PROJET - SCIENCE DES DONNÉES

Par Rayan CHENNAOUI - Aéro 3PD2

1 Introduction

December 14, 2025

Le but de ce mini-projet est d'appliquer des méthodes statistiques sur un ensemble de données réelles afin de mieux comprendre les concepts d'analyse univariée et bivariée. Cette analyse commence par l'étude de séries statistiques simples, avec des indicateurs comme la médiane et l'écart-type. La suite portera sur les performances des avions pour évaluer leur efficacité et leur rentabilité, avant d'aborder l'estimation ponctuelle et les intervalles de confiance pour une analyse statistique rigoureuse.

2 Régime alimentaire des vaches

2.1 Analyse univariée

L'analyse univariée a pour but de décrire et mesurer la répartition des valeurs que peut prendre une variable. La population étudiée est de 15 individus. Dans notre cas, les variables sont continues. Voici la Table 1 récapitulative :

Statistique	**Série X**	**Série Y**
Minimum	23.4	25.6
Maximum	32.4	32.0
Étendue	9.0	6.4
Moyenne	28.04	28.91
Médiane	28.2	28.8
Quartile 25% (Q1)	27.0	28.0
Quartile 75% (Q3)	29.4	29.6
Écart Interquartile	2.4	1.6
Variance	4.53	2.91
Ecart-type	2.13	1.71

Table 1: Résultats des calculs statistiques pour les séries X et Y.

Les calculs effectués sur Python montrent que la variance de X (4.53) indique une dispersion modérée, tandis que celle de Y (2.91) est plus faible, montrant une production homogène avec l'aliment Y. L'écart-type confirme cette observation.

La médiane de l'aliment X est légèrement inférieure à celle de l'aliment Y, indiquant que les vaches nourries avec Y produisent en moyenne plus de lait. Le boxplot montre une plus grande variabilité pour X, confirmant une dispersion supérieure à celle de Y, dont la production est plus stable. En conclusion, l'aliment Y favorise une production régulière et moins sujette aux variations extrêmes.

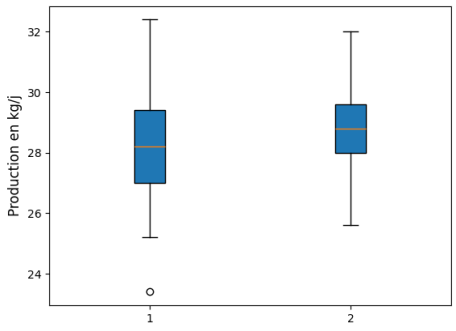


Figure 1: Boxplots pour les séries X et Y.

Le premier histogramme montre une plus grande variabilité dans X, tandis que le second illustre une production plus stable, et meilleure, avec Y :

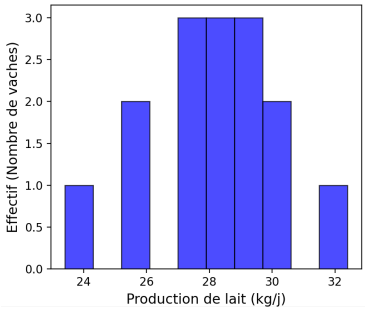


Figure 2: Production de lait de la série X

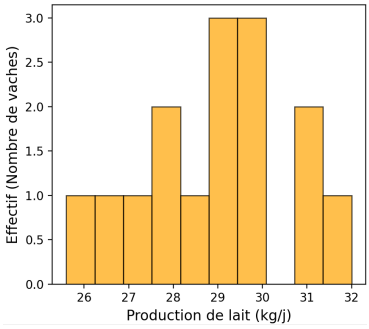


Figure 3: Production de lait de la série Y

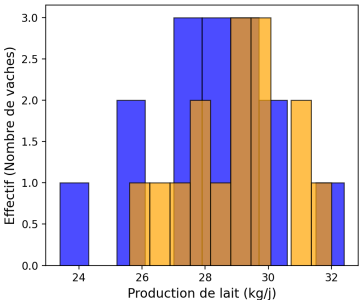


Figure 4: Comparaison visuelle des deux séries

2.2 Analyse bivariée

Cette partie explore les relations entre différentes variables clés de l'aviation, en combinant calculs statistiques, représentations graphiques et interprétations. Le coefficient de corrélation linéaire entre les 2 aliments est r=0.81, ce qui indique une bonne corrélation linéaire (voir Figure 7). Cela signifie que les productions de lait des vaches nourries avec l'aliment X et celles nourries avec l'aliment Y sont étroitement liées : lorsque la production d'une série augmente, celle de l'autre série tend également à augmenter de manière similaire. On peut également déterminer la covariance entre X et Y : cov(X,Y) = 2.93.

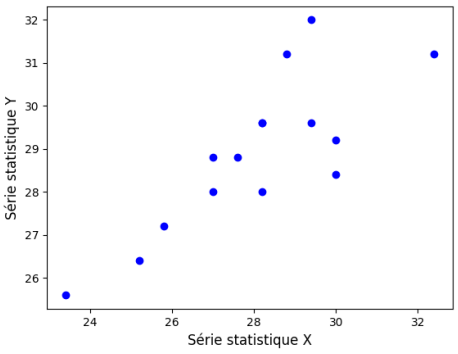


Figure 5: Nuage de points : Y en fonction de X, sans point moyen.

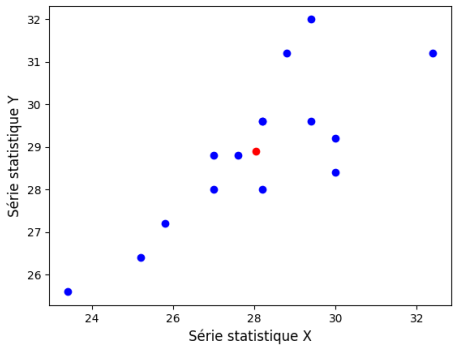
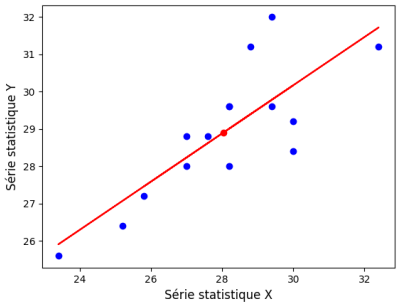
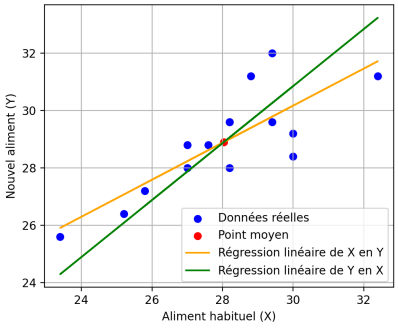


Figure 6: Nuage de points : Y en fonction de X, avec le point moyen.

Il est possible de trouver les coefficients avec plusieurs méthodes : la première méthode en utilisant la covariance, la variance et les valeurs moyennes, on trouve assez aisément a et b tels que : $a = \frac{\text{cov}(X,Y)}{\text{Var}(X)}$ et $b = \overline{Y} - a \cdot \overline{X}$. Puis, on peut également faciliter les calculs en utiliser numpy et l'outil polyfit. Enfin, on peut utiliser la méthode des moindres carrés (cf. Partiel 15/01/2025 Ex2, 2.a) : $\begin{pmatrix} a \\ b \end{pmatrix} = \left(X^\top X\right)^{-1} X^\top Y$. Dans les 3 cas, on trouve $\mathbf{y = ax + b = 0.65x + 10.79}$



(a) Droite de régression linéaire de Y en X



(b) Droites de régression linéaire (Y en X et X en Y)

Figure 7: Comparaison des droites de régression linéaire

On note que la droite de régression de Y en X croît plus vite. Puis, à l'aide de la méthode des moindres carrés, on peut déterminer les valeurs ajustées $\hat{Y}_i$ ainsi que les distances H_i de chaque point par rapport à la droite d'ajustement :

Point	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\hat{Y}_i$	28.62	25.91	27.07	29.01	29.40	27.46	28.23	28.23	29.79	29.01	30.17	29.01	31.72	29.79	30.17
Distances	0.18	0.31	0.67	1.01	1.80	0.26	0.57	0.23	0.19	0.59	1.77	0.59	0.52	2.21	0.97

Table 2: Valeurs ajustées et distances par rapport à la droite de régression (horizontal), pour Y

Puis, les variances expliquées et résiduelles pour l'aliment Y sont : $V(Y)_e = 1.89$ et $V(Y)_r = 1.02$. Et pour l'aliment X : $V(X)_e = 2.95$ et $V(X)_r = 1.59$.

Point	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\hat{X}_i$	27.93	24.71	25.52	27.13	30.35	26.32	27.93	27.13	28.74	28.74	27.53	28.74	30.35	31.15	28.34
Distances	0.33	1.31	0.32	1.07	1.55	0.52	0.93	0.13	0.66	0.54	2.47	0.54	2.05	1.75	1.66

Table 3: Valeurs ajustées et distances par rapport à la droite de régression (horizontal), pour X

On remarque que la variance résiduelle est plus élevée dans le premier modèle, or *"plus la variance résiduelle d'un modèle est élevée, moins le modèle est capable d'expliquer la variation des données"* [1], ainsi nous pouvons en déduire que le deuxième modèle est alors le plus précis. Voici 2 estimations :

- La prédiction pour Y lorsque $X = 25$ est $Y = 27.04$. Lorsque la production avec l'aliment habituel est de 25kg, la production avec le nouvel aliment est de 27.04.
- La prédiction pour X lorsque $Y = 30$ est $X = 29.26$. Lorsque la production avec le nouvel aliment est de 30kg, la production avec l'aliment habituel est de 29.26.

3 Performance des avions

Les performances des avions, telles que la vitesse de croisière, la consommation de carburant et la portée maximale, sont essentielles pour évaluer leur efficacité. Cette section analyse ces données à travers trois approches statistiques : univariée, bivariée et multidimensionnelle, pour mieux comprendre les capacités des différents modèles.

3.1 Analyse univariée

L'analyse univariée a pour but de décrire et mesurer la répartition des valeurs que peut prendre une variable. La population étudiée est de 10000 avions. Dans notre cas, les variables sont continues.

Statistique	**Cruise speed (km/h)**	**Fuel consumption (kg/h)**	**Max range (km)**
Moyenne	849.79	2500.87	5006.52
Médiane	850.24	2498.53	5005
Écart-type	50.58	298.24	698.62
Minimale	646.77	1244.16	2388
Maximale	1038.87	3789.66	7429

Table 4: Paramètres statistiques des performances des avions

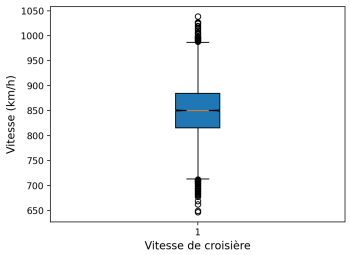


Figure 8: Boxplot vitesse de croisière

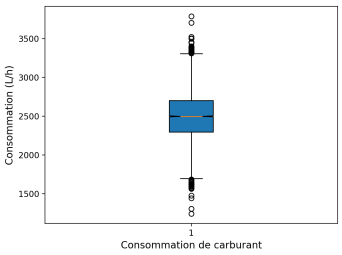


Figure 9: Boxplot consommation carburant

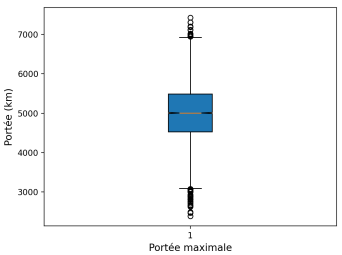


Figure 10: Boxplot portée maximale

[1] https://www.statology.org/residual-variance/

La vitesse moyenne des avions est de **849.79 km/h**, avec une médiane proche (**850.24 km/h**), indiquant une distribution symétrique. L'écart-type de **50.58 km/h** reflète une variabilité modérée, tandis que les valeurs extrêmes (**646.77 km/h** et **1038.87 km/h**) pourraient représenter des modèles atypiques (voir Table 4).

Avec une moyenne de **2500.87 kg/h** et une médiane similaire, la consommation de carburant est également symétrique. L'étendue significative, de **1244.16 kg/h** à **3789.66 kg/h**, met en évidence des différences notables entre les modèles, malgré une variabilité globalement modérée (écart-type de **298.24 kg/h**) (voir Table 4).

La portée moyenne est de **5006.52 km**, avec une médiane proche et un écart-type plus marqué (**698.62 km**), témoignant d'une variabilité notable. Les valeurs extrêmes (**2388** et **7429 km**) indiquent des écarts importants dans les capacités des différents modèles d'avions (voir Table 4).

Afin d'identifier la variable avec la plus grande variabilité, on peut utiliser l'écart-type en valeur relative, dont l'explication est détaillée dans la Table 5.

Variable	**Écart-type**	**Moyenne**	**Variabilité Relative (%)**
Vitesse croisière	50.58	849.79	$\frac{50.58}{849.79} \times 100 \approx 5.95\%$
Consommation de carburant	298.24	2500.87	$\frac{298.24}{2500.87} \times 100 \approx 11.93\%$
Portée maximale	698.62	5006.52	$\frac{698.62}{5006.52} \times 100 \approx 13.95\%$

Table 5: Variabilité relative des variables

La **portée maximale** présente la plus grande variabilité relative (**13.95 %**), ce qui peut refléter des écarts importants entre les modèles d'avions. Cette variabilité est essentielle pour les **stratégies commerciales** (ciblage de marchés), l'**optimisation de la flotte** (harmonisation des capacités) et le **positionnement sur le marché**.

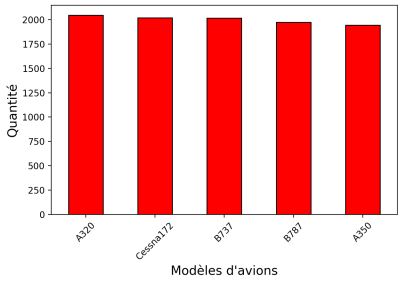


Figure 11: Répartition des modèles d'avions

La répartition des modèles est homogène, avec une légère sur-représentation de l'**Airbus A320** (voir Figure 11). Bien que ce modèle soit légèrement plus fréquent, son influence sur l'analyse demeure marginale. Cette répartition permet d'examiner les caractéristiques globales des avions, comme ses **performances**, mais le déséquilibre doit être pris en compte pour des analyses plus spécifiques. Une analyse séparée de chaque modèle serait pertinente pour garantir une neutralité .

3.2 Analyse bivariée

Le coefficient de corrélation de Pearson de **-0.01** (voir Figure 12) indique une absence de relation linéaire entre la vitesse de croisière et la consommation de carburant. Cela suggère que d'autres facteurs, comme l'aérodynamique, la charge ou les caractéristiques du moteur, influencent davantage la consommation.

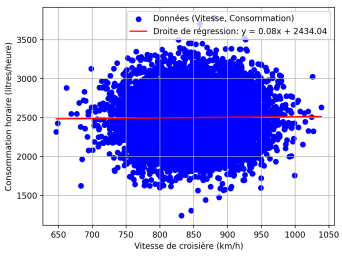


Figure 12: Nuage de points de la consommation horaire en fonction de la vitesse de croisière

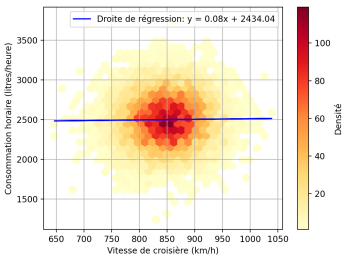


Figure 13: Représentation plus adaptée du graphique précédent

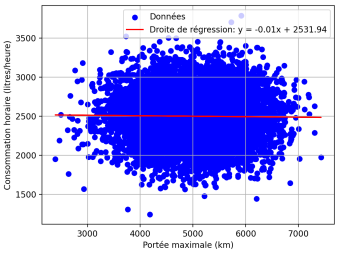


Figure 14: Nuage de points de la consommation horaire en fonction de la portée maximale

La tendance observée est surprenante : selon les principes de l'aviation, une hausse de la consommation de carburant devrait réduire la portée maximale. Plus un avion consomme, moins il parcourt de distance. La relation presque horizontale entre ces variables suggère l'influence d'autres facteurs rendant la portée plus complexe que prévu.

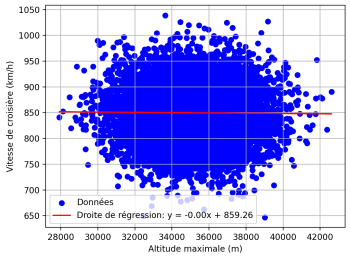


Figure 15: Nuage de points de la vitesse de croisière en fonction de l'altitude maximale

Le graphique précédent montre qu'il existe peu de relation entre l'altitude maximale et la vitesse de croisière, la droite de régression étant presque horizontale. Ainsi, malgré les variations d'altitude, la vitesse reste stable autour de 859 km/h. Cette absence de corrélation suggère que d'autres facteurs, comme le type d'avion ou les moteurs, influencent davantage cette vitesse.

Les avions à haute altitude ne sont pas forcément plus rapides, mais peuvent atteindre de plus grandes vitesses de croisière grâce à une résistance de l'air réduite. L'air moins dense diminue la consommation de carburant pour une vitesse similaire. Certains avions atteignent ces altitudes avec des moteurs plus puissants, une structure renforcée et un système de pressurisation, assurant des conditions sûres dans des zones à faible densité d'air.

3.3 Analyse multidimensionnelle

Cette section analyse les relations entre la vitesse de croisière, l'altitude maximale et la consommation horaire des avions, en explorant leurs interactions et leur impact sur la performance globale. Nous identifierons également les modèles les plus performants. **Le choix a été fait de considérer la consommation horaire en fonction de la vitesse de croisière et de l'altitude**.

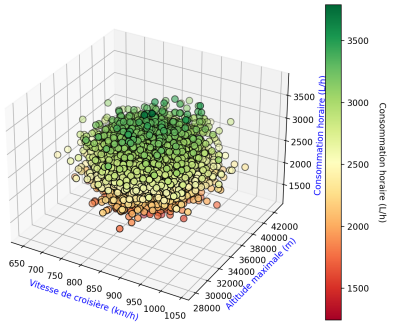


Figure 16: Graphique 3D représentant les interactions entre les 3 variables

Soit $z = \alpha x + \beta y + \gamma$. On réécrit cette équation sous forme matricielle et considérons qu'il est impossible de trouver une solution exacte au système. On cherche alors α, β, γ tels que la norme suivante soit minimisée : $\|AX - Z\|_2$, avec l'équation normale associée donnée par $A^\top AX = A^\top Z$. La solution est alors donnée par $X = \left(A^\top A\right)^{-1} \cdot \left(A^\top Z\right)$. Pour évaluer la qualité de l'ajustement, on utilise l'erreur relative : $\frac{\|z - z_a\|}{\|z\|}$. Si cette erreur est importante, il est néanmoins possible d'obtenir des valeurs cohérentes pour certains modèles d'avion. On pourrait alors chercher, parmi 10 000 modèles, celui qui minimise la norme absolue suivante : $\min(|z - z_a|)$. On trouve alors les coefficients $\alpha = 2449$, $\beta = 0.078$, $\gamma \approx 0$.

Bien que de légères différences numériques soient présentes, les performances globales des différents modèles sont similaires, sans différences significatives dans leur comportement. Le B787 se distingue globalement par sa faible consommation horaire (2 495,6 litres) et sa grande portée (5 017,4 km), en faisant un modèle optimisé pour l'efficacité et les vols long-courriers. Le Cessna172, quant à lui, présente une portée plus modeste (699,7 km), mais avec une faible variabilité, indiquant des performances fiables. Le B787 est le modèle le plus performant.

4 Estimation ponctuelle et intervalle de confiance

Cette section aborde l'estimation ponctuelle et les intervalles de confiance à travers deux études : une expérience de chute libre pour estimer le temps de chute d'un objet, et une analyse des lois de Kepler.

4.1 Chute libre d'un objet

Cette expérience consiste à mesurer le temps de chute d'un objet depuis une hauteur donnée, en estimant le temps moyen et en calculant l'intervalle de confiance pour quantifier l'incertitude des mesures.

La loi normale est utilisée pour estimer la moyenne et l'écart-type si n est grand ; sinon, on utilise la loi T de Student pour un petit échantillon ($n < 30$). La moyenne suit une loi normale pour n grand, grâce au théorème central limite. Pour un petit échantillon, ce n'est pas garanti. La loi T est utilisée quand on a un petit échantillon et une variance inconnue, ajustant l'estimation de la moyenne en fonction de l'incertitude. Aussi, si $n > 30$, la loi normale est applicable pour

l'intervalle de confiance, car la variance est mieux estimée et la précision augmente.

La moyenne des mesures est Moyenne $= \frac{0.64+0.63+\cdots+0.61}{10} = 0.626\,\text{s}$, et l'écart-type estimé est : $s \approx 0.018\,\text{s}$. L'intervalle de confiance est donc : IC $= [0.612, 0.640]$. Cela signifie que la véritable moyenne se trouve dans cet intervalle avec 95% de confiance. Un échantillon plus grand donnerait une estimation plus précise de la moyenne, avec un intervalle de confiance plus étroit.

4.2 Lois de Kepler & gravitation universelle

La 3ième loi de Kepler, $T^2 = k \cdot a^3$, relie la période orbitale d'une planète à la distance moyenne au Soleil, permettant de comprendre les mouvements planétaires. Les données expérimentales permettent de mesurer et estimer les incertitudes liées à ces phénomènes physiques.
Afin de vérifier la loi de Kepler, on vérifie si T^2 est proportionnel à a^3 en comparant les valeurs expérimentales. Si la loi est respectée, la constante k doit être similaire pour toutes les planètes, c'est le cas (voir Table 6), car on trouve bien le même k (2.97×10^{-19}).
On peut également vérifier visuellement que la relation est linéaire. Le nuage de points est aligné, cela confirme la loi de Kepler (voir Table 7).

Enfin, on peut appliquer une régression linéaire pour ajuster les données et quantifier la relation entre T^2 et a^3. Avec l'outil `linregress`, on trouve que $r \approx 1$, cela indique une corrélation presque parfaite (voir Table 7).

Planète	T^2 (s^2)	a^3 (m^3)	k (s^2/m^3)
Mercure	5.78×10^{13}	1.94×10^{32}	2.97×10^{-19}
Vénus	3.77×10^{14}	1.27×10^{33}	2.98×10^{-19}
Terre	9.96×10^{14}	3.35×10^{33}	2.97×10^{-19}
Mars	3.52×10^{15}	1.18×10^{34}	2.98×10^{-19}
Jupiter	1.40×10^{17}	4.72×10^{35}	2.97×10^{-19}
Saturne	8.64×10^{17}	2.95×10^{36}	2.93×10^{-19}

Table 6: Valeurs de T^2, a^3 et $k = \frac{T^2}{a^3}$ pour chaque planète

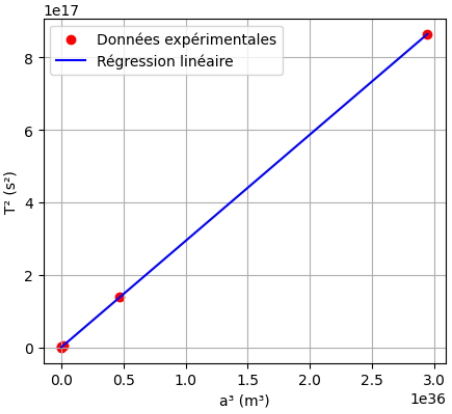


Table 7: Vérification de la troisième loi de Kepler

5 Modélisation statique des temps de fabrication

L'analyse des temps de fabrication de 50 produits vise à évaluer si ces données peuvent être modélisées par une loi normale. Les statistiques descriptives indiquent une moyenne de $\mu = 24.82$ minutes, un écart-type de $\sigma = 2.93$ minutes et une médiane de 24.0 minutes. Le graphique montre un histogramme des données, avec la densité empirique (courbe verte) et la densité théorique ajustée selon une loi normale (courbe rouge pointillée). Bien que la densité empirique suive approximativement la forme normale, quelques écarts sont notés aux extrémités. Ces observations suggèrent qu'une modélisation par une loi normale est plausible.

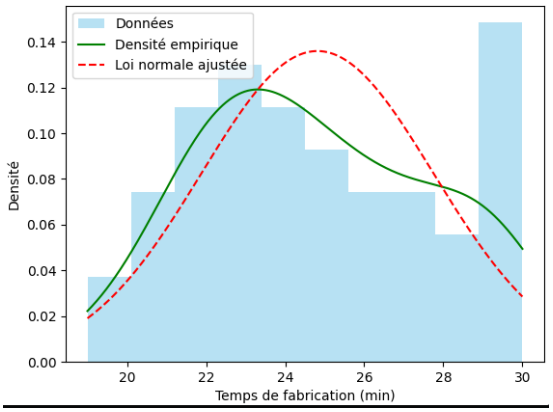


Figure 17: Histogramme et ajustement par une loi normale

6 Conclusion

Ce projet a permis d'explorer l'analyse statistique à travers des données réelles, en appliquant des méthodes univariées, bivariées et multidimensionnelles. Les résultats ont montré des différences importantes dans la production laitière selon le type d'aliment, avec une plus grande variabilité pour l'aliment X. En ce qui concerne la performance des avions, les analyses ont révélé des relations complexes entre la consommation de carburant, la vitesse de croisière et la portée. Ce travail met en lumière l'importance de comprendre les données et les interrelations pour prendre des décisions éclairées dans des domaines variés.