École Supérieure de Physique et de
Chimie Industrielles de la Ville de Paris

Statistiques Appliquées

Tests statistiques

Néo POTRON
Alex CASSI

143

Mai 2025

Table des matières

1 Tests à partir de données synthétiques

Dans cette partie nous allons nous même engendrer les données afin d'expérimenter le theorème de la limite centrale (ou TCL). Le théorème de la limite centrale, de façon informelle, donne une estimation précise de l'erreur que l'on commet en approchant l'espérance mathématique par la moyenne arithmétique (historiquement nommé théorème des erreurs par Gauss).

1.1 Étude de la distribution de l'estimateur de la moyenne par échantillonnage (cf. 4.1)

Théorème Central Limite

Soit X_n une suite de variables aléatoires de même loi, d'espérance μ et d'écart-type σ. Alors, selon le théorème central limite (TCL), la variable aléatoire suivante converge en loi vers une loi normale centrée réduite :

$$Z_n = \frac{(X_1 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}} \longrightarrow \mathcal{N}(0,1)$$

Autrement dit :

$$X_1 + \cdots + X_n \longrightarrow \mathcal{N}(n\mu, \sigma\sqrt{n})$$

Lorsque l'on prend $n = 10$ échantillons au hasard dans une population telle que la caractéristique observée est normée centrée, la moyenne obtenue n'est pas rigoureusement nulle. Ainsi, pour évaluer la variabilité de l'estimateur de la moyenne, nous avons simulé un grand nombre M d'échantillons indépendants de taille n. L'estimateur de la moyenne, calculé pour chaque échantillon, est lui-même une variable aléatoire. Théoriquement, sa distribution est une loi normale centrée en 0 avec un écart-type égal à $\sigma/\sqrt{n} = 1/\sqrt{10} \approx 0{,}316$. Pour visualiser cette distribution, nous avons simulé successivement $M = 100$, 1 000 et 10 000 échantillons, calculé leur moyenne et leur variance, puis tracé l'histogramme des valeurs obtenues. Ces histogrammes ont été comparés à la densité théorique de la loi normale correspondante.

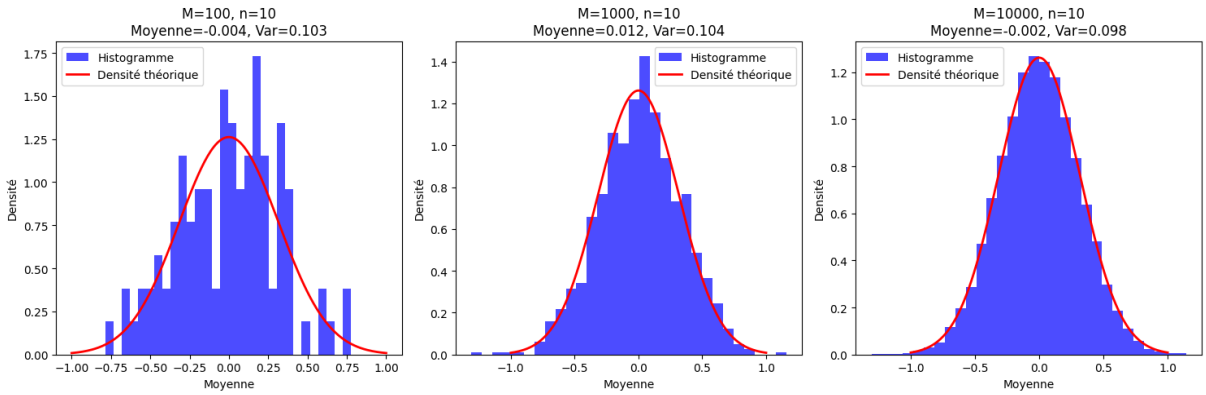

Figure 1 – *Histogrammes des moyennes obtenues pour $n = 10$ échantillons suivant une loi normale centrée réduite*

Les résultats montrent que, pour un petit nombre de répétitions ($M = 100$), l'estimation

de la distribution de la moyenne est assez approximative, avec des fluctuations visibles. En augmentant M, la distribution empirique des moyennes converge rapidement vers la densité théorique, ce qui illustre la loi des grands nombres et la convergence en loi de l'estimateur de la moyenne.

Si l'on augmente la taille n de chaque échantillon, la variance de l'estimateur $\overline{X}$ diminue proportionnellement à $1/n$. Autrement dit, plus n est grand, plus l'estimation de la moyenne devient précise (moins variable), ce qui améliore significativement la fiabilité de l'estimateur.

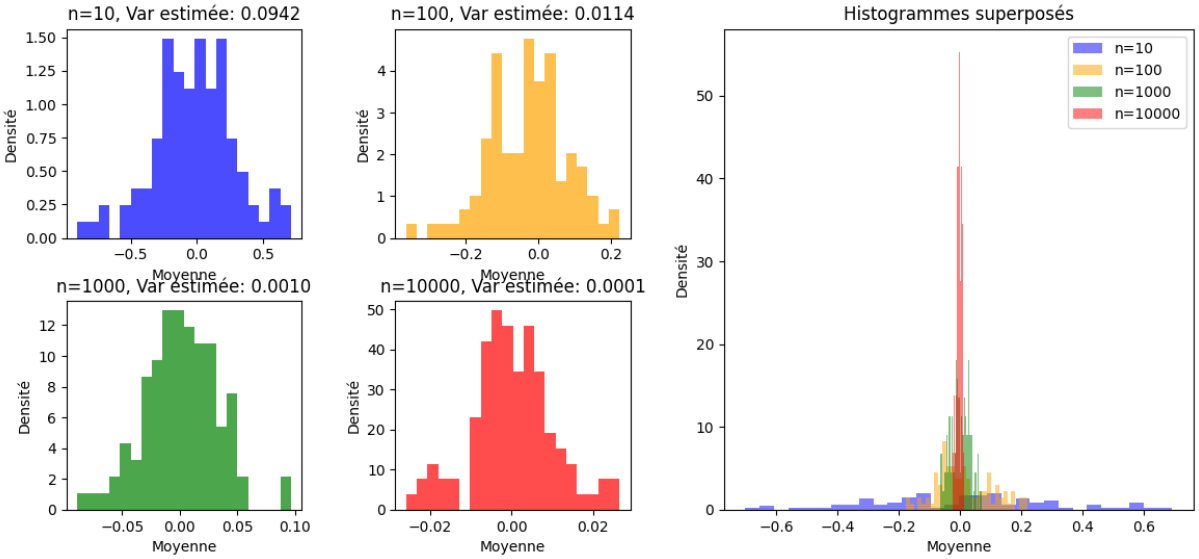

Figure 2 – *Histogrammes des moyennes obtenues pour n échantillons variant de 10 à 10 000*

Ainsi, on constate que pour un nombre d'échantillonnages M élevé, la moyenne empirique des estimateurs tend vers zéro, conforme à la valeur théorique. La variance empirique des moyennes diminue fortement lorsque la taille de l'échantillon n augmente, ce qui confirme que la précision de l'estimation s'améliore avec la taille de l'échantillon.
En effet, lorsque $n = 10$, la variance est autour de 0,1, tandis qu'elle chute à 10^{-4} pour $n = 10000$. Cette décroissance de la variance est en accord avec la relation théorique :

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n},$$

Par conséquent, augmenter la taille de l'échantillon réduit significativement l'incertitude sur l'estimation de la moyenne, renforçant la confiance dans la précision de cette estimation. À ce stade, il est donc naturel de se demander : mon estimation ponctuelle de la moyenne est-elle suffisamment fiable ? Il s'agit donc de déterminer un intervalle de confiance, qui encadre la moyenne avec un certain degré de certitude.

1.2 Intervalle de confiance (cf. 4.1)

Intervalle de confiance

On souhaite estimer un paramètre inconnu θ à partir d'un échantillon de n observations. Soit $\alpha \in]0,1[$ le niveau de risque. S'il existe deux variables aléatoires $\theta_{\min}(X_1, ..., X_n)$ et $\theta_{\max}(X_1, ..., X_n)$ telles que :

$$P\left(\theta \in [\theta_{\min}, \theta_{\max}]\right) = 1 - \alpha,$$

Alors l'intervalle $[\theta_{\min}, \theta_{\max}]$ est appelé intervalle de confiance pour θ, de niveau de confiance $1 - \alpha$, noté $IC_{1-\alpha}(\theta)$. Dit autrement, si l'on répétait de nombreux échantillonnages, la proportion des cas où l'intervalle contiendrait la vraie valeur θ serait proche de $1 - \alpha$.

1.2.1 Cas où la variance de la population est connue

Dans notre cas, on a $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ la moyenne empirique d'un échantillon de n variables aléatoires de moyenne μ et d'écart-type σ. Le théorème central limite donne :

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \longrightarrow \mathcal{N}(0,1)$$

On cherche alors les bornes $z_{\alpha/2}$ telles que :

$$P\left(-z_{1-\alpha/2} < Z_n < z_{1-\alpha/2}\right) = 1 - \alpha$$

Ce qui permet de construire l'intervalle de confiance suivant pour μ :

$$IC_{1-\alpha}(\mu) = \left[\bar{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Pour un intervalle de confiance de 95 % (*i.e.* un niveau de risque $\alpha = 0{,}05$), le quantile critique inférieur vaut $z_{1-\alpha/2} \approx 1{,}96$. Cela permet de définir l'intervalle de confiance à $1 - \alpha$ près autour de la moyenne empirique. Une fois l'intervalle de confiance $IC_{1-\alpha}(Z_n)$ établi sur la variable normalisée, on revient à la moyenne empirique $\overline{X}_n$ pour obtenir :

$$IC_{1-\alpha}(\mu) = \left[\overline{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\ ;\ \overline{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

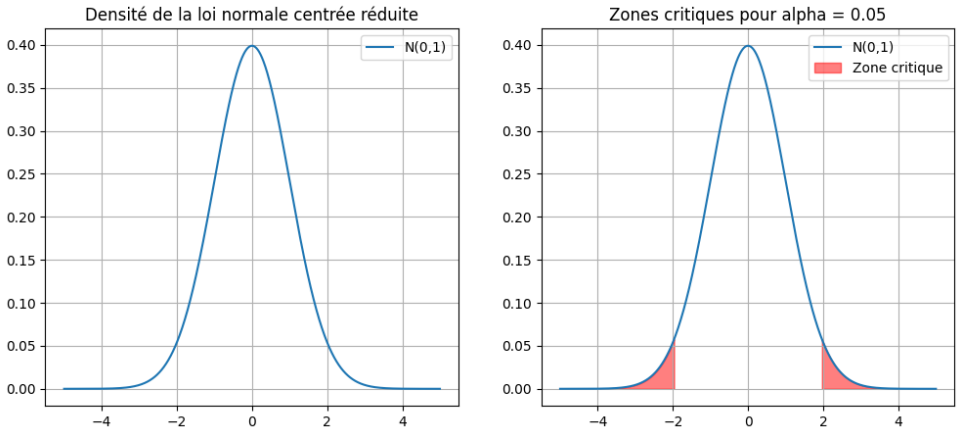

Figure 3 – *Visualisation des zones critiques pour une loi normale centrée réduite*

L'intervalle de confiance dépend donc de trois paramètres : le niveau de risque α, la taille de l'échantillon n, et l'écart-type σ. Lorsque la taille de l'échantillon n augmente, l'intervalle se resserre naturellement autour de la valeur estimée, ce qui améliore la précision de l'estimation. À l'inverse, un écart-type σ élevé reflète une variabilité importante du phénomène étudié, rendant plus difficile l'obtention d'une estimation fiable.

Dans notre cas, l'estimation réalisée sur un échantillon de $n = 30$ valeurs a donné une moyenne $\mu = 0{,}0981$ et un écart-type $\sigma = 1{,}0256$. Sur 10 000 répétitions du procédé d'échantillonnage, le taux d'erreur empirique observé, c'est-à-dire la fréquence à laquelle la vraie moyenne n'est pas contenue dans l'intervalle de confiance, est de **6 %**, légèrement supérieur au niveau de risque théorique de 5 %. Ce décalage illustre bien la nature probabiliste de l'intervalle de confiance, qui garantit une couverture de la vraie moyenne dans la majorité des cas, mais pas systématiquement pour un seul échantillon.

1.2.2 Cas où la variance est inconnue

Jusqu'à présent, la construction de l'intervalle de confiance reposait sur l'hypothèse que la variance σ^2 de la population était connue. Or, dans la plupart des situations réelles, cette information n'est pas accessible, et il faut l'estimer à partir de l'échantillon. Cette estimation introduit une incertitude supplémentaire qui modifie la loi suivie par la statistique normalisée. En effet, lorsque la variance est inconnue, la variable $T = \frac{\overline{X}_n - \mu}{S/\sqrt{n}}$ ne suit plus une loi normale centrée réduite, mais une loi de Student à $n-1$ degrés de liberté, où S est l'écart-type corrigé de l'échantillon calculé avec le facteur de Bessel $(n-1)$. Ainsi, l'intervalle de confiance à un niveau de confiance $1-\alpha$ pour la moyenne μ s'écrit :

$$IC_{1-\alpha}(\mu) = \left[\overline{X}_n - t_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} \ ; \ \overline{X}_n + t_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right]$$

où : $t_{1-\frac{\alpha}{2},n-1}$ est la valeur critique de la loi de Student à $n-1$ degrés de liberté.

Dans notre étude, avec un échantillon de taille $n = 10$ et un niveau de risque $\alpha = 0{,}05$, on calcule la valeur critique : $t_{1-\frac{\alpha}{2},9} \approx 2{,}2622$. En réalisant $T = 10\,000$ répétitions de l'échantillonnage et du calcul d'intervalle, le taux d'erreur empirique observé est de : 5,15%. Ce résultat confirme que l'utilisation de la loi de Student permet d'obtenir un intervalle de confiance adapté à l'incertitude induite par l'estimation de la variance, respectant ainsi le niveau de risque théorique. On observe que ce taux d'erreur est plus proche du niveau nominal α que celui obtenu avec la loi normale lorsque la variance est inconnue, soulignant l'importance de ce choix dans les analyses statistiques.

2 Analyse statistique de données

Dans son article fondateur sur la régression, Francis Galton a étudié les tailles moyennes de 205 couples ainsi que celles de leurs enfants. Nous nous concentrons ici uniquement sur les tailles des parents. L'objectif de cette partie est d'analyser ces données, en estimant notamment les caractéristiques statistiques telles que la moyenne et la variance des tailles, et d'étudier la fiabilité de ces estimations via des intervalles de confiance adaptés. Cette démarche permettra de mieux comprendre la variabilité des tailles parentales et d'évaluer la précision des mesures obtenues à partir d'échantillons.

2.1 Statistiques descriptives des tailles parentales (cf. 4.2)

Définitions des mesures statistiques

Soit un échantillon $x_1, x_2, \ldots, x_n \in \mathbb{R}$, on note :

• **Variance biaisée et corrigée** :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \text{et} \qquad \hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

• **Coefficient d'asymétrie (skewness)** :

$$\text{Skewness biaisée} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3$$

La version corrigée utilise des estimateurs corrigés appelés *k-statistics*, qui tiennent compte de la taille de l'échantillon.

• **Coefficient d'aplatissement (kurtosis)** :

$$\text{Kurtosis biaisée} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 - 3$$

Ici aussi, la version corrigée est obtenue via les *k-statistics*. Elle fournit une meilleure estimation de la kurtosis de la population, surtout pour les petits échantillons. Le -3 permet de centrer la valeur à 0 pour une loi normale (définition de Fisher).

Nous avons extrait et analysé les tailles des pères et des mères à partir des données fournies. Voici les principales statistiques descriptives obtenues :

— **Tailles moyennes** : La moyenne des tailles des pères est de **69,32** pouces, tandis que celle des mères est de **64,00** pouces.

— **Variances et écarts-types** : L'estimation non biaisée de la variance des tailles des pères est de **7,01**, avec un écart-type de **2,65**. Pour les mères, la variance non biaisée vaut **5,44**, et l'écart-type associé est **2,33**.

— **Asymétrie (skewness)** : Les deux distributions sont globalement symétriques : l'asymétrie est proche de zéro pour les pères (**0,01**) et légèrement négative pour les mères (**-0,18**).

— **Aplatissement (kurtosis)** : L'aplatissement des tailles des pères (**0,41**) suggère une distribution légèrement plus "pointue" qu'une loi normale, tandis que celui des mères (**-0,02**) est très proche de celui de la loi normale.

Statistique	**Pères**	**Mères**
Moyenne	69,32	64,00
Variance (non biaisée)	7,01	5,44
Écart-type (non biaisé)	2,65	2,33
Variance (biaisée)	6,97	5,42
Écart-type (biaisé)	2,64	2,33
Asymétrie (non biaisée)	0,01	-0,18
Aplatissement (non biaisé)	0,41	-0,02

Table 1 – *Récapitulatif des statistiques descriptives*

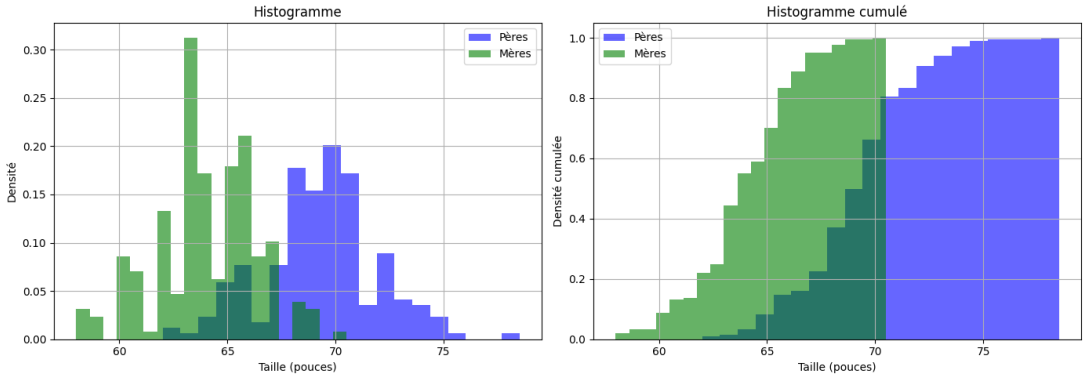

Figure 4 – *Histogrammes et histogrammes cumulés des tailles des pères et des mères*

Nous pouvons visualiser les données grâce à une boîte à moustaches (ou *boxplot*). Elle met en évidence :

— la médiane (trait central),

— les premier et troisième quartiles (bords de la boîte),

— les valeurs extrêmes dites « moustaches »,

— les éventuels points atypiques (outliers).

Cette visualisation permet de comparer la dispersion et la symétrie des distributions de taille chez les pères et les mères. Par exemple, une boîte plus étroite indique une variabilité plus faible, et un décalage de la médiane signale une asymétrie.

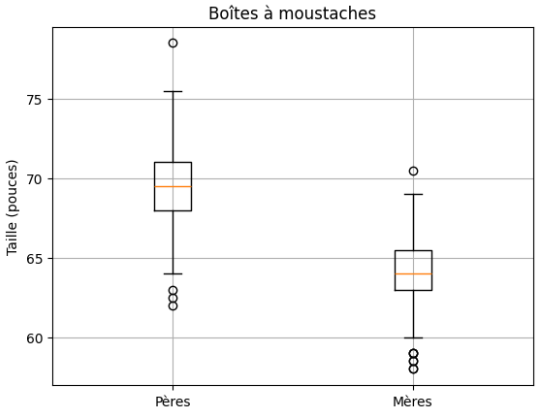

Figure 5 – *Boîtes à moustaches des tailles des pères et des mères*

2.2 Tests de conformité (cf. 4.2)

2.2.1 Test du χ^2 : Analyse de la normalité

Test du χ^2 d'adéquation

Le test du χ^2 d'adéquation (*chi-square goodness-of-fit test*) permet de vérifier si un échantillon peut raisonnablement être considéré comme issu d'une distribution théorique donnée, ici la loi normale. Il consiste à regrouper les observations en classes (généralement définies à partir de l'histogramme) et à comparer les effectifs observés O_i aux effectifs attendus E_i sous l'hypothèse de distribution choisie. La statistique de test est donnée par :

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

Elle suit une loi du χ^2 asymptotique lorsque les effectifs sont suffisants, généralement lorsque $E_i \geq 5$ pour tout i.

Pour évaluer la normalité des distributions des tailles des pères et des mères, nous avons implémenté un test du χ^2 de conformité, adapté pour prendre en compte l'exigence $E_i \geq 5$ dans chaque classe. Ce test permet de comparer la distribution empirique à une loi normale dont les paramètres (moyenne et écart-type) sont estimés à partir des données. La méthode consiste à :

- Estimer la moyenne et l'écart-type de l'échantillon.
- Définir des intervalles (ou *bins*) de largeur constante.
- Calculer les effectifs observés O_i dans chaque bin à l'aide d'un histogramme.
- Calculer les effectifs attendus E_i dans chaque bin à partir de la fonction de répartition d'une loi normale : $E_i = N \cdot \mathbb{P}(X \in [a_i, b_i]) = N \cdot (F(b_i) - F(a_i))$ avec F la fonction de répartition de la loi normale estimée.
- Fusionner les classes aux extrémités si certains $E_i < 5$ pour garantir la validité asymptotique du test.
- Calculer la statistique de test avec le nombre final de bins après fusion.
- Calculer le nombre de degrés de liberté : ddl $= r - 1 - m$, où $m = 2$ (moyenne et écart-type estimés).
- En déduire la *p-value* et conclure.

Cette méthode a été appliquée aux données de tailles des pères et des mères. Voici les résultats :

- **Pères** : $\chi^2 = 7.39$, ddl $= 4$, *p-value* $= 0.1165$. Ainsi, on ne rejette pas l'hypothèse nulle : les données peuvent être considérées comme issues d'une loi normale.
- **Mères** : $\chi^2 = 25.89$, ddl $= 5$, *p-value* $= 0.0001$. Ainsi, on rejette l'hypothèse nulle : la distribution n'est pas gaussienne.

Le code implémente également une fonction `merge_bins` fusionnant les classes aux extrémités lorsque les effectifs attendus sont trop faibles, ce qui garantit une meilleure fiabilité du test statistique. Cela permet de représenter graphiquement, pour chaque classe finale, les effectifs observés et attendus :

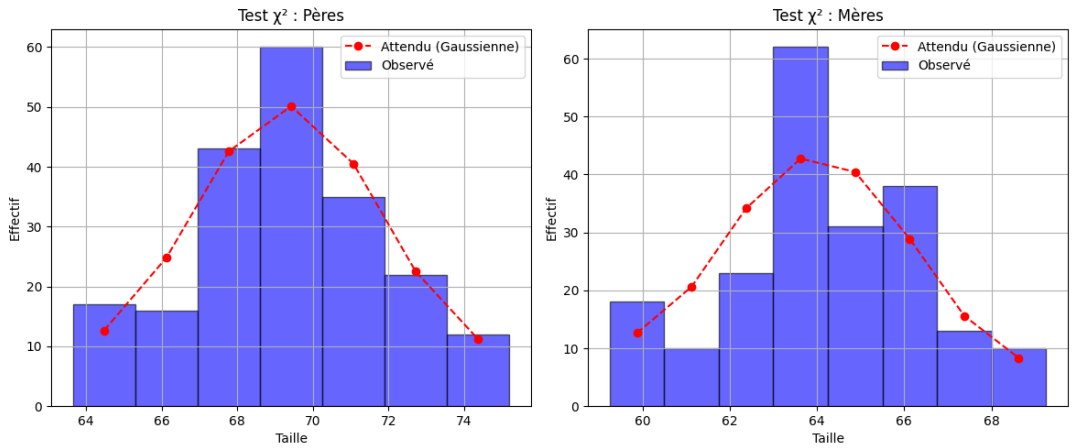

Figure 6 – *Histogrammes des tailles pour le test du χ^2 avec comparaison des O et E*

2.2.2 Graphique Quantile-Quantile et droite de Henri

La droite de Henri est la droite théorique sur laquelle les points devraient s'aligner si les données suivaient parfaitement une loi normale. Elle est obtenue en traçant les quantiles théoriques en abscisse et les quantiles empiriques en ordonnée. Une bonne superposition entre les points et la droite valide l'hypothèse de normalité. Le graphique Quantile-Quantile consiste à :

— Trier les données $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

— Calculer les quantiles théoriques de la loi normale standard associés à chaque rang, typiquement via $z_i = \Phi^{-1}\left(\frac{i-0,5}{n}\right)$ où Φ^{-1} est la fonction quantile de la loi normale.

— Tracer les couples $(z_i, x_{(i)})$.

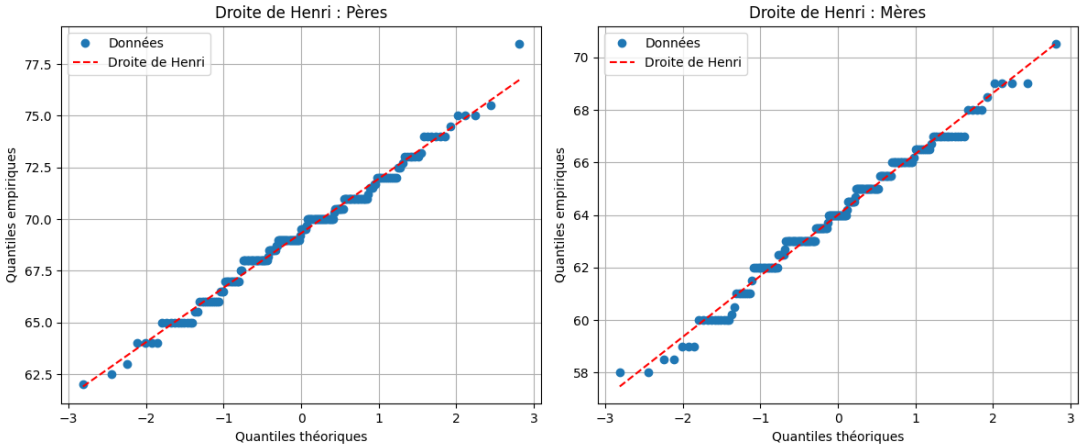

Figure 7 – *Graphiques Quantile-Quantile et droite de Henri pour les deux populations*

On observe que pour les pères, les points suivent globalement la droite de Henri : l'ajustement à une loi normale est visuellement satisfaisant. Tandis que pour les mères, des écarts importants apparaissent aux extrémités du nuage de points, révélant une queue plus lourde ou un aplatissement — confirmant le résultat du test χ^2 : la distribution n'est pas gaussienne.

2.2.3 Test de Levene, de Student et de Welch

Test de Levene

Le test de Levene permet de vérifier l'égalité des variances entre deux populations, sans supposer que les données suivent une loi normale stricte. Il est donc plus robuste que le test de Fisher face aux écarts à la normalité. On souhaite tester l'hypothèse nulle : $H_0 : \sigma_1^2 = \sigma_2^2$. Le test s'appuie sur la transformation des données en considérant les écarts absolus des observations par rapport à une mesure de tendance centrale (médiane ou moyenne) dans chaque groupe. Plus précisément, pour chaque observation X_{ij} du groupe j ($j = 1, 2$), on calcule :

$$Z_{ij} = |X_{ij} - \tilde{X}_j|$$

où : $\tilde{X}_j$ est la médiane du groupe j. Le test consiste alors à effectuer une analyse de variance sur les valeurs Z_{ij} afin de déterminer si les dispersions des deux groupes diffèrent significativement. On rejette l'hypothèse nulle au niveau α si la statistique de test W est supérieure au quantile d'ordre $1 - \alpha$ de la loi F correspondante, ou si elle est inférieure au quantile d'ordre α (cas bilatéral).

Ces intervalles ne se recouvrent pas, ce qui suggère déjà une différence notable entre les moyennes. Nous avons utilisé le test de Levene, plus robuste que le test F classique, afin de tester l'égalité des variances sans supposer a priori que les données sont parfaitement normales. Celui-ci donne une statistique $F = 1{,}918$ avec une p-value de 0,1668. Comme cette p-value est supérieure au seuil de 0,05, on ne rejette pas l'hypothèse nulle d'égalité des variances, ce qui justifie l'utilisation d'un test t classique (Student) avec variances supposées égales.

Test de Student : Variances égales

Le test de Student permet de vérifier si deux populations ont la même espérance, sous les hypothèses que ces populations sont normalement distribuées et que leurs variances sont égales. On souhaite tester l'hypothèse nulle $H_0 : \mu = \mu_0$. La statistique de test est telle que :

$$Z = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S^*/\sqrt{n}}$$

Avec :

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

Pour un risque d'erreur de première espèce α, on compare la statistique Z aux quantiles de la loi de Student à $n - 1$ degrés de liberté :

- Si Z est compris entre les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$, alors on ne rejette pas $H_0 : \mu = \mu_0$.
- Si Z est inférieur au quantile d'ordre $\alpha/2$, alors on rejette H_0 et on conclut $\mu < \mu_0$.
- Si Z est supérieur au quantile d'ordre $1-\alpha/2$, alors on rejette H_0 et on conclut $\mu > \mu_0$.

Le test t de Student pour la comparaison des moyennes renvoie une statistique $t = 21{,}563$ avec une p-value proche de zéro. Cela conduit à rejeter l'hypothèse nulle d'égalité des moyennes au seuil de 5 %, indiquant que la taille moyenne des pères est significativement différente de celle des mères. Ainsi, malgré des variances comparables entre les deux populations, les tailles moyennes des pères et des mères diffèrent de manière statistiquement significative.

Dans le cas où les deux variances auraient été inégales, nous aurions effectué un test de Welch décrit ci-dessous.

Test de Welch : Variances inégales

Le test de Welch permet de vérifier si deux populations ont la même espérance, sous l'hypothèse que ces populations sont normalement distribuées. Il est particulièrement utilisé lorsque les variances sont inégales. On souhaite tester l'hypothèse nulle $H_0 : \mu_1 = \mu_2$. La statistique de test est telle que :

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Si H_0 est vraie, Z suit une loi de Student à ν degrés de liberté calculés par la formule de Welch-Satterthwaite :

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}}.$$

Une fois Z et ν calculés, on compare Z aux quantiles de la loi de Student à ν degrés de liberté pour décider d'accepter ou de rejeter H_0.

3 Hypercholestérolémie primaire en monothérapie (cf. 4.2)

L'étude porte sur 827 patients traités à l'hypocholestérolémiant et 892 patients sous placebo. Pour chacun, on a mesuré la variation du taux de cholestérol LDL entre le début et la fin du traitement.

	Traitement	**Placebo**
Femmes	428	502
Hommes	399	390
25–44 ans	162	210
45–64 ans	403	423
> 65 ans	262	259
$\sum \delta_k$ (g/L)	$-528{,}4477$	$-116{,}0837$
$\sum(\delta_k - \bar{\delta})^2$	76,1925	35,8899

Table 2 – *Résultats de l'étude clinique*

3.1 Équilibre des groupes selon le sexe et l'âge

Afin de mieux visualiser la répartition des effectifs selon le sexe et les tranches d'âge dans les groupes traitement et placebo, nous avons tracé les histogrammes ci-dessous.

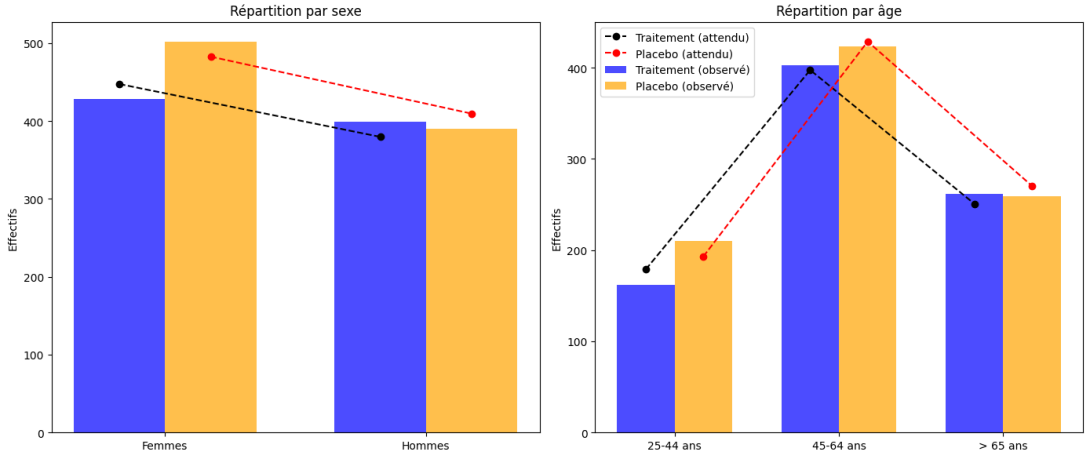

Figure 8 – *Comparaison des effectifs dans les groupes traitement et placebo*

On observe que la distribution des effectifs est globalement équilibrée entre les deux groupes, tant pour la variable sexe (femmes : 428 vs 502, hommes : 399 vs 390) que pour les tranches d'âge. Ceci est important pour garantir la comparabilité des groupes avant d'analyser l'effet du traitement. Pour cela, nous effectuons deux tests du χ^2 d'indépendance au seuil $\alpha = 5\%$.

3.1.1 Test du χ^2 d'indépendance pour le sexe

Hypothèses :

- Hypothèses non remises en question : taille d'échantillon suffisante, observations indépendantes.
- H_0 : la répartition par sexe est indépendante du groupe (traitement ou placebo).

— H_1 : la répartition par sexe dépend du groupe.

Résultat du test : $\chi^2 = 3.358$, ddl $= 1$, $p = 0.0669$
Conclusion : Au seuil de 5%, on ne rejette pas H_0. Il n'y a pas de différence significative de répartition entre les sexes dans les deux groupes.

Effectifs attendus sous H_0 : $\begin{pmatrix} 447.42 & 482.58 \\ 379.58 & 409.42 \end{pmatrix}$ (cf. Figure 8).

3.1.2 Test du χ^2 d'indépendance pour l'âge

Hypothèses :

— Hypothèses non remises en question : tranches d'âge exhaustives et non chevauchantes, effectifs suffisants.

— H_0 : la répartition par tranche d'âge est indépendante du groupe.

— H_1 : la répartition par tranche d'âge dépend du groupe.

Résultat du test : $\chi^2 = 4.243$, ddl $= 2$, $p = 0.1198$
Conclusion : Au seuil de 5%, on ne rejette pas H_0. Les groupes sont comparables en ce qui concerne la distribution par âge.

Table des effectifs attendus sous H_0 : $\begin{pmatrix} 178.97 & 193.03 \\ 397.38 & 428.62 \\ 250.65 & 270.35 \end{pmatrix}$ (cf. Figure 8).

Ainsi, les deux groupes sont statistiquement équilibrés selon le sexe et l'âge, ce qui permet une comparaison fiable de l'effet du traitement.

3.2 Analyse des variations du taux de cholestérol LDL

Les variations du taux de cholestérol LDL ont été analysées à l'aide de boîtes à moustaches afin de visualiser la distribution, la médiane, ainsi que l'étendue des données.

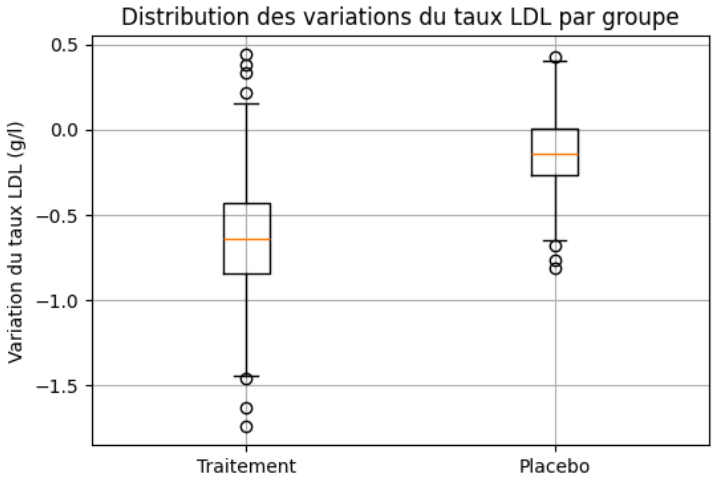

Figure 9 – *Boîtes à moustaches des variations du taux LDL pour les deux groupes d'étude*

Résumé des estimations statistiques ponctuelles obtenues :

Groupe	Espérance	Variance	Écart-type
Traitement	-0.6390 g/L	0.0922	0.3037 g/L
Placebo	-0.1301 g/L	0.0403	0.2007 g/L

Table 3 – *Estimation ponctuelle des paramètres statistiques de la variation du taux de cholestérol LDL*

3.3 Efficacité de l'hypocholestérolémiant

Pour déterminer si le traitement hypocholestérolémiant est plus efficace que le placebo dans la diminution du taux de LDL, nous avons comparé les moyennes des variations observées dans deux groupes indépendants. Soient μ_1 et μ_2 les espérances des variations de LDL dans les groupes traitement et placebo respectivement. Nous testons :

$$\begin{cases} H_0 : \mu_1 - \mu_2 \geq 0 & \text{(le traitement n'est pas plus efficace)} \\ H_1 : \mu_1 - \mu_2 < 0 & \text{(le traitement est plus efficace)} \end{cases}$$

Ce test est un test unilatéral à gauche sur la différence des moyennes. La statistique de test utilisée est :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Où : $\bar{x}_i$ sont les moyennes observées, s_i^2 les variances échantillonnales, et n_i les tailles d'échantillon respectives. Le seuil de signification est fixé à $\alpha = 5\%$. Le rejet de l'hypothèse nulle H_0 se fait si la p-value du test est inférieure à ce seuil, ce qui indique une différence significative en faveur du traitement.

À l'issue du test, les résultats suivants ont été obtenus :

— Les moyennes des diminutions sont calculées par :

$$\bar{x}_1 = \frac{S_1}{n_1} = -0{,}6389 \text{ g/L}, \quad \bar{x}_2 = \frac{S_2}{n_2} = -0{,}1301 \text{ g/L}$$

— La différence ponctuelle des moyennes est :

$$\hat{\delta} = \bar{x}_1 - \bar{x}_2 = -0{,}5089 \text{ g/L}$$

— Les variances sont estimées par :

$$s_1^2 = \frac{SC_1}{n_1 - 1} = 0{,}0922, \quad s_2^2 = \frac{SC_2}{n_2 - 1} = 0{,}0403$$

— L'erreur standard de la différence est calculée par :

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0{,}0221$$

— Nous construisons un intervalle de confiance à 95 % sous l'hypothèse d'une distribution normale approchée (taille d'échantillons importante), avec la statistique critique $z_{0,975} = 1{,}96$:

$$IC_{95\%} = \hat{\delta} \pm z_{0,975} \times SE = [-0{,}5334; -0{,}4843] \text{ g/L}$$

— La statistique de test pour vérifier si le traitement est plus efficace que le placebo (test unilatéral gauche) est :

$$Z = \frac{\hat{\delta}}{SE} = -40{,}6504$$

— La p-value associée est :

$$p = \Phi(Z) \approx 0$$

L'intervalle de confiance étant strictement négatif, et la p-value étant largement inférieure au seuil de 5 %, on rejette l'hypothèse nulle H_0.

Conclusion : Le test est hautement significatif, ce qui confirme que le traitement hypocholestérolémiant est statistiquement plus efficace que le placebo dans la diminution du taux de LDL chez les patients étudiés.