l'estimateur : $\sigma(\widehat{P}) = \sqrt{\frac{\widehat{pq}}{n}(\frac{N-n}{N-1})} = 0.059$

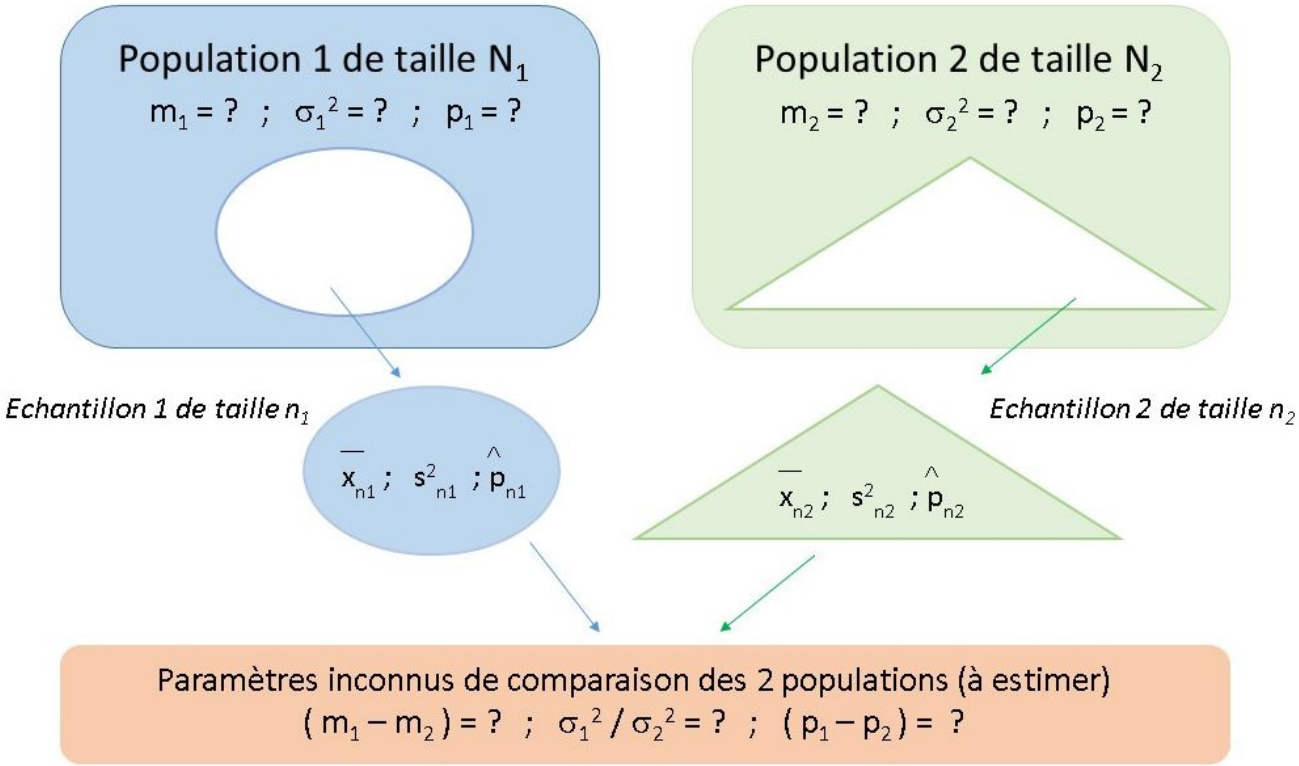
  $$P[\widehat{P} > \tfrac{30}{50}] = 1 - P[\widehat{P} \leq 0.6] = 1 - P[U = \frac{(\widehat{P}-p)}{\sigma(\widehat{P})} \leq \frac{(0.6-0.64)}{0.059}]$$
  $$= 1 - \Phi(-0.68) = \Phi(0.68) = 75.17\% \text{ cf. Table } N(0\,,1).$$

Comparaisons de 2 échantillons indépendants

- 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 du rapport de 2 variances observées .

Comparaisons de 2 échantillons indépendants



- Indépendants signifie que l'échantillon 1 est constitué de manière indépendante de l'échantillon 2 (par opposition aux échantillons appariés) :
- Les individus de l'échantillon 1 ne sont pas les mêmes que ceux de l'échantillon 2.

Estimation de la différence de 2 moyennes

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

  2 Populations, 2 échantillons indépendants

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

- On suppose normales les 2 populations, avec respectivement des moyennes m_x et m_y, et des variances σ_x^2 et σ_y^2.

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

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 : $\overline{X}_n - \overline{Y}_p$

   Estimateur **sans biais** :

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

   Estimateur **convergent** :

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

   La différence des moyennes $(\overline{x}_n - \overline{y}_p)$ 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 $(\widehat{p}_x - \widehat{p}_y)$ observée sur les échantillons est une estimation ponctuelle de la différence des proportions $(p_x - p_y)$ des populations.

Intervalle de confiance - Différence de 2 moyennes : $m_x - m_y$

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

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

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

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

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

- La statistique de test :

$$\frac{(\overline{X}_n - \overline{Y}_p) - (m_x - m_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}}} \rightarrow N(0;1)$$

- Ce qui fournit aisément un intervalle de confiance de niveau $(1-\alpha)$ pour la différence $(m_x - m_y)$:

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

- *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

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

Exemple d'application 6 - Solution

- Remarques : Petits échantillons $n_A = n_B = 9$ pièces mais le temps de fabrication est supposé normalement distribué. Les variances $\sigma_A^2 = 25$ et $\sigma_B^2 = 36$ sont connues.
- Statistique de test : $\frac{(\overline{X}_A - \overline{X}_B) - (m_A - m_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \rightarrow N(0;1)$.
- Les données :
  $n_A = n_B = n = 9$.
  Niveau de confiance : $1 - \alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.
  $u_{\frac{\alpha}{2}} = u_{2.5\%} = \pm 1.96$ cf. Table de la loi normale $N(0\,,\,1)$
  Marge d'erreur dans l'estimation de $(m_A - m_B)$: $E = u_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{n}} = 1.96\sqrt{\frac{25+36}{9}} = 5.10mn$
  Estimation ponctuelle de la différence $(m_A - m_B)$: $\overline{x}_A - \overline{x}_B = 50 - 45 = 5mn$.
- Intervalle de confiance de niveau 95% de $(m_A - m_B)$:
  $$5 - 5.10 = -0.10 \leq (m_A - m_B) \leq 5 + 5.10 = 10.10$$
  $$(m_A - m_B) \in [-0.10mn\,,\,10.10mn]$$
- Conclusion : $0 \in I.C._{95\%}$, donc la différence de 5 mn observée sur les échantillons n'est pas significative (avec un risque d'erreur de 5%), on peut donc considérer que ces deux machines ont des performances identiques.
- Question : oui, la machine B est aussi performante que la machine A, l'écart observé de 5 mn n'est pas significatif, il est dû aux fluctuations d'échantillonnage.

Intervalle de confiance - Différence de 2 moyennes : $m_x - m_y$

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

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

- 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}. Dans ce cas,
- La distribution d'échantillonnage de la différence $(\overline{X}_n - \overline{Y}_p)$ est approximativement normale. La statistique de test :

$$\frac{(\overline{X}_n - \overline{Y}_p) - (m_x - m_y)}{\sqrt{\frac{s_x^{*2}}{n} + \frac{s_y^{*2}}{p}}} \rightarrow N(0;1)$$

- Ce qui fournit aisément un intervalle de confiance de niveau $(1-\alpha)$ pour la différence $(m_x - m_y)$:

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

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 Ecarts	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 y avoir une différence significative entre ces deux groupes ?

Exemple d'application 7 - Solution

- Remarques : Grands échantillons $n_1 = 34$ et $n_2 = 32$ indépendants. Les variances σ_1^2 et σ_2^2 sont inconnues.
- Statistique de test : $$\frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}}} \hookrightarrow N(0;1).$$
- Les données : $n_1 = 34$ et $n_2 = 32$.
  Niveau de confiance : $1 - \alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.
  Fractile de la loi normale : $u_{\frac{\alpha}{2}} = u_{2.5\%} = \pm 1.96$ cf. Table de la loi normale $N(0\,,\,1)$
  Estimation des variances : $s_1^{*2} = \frac{SCE_1}{n_1 - 1} = \frac{15774}{33} = 478$ et $s_2^{*2} = \frac{SCE_2}{n_2 - 1} = \frac{9858}{31} = 318$.
  Marge d'erreur dans l'estimation de $(m_A - m_B)$: $E = u_{\frac{\alpha}{2}} \sqrt{\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}} = 1.96 \sqrt{\frac{478}{34} + \frac{318}{32}} = 9.6$
  Estimation ponctuelle de la différence $(m_1 - m_2)$: $\overline{x}_1 - \overline{x}_2 = 184 - 178 = 6$.
- Intervalle de confiance de niveau 95% de $(m_1 - m_2)$:
  $6 - 9.6 = -3.6 \leq (m_1 - m_2) \leq 6 + 9.6 = 15.6 \Rightarrow (m_1 - m_2) \in [-3.60\,,\,15.60]$
- Conclusion : $0 \in I.C._{95\%}$, donc la différence de 6 points observée sur les appréciations moyennes n'est pas significative (avec un risque d'erreur de 5%), on peut donc considérer que les deux groupes de cadres ont des appréciations globales identiques.
- Question : oui, en moyenne, la performance est identique entre ces deux groupes de cadres. L'écart observé de 6 points est attribuable aux fluctuations d'échantillonnage.

Intervalle de confiance - Différence de 2 moyennes : $m_x - m_y$

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 - Différence de 2 moyennes : $m_x - m_y$

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.

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

- La statistique de test :

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

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

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

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 :

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

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

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

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

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

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

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

Propriété, $\left. \begin{array}{c} X \rightarrow \chi^2_m \\ Y \rightarrow \chi^2_n \\ X \text{ et } Y \text{ indépendantes} \end{array} \right\} \Rightarrow X + Y \rightarrow \chi^2_{m+n}$

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

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

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

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

on a alors, $\dfrac{\textbf{(1)}}{\sqrt{\frac{\textbf{(2)}}{n+p-2}}} = \dfrac{\frac{(\overline{X}_n-\overline{Y}_p)-(m_x-m_y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{p}}}}{\sqrt{\frac{\frac{(n-1)s_x^{*2}+(p-1)s_y^{*2}}{\sigma^2}}{n+p-2}}} = \dfrac{(\overline{X}_n-\overline{Y}_p)-(m_x-m_y)}{\sqrt{\frac{(n-1)s_x^{*2}+(p-1)s_y^{*2}}{n+p-2}}\sqrt{\frac{1}{n}+\frac{1}{p}}} = \dfrac{(\overline{X}_n-\overline{Y}_p)-(m_x-m_y)}{s^*\sqrt{\frac{1}{n}+\frac{1}{p}}}$

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

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 inconnues 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 Ecarts	2400	2800

1. Déterminer un intervalle de confiance de niveau 95% de la différence des durées de vie moyennes des ampoules de ces deux fabricants.
2. 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 ?

Exemple d'application 8 - Solution

- Remarques : petits échantillons $n_1 = n_2 = n = 21$ indépendants. Les variances σ_1^2 et σ_2^2 sont inconnues mais supposées égales : $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Statistique de test : $\frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{s^* \sqrt{\frac{2}{n}}} \mapsto T_{2(n-1)=40\, d.d.l.}$
- Les données :
  $n_1 = n_2 = n = 21$.
  Niveau de confiance : $1 - \alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.
  $t_{\frac{\alpha}{2}} = t_{2.5\%} = \pm 2.021$ cf. Table de la loi de Student à 40 d.d.l.
  Estimation de la variance commune : $s^{*2} = \frac{SCE_1 + SCE_2}{2(n-1)} = \frac{2400+2800}{40} = 11.40^2$.
  Marge d'erreur dans l'estimation de $(m_1 - m_2)$: $E = t_{\frac{\alpha}{2}}\, s^* \sqrt{\frac{2}{n}} = 2.021 \times 11.40 \sqrt{\frac{2}{21}} = 7.11h$.
  Estimation ponctuelle de la différence $(m_1 - m_2)$: $\overline{x}_1 - \overline{x}_2 = 1025 - 1070 = -45h$.
- Intervalle de confiance de niveau 95% de $(m_1 - m_2)$:
  $$-45 - 7.11 = -52.11 \leq (m_1 - m_2) \leq -45 + 7.11 = -37.89$$
  $$(m_1 - m_2) \in [-52.11h\,,\ -37.89h]$$
- Conclusion : 0 n'appartient pas à $I.C._{95\%}$, l'écart de - 45 h observé sur les durées de vie moyennes est significatif (avec un risque d'erreur de 5%). Cet écart n'est donc pas attribuable aux fluctuations d'échantillonnage.
- Question : oui, la revue doit conclure, avec un risque d'erreur de 5%, que les durées de vie des ampoules de ces deux fabricants ne sont pas identiques : $m_1 \neq m_2$.

Intervalle de confiance - Différence de 2 moyennes : $m_x - m_y$

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

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

!!! CE CAS NE FAIT PAS PARTIE DU PROGRAMME POUR L'EXAMEN !!!

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

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

- Il faut déterminer ν pour pouvoir lire le fractile $t_{(\frac{\alpha}{2}\,;\,\nu)}$ dans la table de la loi de Student.

- Le calcul de ν est approché par l'équation de Welch-Satterthwaite :

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

Intervalle de confiance - Différence de 2 moyennes : $m_x - m_y$

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

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

- 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 $(m_x - m_y)$:

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

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 de rendements	$n_1 = 8$	$n_2 = 10$
rendement moyen observé (%)	$\overline{x}_1 = 9.8$	$\overline{x}_2 = 6.9$
Somme des Carrés des Ecarts	$SCE_1 = 250$	$SCE_2 = 61$

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 d'application 9 - Solution

- 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 : $$\frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{s_1^{2*}}{n_1} + \frac{s_2^{2*}}{n_2}}} \mapsto T_{\nu\ d.d.l.}$$
- 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 ν à partir de l'expression de Welch-Satterthwaite :
  $$\nu = \frac{(\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2})^2}{\frac{s_1^{*4}}{n_1^2(n_1 - 1)} + \frac{s_2^{*4}}{n_2^2(n_2 - 1)}} = \frac{(\frac{35.71}{8} + \frac{6.78}{10})^2}{\frac{35.71^4}{8^2 \times 7} + \frac{6.78^4}{10^2 \times 9}} = 9.1233$$
  Interpolation linéaire du fractile de la loi de Student : cf. Table à $\nu = 9$ d.d.l. et à $\nu = 10$ d.d.l.
  $t_{(2.5\%\ ;\ \nu=10\ d.d.l.)} = 2.2281$ et $t_{(2.5\%\ ;\ \nu=9\ d.d.l.)} = 2.2622 \Rightarrow t_{(\frac{\alpha}{2}=2.5\%\ ;\ \nu=9.12\ d.d.l.)} = \mp 2.2580$
  Marge d'erreur : $E = t_{(\frac{\alpha}{2}\ ;\ \nu=9.12)} \sqrt{\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}} = 2.2580 \times \sqrt{\frac{35.71}{8} + \frac{6.78}{10}} = 5.12\%$.
  Estimation ponctuelle de la différence $(m_1 - m_2)$: $\overline{x}_1 - \overline{x}_2 = 9.8 - 6.9 = 2.9\%$.
- 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 : $m_1 \approxeq m_2$.

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 (couples)

La méthode pour comparer 2 échantillons appariés consiste à établir :

- pour chaque paire, la différence $d_i = x_i - y_i$, on se ramène ainsi à un seul échantillon différence $(d_1, d_2, \ldots, d_n)$ de taille n,
- $D_1 = (X_1 - Y_1), \ldots, D_n = (X_n - Y_n)$ n variables aléatoires i.i.d.,
- Remarque : si la taille d'échantillon n est petite, $n < 30$, il faut que la différence $D = X - Y$ suive une loi normale.
- $\overline{D}_n = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)$ est l'estimateur sans biais et convergent de la moyenne différence m_d,
- la distribution d'échantillonnage $\overline{D}_n$ de la différence a pour moyenne $E(\overline{D}_n) = m_d = (m_x - m_y)$ et pour variance $V(\overline{D}_n) = \sigma_d^2 = \sigma_{x-y}^2$ inconnues,
- $\overline{d}_n = \overline{x}_n - \overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} d_i$, la moyenne calculée sur l'échantillon différence $(d_1, d_2, \ldots, d_n)$ est une estimation ponctuelle de la moyenne différence m_d,

Différence de 2 moyennes : Echantillons appariés

Echantillons dépendants - Données associées par paires

- $s_d^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2$ 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 :

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

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

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

Exemple d'application 10

- On mesure 12 pièces avec deux méthodes différentes. La différence des mesures est supposée normalement distribuée. On a obtenu les résultats suivants :

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

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

Exemple d'application 10 - Solution

- Echantillons appariés (dépendants). Conditions d'application : la mesure différence Z = X - Y est supposée normalement distribuée.
- Statistique de test : $\frac{\overline{Z}_n - m_z}{S^*_n/\sqrt{n}} \rightarrow T_{(n-1=11\ d.d.l.)}$
- Les données :
  $n = 12 \Rightarrow \nu = n - 1 = 11\ d.d.l.$
  $\overline{z}_{12} = \overline{x}_{12} - \overline{y}_{12} = 1 - 2.08 = -1.08$: moyenne calculée sur l'échantillon différence de taille $n = 12$ (estimation ponctuelle de m_z.)
  $s^{*2}_{12} = \frac{SCE_{z=x-y}}{n-1} = \frac{14.58}{11} = 1.3254 = 1.151^2$
  Seuil de signification : $\alpha = 5\%$
  $t_{\frac{\alpha}{2}} = t_{2.5\%} = \pm 2.201$ cf. Table de Student à $\nu = n - 1 = 11\ d.d.l.$
- Marge d'erreur dans l'estimation de m : $E = t_{\frac{\alpha}{2}} \frac{s*_{12}}{\sqrt{n}} = 2.201 \frac{1.151}{\sqrt{12}} = 0.7315$
- Intervalle de confiance de niveau 95% de m (variance σ^2_z inconnue) :
  $-1.08 - 0.7315 = -1.811 \leq m_z = m_x - m_y \leq -1.08 + 0.7315 = -0.3485$
  $m_z = (m_x - m_y) \in [-1.811\ ,\ -0.3485]$
- Conclusion : 0 n'appartient pas à $I.C._{95\%}$, l'écart de - 1.08 observé est significatif (avec un risque d'erreur de 5%). On peut donc conclure que $m_z = (m_x - m_y) \neq 0 \Leftrightarrow m_x \neq m_y$; les deux méthodes de mesures sont différentes.
- Remarque importante : Si on fait l'erreur de considérer ces deux échantillons de mesures comme des échantillons indépendants, on trouve un intervalle de confiance de niveau 95% de $(m_x - m_y) \in [-9.72\ ,\ 7.56]$.
  Dans ce cas, $0 \in I.C._{95\%}$ c'est-à-dire que $m_x \approx m_y$; les deux méthodes de mesures sont identiques.

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 - Statistique de test n°1)

- On sait déjà que :

$$\sum_{i=1}^{n_1} \frac{(X_i - \overline{X}_1)^2}{\sigma_1^2} = (n_1 - 1)\frac{S_1^{*2}}{\sigma_1^2} \rightarrow \chi^2_{\nu_1 = n_1 - 1 \ \ d.d.l.}$$

$$\sum_{i=1}^{n_2} \frac{(X_i - \overline{X}_2)^2}{\sigma_2^2} = (n_2 - 1)\frac{S_2^{*2}}{\sigma_2^2} \rightarrow \chi^2_{\nu_2 = n_2 - 1 \ \ d.d.l.}$$

- On peut alors montrer que la statistique de test :

$$\boxed{\frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^{*2}}{S_2^{*2}} \rightarrow F_{\nu_1 = n_1 - 1\,,\nu_2 = n_2 - 1} d.d.l.}$$

- On en déduit, au niveau $(1-\alpha)$, un intervalle de confiance pour le rapport $\frac{\sigma_2^2}{\sigma_1^2}$: $\quad f_1 \frac{s_2^{*2}}{s_1^{*2}} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq f_2 \frac{s_2^{*2}}{s_1^{*2}}$

  où, $\quad f_1 = f_{1-\frac{\alpha}{2}} = P(F_{(\nu_1,\nu_2)} > f_1) = 1 - \frac{\alpha}{2}$

  et $\quad f_2 = f_{\frac{\alpha}{2}} = P(F_{(\nu_1,\nu_2)} > f_2) = \frac{\alpha}{2}$

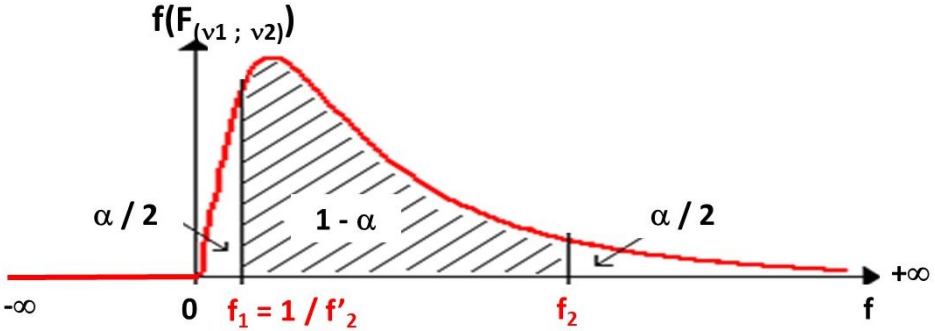
  sont les fractiles de la de Fisher-Snédécor à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté (cf. table).

Loi de Fisher-Snedecor

- Propriété :

$$X = \frac{\frac{\chi^2_{\nu_1}}{\nu_1}}{\frac{\chi^2_{\nu_2}}{\nu_2}} \rightarrow F_{(\nu_1,\nu_2)} \quad ; \quad \frac{1}{X} = \frac{\frac{\chi^2_{\nu_2}}{\nu_2}}{\frac{\chi^2_{\nu_1}}{\nu_1}} \rightarrow F_{(\nu_2,\nu_1)}$$

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



- $F(f_2) = P(F_{(\nu_1,\nu_2)} \leq f_2) = 1 - \frac{\alpha}{2}$: f_2 se lit sur la table $F_{(\nu_1,\nu_2)}$.
- $F(f_1) = P(F_{(\nu_1,\nu_2)} \leq f_1) = \frac{\alpha}{2}$

  $\Rightarrow P(\frac{1}{F_{(\nu_1,\nu_2)}} \leq \frac{1}{f_1}) = 1 - \frac{\alpha}{2}$

  $\Rightarrow P(F_{(\nu_2,\nu_1)} \leq \frac{1}{f_1} = f'_2) = 1 - \frac{\alpha}{2}$: $f'_2 = \frac{1}{f_1}$ se lit sur la table $F_{(\nu_2,\nu_1)}$,

  on en déduit $f_1 = \frac{1}{f'_2}$.

Rapport de 2 variances

(comparaison de 2 variances - Statistique de test n°2)

- On peut aussi montrer que la statistique de test :

$$\frac{\sigma_1^2}{\sigma_2^2}\frac{S_2^{*2}}{S_1^{*2}} \rightarrow F_{\nu_2=n_2-1\,,\nu_1=n_1-1}\, d.d.l.$$

- Cas particuliers :

  Si $n_1 = n_2 \Rightarrow \nu_1 = \nu_2$: $f_1 = \frac{1}{f_2}$

  Si $n_1 = n_2$ et $s_1^2 = s_2^2$: $f_1 \leq \frac{\sigma_2^2}{\sigma_1^2} \leq f_2 \Leftrightarrow f_1 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq f_2$

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 Ecarts	2400	2800

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

Exemple d'application 11 - Solution

- Remarques : petits échantillons $n_1 = n_2 = n = 21$ indépendants.
- Statistique de test : $\frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^{*2}}{S_2^{*2}} \to F_{(n_1-1=20\,;\,n_2-1=20)\ d.d.l.}$.
- Les données : niveau de confiance : $1-\alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.
  $f_2 = f_{2.5\%} = 2.464$ et $f_1 = f_{97.5\%} = \frac{1}{f_2} = \frac{1}{2.464} = 0.406$ cf. Table de la loi de Fisher $F_{(20;20)}$.
  Estimation des variances : $s_1^{*2} = \frac{SCE_1}{(n_1-1)} = \frac{2400}{20} = 120$ et $s_2^{*2} = \frac{SCE_2}{(n_2-1)} = \frac{2800}{20} = 140$.
- Intervalle de confiance de niveau 95% de $\frac{\sigma_2^2}{\sigma_1^2}$:

$$0.474 = 0.406\frac{140}{120} = f_1\frac{s_2^{*2}}{s_1^{*2}} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq f_2\frac{s_2^{*2}}{s_1^{*2}} = 2.464\frac{140}{120} = 2.875$$

$$\frac{\sigma_2^2}{\sigma_1^2} \in [0.474\ ,\ 2.875]$$

- Conclusion : $1 \in I.C._{95\%}$, il n'y a pas de différence significative (avec un risque d'erreur de 5%) entre les deux variances. On peut donc les supposer égales : $\sigma_1^2 \approx \sigma_2^2$.

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 (%)	$\overline{x}_1 = 9.8$	$\overline{x}_2 = 6.9$
Somme des Carrés des Ecarts	$SCE_1 = 250$	$SCE_2 = 61$

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

Exemple d'application 12 - Solution

- Cas de deux petits échantillons indépendants - $n_1 = 8$ et $n_2 = 10$.
- Statistique de test : $\frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^{*2}}{S_2^{*2}} \rightarrow F_{(n_1-1=7\,;\,n_2-1=9)\ d.d.l.}$.
- Les données : Niveau de confiance : $1-\alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.
  $f_2 = f_{2.5\%} = 4.20$ et $f_1 = f_{97.5\%} = \frac{1}{f_2} = \frac{1}{4.82} = 0.207$ cf. Table de la loi de Fisher $F_{(7;9)}$.
  Estimation des variances : $s_1^{*2} = \frac{SCE_1}{(n_1-1)} = \frac{250}{7} = 35.71$ et $s_2^{*2} = \frac{SCE_2}{(n_2-1)} = \frac{61}{9} = 6.78$.
- Intervalle de confiance de niveau 95% de $\frac{\sigma_2^2}{\sigma_1^2}$:

$$0.039 = 0.207 \times \frac{6.78}{35.71} = f_1\frac{s_2^{*2}}{s_1^{*2}} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq f_2\frac{s_2^{*2}}{s_1^{*2}} = 4.20 \times \frac{6.78}{35.71} = 0.797$$

$$\frac{\sigma_2^2}{\sigma_1^2} \in [0.039\ ,\ 0.797]$$

- Conclusion : $1 \notin I.C._{95\%}$, on peut donc conclure qu'il y a une différence significative entre les variances des rendements de ces deux banques. Avec un risque d'erreur $\alpha = 5\%$, on peut donc les considérer différentes ou inégales : $\sigma_1^2 \neq \sigma_2^2$.

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 $(\widehat{P}_1 - \widehat{P}_2)$ des deux proportions pour estimer, par intervalle de confiance, cette différence.
- Cas unique de grands échantillons prélevés au hasard dans 2 populations indépendantes. Dans ce cas, on sait que :

$$\left.\begin{array}{l} \dfrac{\widehat{P}_1 - p_1}{\sqrt{\dfrac{p_1 q_1}{n_1}}} \rightarrow N(0\,;\,1) \\ \dfrac{\widehat{P}_2 - p_2}{\sqrt{\dfrac{p_2 q_2}{n_2}}} \rightarrow N(0\,;\,1) \\ \widehat{P}_1 \text{ et } \widehat{P}_2 \text{ indépendantes} \end{array}\right\} \Rightarrow \frac{(\widehat{P}_1 - \widehat{P}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}} \rightarrow N(0\,;\,1)$$

Différence de 2 proportions

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

- La statistique de test :

$$\boxed{\frac{(\widehat{P}_1 - \widehat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \hookrightarrow N(0;1)}$$

- Marge d'erreur : $E = \pm u_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_1 \widehat{q}_1}{n_1} + \frac{\widehat{p}_2 \widehat{q}_2}{n_2}}$

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

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

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 $\widehat{p}$ en combinant les proportions observées dans chaque échantillon comme suit :

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

- On peut donc aussi utiliser la statistique de test :

$$\boxed{\frac{(\widehat{P}_1 - \widehat{P}_2) - (p_1 - p_2)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \hookrightarrow N(0;1)}$$

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

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

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 rejetterait 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 ?

Exemple d'application 13 - Solution

- Grands échantillons : $n_1 \geq 30$ et $n_2 \geq 30$).
- Statistique de test : $\dfrac{(\widehat{P}_1 - \widehat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \hookrightarrow N(0;1)$
- Les données : $n_1 = n_2 = 250$.
  Niveau de confiance : $1 - \alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.
  $u_{2.5\%} = \pm 1.96$ cf. Table de la loi Normale N(0,1).
  Estimations ponctuelles des proportions : $\widehat{p}_1 = \frac{110}{250} = 44\%$ et $\widehat{p}_2 = \frac{118}{250} = 47.2\%$.
  Estimation ponctuelle de la différence $(p_1 - p_2)$: $\widehat{p}_1 - \widehat{p}_2 = -3.2\%$.
- Marge d'erreur dans l'estimation de la différence $(p_1 - p_2)$:
  $E = u_{2.5\%} \sqrt{\frac{\widehat{p}_1 \widehat{q}_1}{n_1} + \frac{\widehat{p}_2 \widehat{q}_2}{n_2}} = 1.96 \sqrt{\frac{0.44 \times 0.56}{250} + \frac{0.472 \times 0.528}{250}} = \pm 8.7\%$
- Intervalle de confiance de niveau 95% de $(p_1 - p_2)$:
  $-11.9\% = -3.2\% - 8.7\% \leq (p_1 - p_2) \leq -3.2\% + 8.7\% = 5.5\%$
  $(p_1 - p_2) \in [-11.9\% \, , \, 5.5\%]$
- Conclusion : $0 \in I.C._{95\%}$, il n'y a pas de différence significative (avec un risque d'erreur de 5%) entre les deux proportions. L'écart observé de 3.2% est dû aux fluctuations d'échantillonnage. On peut donc les supposer comme égales : $p_1 \approx p_2$.