SONNTAG Mikail

21800141

Université de Strasbourg

Projet en Statistiques avec Python

2024 - 2025

Contents

1 Exercice 1 :

1.1 Problématique

Nous disposons d'un échantillon composé de garçons et de filles issus d'un district écossais. Cet échantillon est constitué d'observations indépendantes. Les données sont résumées dans le tableau suivant :

	Blond	Roux	Châtain	Brun	Noir de jais
Garçon	592	119	849	504	36
Fille	544	97	677	451	14

Table 1: Tableau de contingence : Répartition des couleurs de cheveux en fonction du sexe.

Nous souhaitons déterminer si la couleur des cheveux est indépendante du sexe à l'aide d'un test d'indépendance du χ^2 (khi-deux). Ce test permet de vérifier si deux variables qualitatives observées sur un échantillon sont indépendantes.

1.2 Hypothèses et seuil α

Posons notre hypothèse nulle et notre hypothèse alternative :

- H_0 : La couleur des cheveux est indépendante du sexe.
- H_1 : La couleur des cheveux dépend du sexe.

Nous fixons également le seuil $\alpha = 5\%$ (soit $\alpha = 0.05$). Ce seuil représente la probabilité de rejeter à tort l'hypothèse nulle en commettant une erreur de type I.

1.3 Vérification des conditions d'applications du test

Avant d'utiliser le test d'indépendance du χ^2 (khi-deux), nous devons vérifier les conditions nécessaires à son application.

1. Les individus composant l'échantillon doivent être choisis aléatoirement : c'est le cas grâce au protocole de récolte des données.
2. Les classes des variables doivent être exclusives : c'est le cas, car chaque individu appartient à une seule catégorie pour chaque variable.
3. Plus de 80% des effectifs théoriques doivent être supérieurs ou égaux à 5. Le tableau suivant nous permet de vérifier cette condition :

	Blond	Roux	Châtain	Brun	Noir de jais
Garçon	614,37	116,82	825,29	516,48	27,04
Fille	521,63	99,18	700,71	438,52	22,96

Table 2: Tableau de contingence des effectifs théoriques : Répartition des couleurs de cheveux en fonction du sexe.

4. La taille de l'échantillon doit être suffisamment grande : ici, elle est supérieure à 30 (qui est une valeur choisie arbitrairement).

Les conditions étant vérifiées, nous pouvons maintenant calculer la statistique de décision, ainsi que la p-valeur, pour déterminer si les différences observées sont suffisamment significatives pour rejeter l'hypothèse nulle.

Avant de donner les résultats, il est important de définir ces quelques notions :

- La **statistique de décision** (ou **statistique de test**) est une valeur calculée à partir des données observées, qui permet de déterminer si l'hypothèse nulle H_0 doit être rejetée. Elle est comparée à une valeur critique au seuil α.
- La **p-valeur** mesure la probabilité d'obtenir un résultat au moins aussi extrême que celui observé, sous l'hypothèse nulle. Si cette valeur est faible (inférieure à α), cela signifie que le résultat est très peu probable si l'hypothèse nulle était vraie, ce qui nous pousse à la rejeter. La décision du test est prise grâce à la p-valeur.
- Sous H_0, les **degrés de liberté** dépendent du nombre d'observations et du nombre de groupes ou de paramètres estimés. Ils servent à déterminer la forme de la distribution utilisée pour le test, afin de calculer la statistique de test et de prendre une décision.

1.4 Résultats du test

- La statistique χ^2 calculée est de $\mathbf{X^2 \approx 10.47}$.
- Sous H_0, elle suit une loi du χ^2 avec **4 degrés de liberté**.
- La p-valeur associée est de $\mathbf{p \approx 0.033}$.

1.5 Conclusion

Au seuil de significativité $\alpha = 0.05$, la p-valeur ($p \approx 0.033$) est inférieure à α. Nous rejetons donc H_0. Il y a suffisamment de preuves statistiques pour conclure que **la couleur des cheveux dépend du sexe**.

2 Exercice 2 :

2.1 Problématique

Alice, récemment arrivée en Alsace, souhaite savoir si les Manele et les Manala diffèrent en poids. Elle a acheté au hasard des brioches en forme de bonhomme chez divers artisans boulangers d'Alsace, puis les a pesées avec la même balance. Les données collectées sont exprimées en grammes.

Pour répondre à sa question, nous utiliserons un test de Student, qui permet de comparer les moyennes de deux groupes afin de déterminer s'il existe une différence statistiquement significative entre eux. Nos deux échantillons seront les suivants : l'échantillon 1 contenant les poids des Manele et l'échantillon 2 contenant les poids des Manala.

2.2 Statistiques descriptives

Avant de commencer à poser les hypothèses du test, il est important de calculer les statistiques descriptives pour avoir une idée globale de nos données, grâce à la moyenne, la médiane, l'écart-type, les quartiles... Cela nous permet de mieux comprendre la structure des données avant d'effectuer des tests statistiques.

Statistique	**Valeur**
Nombre d'observations (n)	27
Minimum	96.62
Maximum	124.27
Moyenne	110.39
Variance	44.82
Écart-type	6.70
Asymétrie (Skewness)	0.04
Kurtose (Kurtosis)	-0.45
Quartile inférieur (Q1)	106.43
Médiane	110.4
Quartile supérieur (Q3)	114.87

Table 3: Statistiques descriptives des données Manele

Statistique	**Valeur**
Nombre d'observations (n)	24
Minimum	94.41
Maximum	131.22
Moyenne	112.39
Variance	65.89
Écart-type	8.12
Asymétrie (Skewness)	0.35
Kurtose (Kurtosis)	0.19
Quartile inférieur (Q1)	106.74
Médiane	111.22
Quartile supérieur (Q3)	117.19

Table 4: Statistiques descriptives des données Manala

2.3 Hypothèses et seuil α

Posons notre hypothèse nulle et notre hypothèse alternative :

- H_0 : Le poids moyen des Manele est égal au poids moyen des Manala.
- H_1 : Le poids moyen des Manele est différent de celui des Manala.

Fixons comme avant le seuil $\alpha = 5\%$ (soit $\alpha = 0.05$).

2.4 Vérification des conditions d'applications du test

Avant d'utiliser le test de Student, nous devons vérifier les conditions nécessaires à son application.

1. Chaque échantillon doit être constitué d'observations indépendantes : c'est bien le cas de nos données car, chaque brioche est achetée auprès d'un artisan différent et chaque poids observé est une mesure indépendante des autres.

2. Les deux échantillons doivent être indépendants : c'est également le cas, car les échantillons des Manele et des Manala proviennent de différents artisans, et le poids des Manele n'influence pas celui des Manala

3. L'échantillon 1 doit suivre une distribution normale : $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$. Pour vérifier cette condition nous devons utiliser un test de Shapiro-Wilk qui est un test d'ajustement à une loi normale.

   - Hypothèses :
     - H_0 : L'échantillon 1 est issu d'une population normalement distribuée.

- H_1 : L'échantillon 1 n'est pas issu d'une population normalement distribuée.
  On fixe le seuil $\alpha = 5\%$ (soit $\alpha = 0.05$).

- Vérification : Afin d'utiliser ce test, l'échantillon doit être composé d'observations indépendantes, ce qui est le cas ici.
- Résultats : La statistique du test calculée est $\approx \mathbf{0.99}$ et la p-valeur associée est de $\mathbf{p \approx 0.99}$.
- Conclusion : Au seuil de significativité $\alpha = 0.05$, la p-valeur ($p \approx 0.99$) est supérieur à α. Nous ne rejetons pas H_0, nous concluons donc que **l'échantillon 1 suit une distribution normale**.

4. L'échantillon 2 doit également suivre une distribution normale : $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$. Nous utiliserons encore un test de Shapiro-Wilk pour vérifier cette condition.

   - Hypothèses :
     - H_0 : l'échantillon 2 est issue d'une population normalement distribuée.
     - H_1 : L'échantillon 2 n'est pas issue d'une population normalement distribuée.
       On fixe le seuil $\alpha = 5\%$ (soit $\alpha = 0.05$).
   - Vérification : Afin d'utiliser ce test, l'échantillon doit être composé d'observations indépendantes, ce qui est le cas ici.
   - Résultats : La statistique du test calculée est $\approx \mathbf{0.96}$ et la p-valeur associée est de $\mathbf{p \approx 0.50}$.
   - Conclusion : Au seuil de significativité $\alpha = 0.05$, la p-valeur ($p \approx 0.50$) est supérieur à α. Nous ne rejetons pas H_0, nous concluons donc que **l'échantillon 2 suit une distribution normale**.

5. σ_1 et σ_2 doivent être inconnus, ce qui est le cas ici, car nous n'avons pas accès aux populations complètes des Manele et des Manala.

6. σ_1 et σ_2 doivent être égaux. Pour vérifier cette condition, nous utiliserons un test de Fisher-Snedecor qui permet de comparer les variances de deux échantillons afin de déterminer si elles sont statistiquement similaires.

   - Hypothèses :
     - H_0 : $\sigma_1 = \sigma_2$.
     - H_1 : $\sigma_1 \neq \sigma_2$.
       On fixe le seuil $\alpha = 5\%$ (soit $\alpha = 0.05$).
   - Vérification : Afin d'utiliser ce test, nous devons vérifier ces conditions d'applications :
     (a) Chaque échantillon doit être composé d'observations indépendantes.
     (b) Les deux échantillons doivent être indépendantes.
     (c) L'échantillon 1 doit suivre une distribution normale : $X_2 \sim \mathcal{N}(\mu_1, \sigma_1)$.
     (d) L'échantillon 2 doit suivre une distribution normale : $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$.

     Toutes ces conditions ont été vérifiées précédemment, nous pouvons donc utiliser ce test.

- Résultats : La statistique du test calculée est $\approx \mathbf{0.68}$. Sous H_0, elle suit une loi de Fischer à **26 et 23 degrés de libertés**. La p-valeur associée est de $\mathbf{p \approx 0.83}$.
- Conclusion : Au seuil de significativité $\alpha = 0.05$, la p-valeur ($p \approx 0.83$) est supérieur à α. Nous ne rejetons pas H_0, nous concluons donc que $\sigma_1 = \sigma_2$.

Les conditions étant vérifiées, nous pouvons maintenant calculer la statistique de décision, ainsi que la p-valeur, pour déterminer si les différences observées sont suffisamment significatives pour rejeter l'hypothèse nulle.

2.5 Résultats du test

- La statistique calculée est $\approx \mathbf{-0.96}$.
- Sous H_0, elle suit une loi de Student à **49 degrés de liberté**.
- La p-valeur associée est de $\mathbf{p \approx 0.34}$.

2.6 Conclusion

Au seuil de significativité $\alpha = 0.05$, la p-valeur ($p \approx 0.34$) est supérieur à α. Nous ne rejetons pas H_0, ce qui signifie qu'il n'y a pas suffisamment de preuves statistiques pour affirmer qu'il existe une différence significative entre le poids moyen des Manele et des Manala. Nous concluons donc que **le poids moyen des Manele est égal au poinds moyen des Manala**.

3 Exercice 3 :

3.1 Problématique

Rémi, philatéliste passionné, souhaite savoir si l'épaisseur des timbres diffère selon leur pays d'origine. Il a sélectionné au hasard des timbres non abîmés de sa collection provenant d'Allemagne, d'Autriche, de Belgique et de France, puis a mesuré leur épaisseur en micromètres.

Pour répondre à sa question, nous utiliserons une analyse de variance (ANOVA), qui permet de comparer les moyennes de plusieurs échantillons pour déterminer s'il existe une différence statistiquement significative entre elles. Les échantillons correspondent ici aux timbres des quatre pays, avec l'échantillon 0 l'Allemagne, l'échantillon 1 l'Autriche, l'échantillon 2 la Belgique et l'échantillon 3 la France.

3.2 Statistiques descriptives

Calculons les statistiques descriptives de nos quatres échantillons.

Statistique	Valeur
Nombre d'observations (n)	19
Minimum	238
Maximum	261
Moyenne	251.63
Variance	36.58
Écart-type	6.05
Asymétrie (Skewness)	-0.31
Kurtose (Kurtosis)	-0.42
Quartile inférieur (Q1)	246.5
Médiane	252
Quartile supérieur (Q3)	256

Table 5: Statistiques descriptives de l'échantillon 0

Statistique	Valeur
Nombre d'observations (n)	25
Minimum	242
Maximum	265
Moyenne	251.81
Variance	39.39
Écart-type	6.28
Asymétrie (Skewness)	0.06
Kurtose (Kurtosis)	0.94
Quartile inférieur (Q1)	246
Médiane	252
Quartile supérieur (Q3)	258

Table 6: Statistiques descriptives de l'échantillon 1

Statistique	Valeur
Nombre d'observations (n)	23
Minimum	237
Maximum	265
Moyenne	253.22
Variance	46.63
Écart-type	6.83
Asymétrie (Skewness)	-0.34
Kurtose (Kurtosis)	-0.30
Quartile inférieur (Q1)	247.5
Médiane	254
Quartile supérieur (Q3)	258

Table 7: Statistiques descriptives de l'échantillon 2

Statistique	Valeur
Nombre d'observations (n)	22
Minimum	196
Maximum	230
Moyenne	210.77
Variance	72.56
Écart-type	8.52
Asymétrie (Skewness)	0.09
Kurtose (Kurtosis)	-0.34
Quartile inférieur (Q1)	202.25
Médiane	211
Quartile supérieur (Q3)	216

Table 8: Statistiques descriptives de l'échantillon 3

3.3 Hypothèses et seuil α

Posons notre hypothèse nulle et notre hypothèse alternative :

- H_0 : L'épaisseur moyenne des timbres est la même pour les quatre pays.
- H_1 : Au moins une des moyennes diffère des autres.

Fixons comme avant le seuil $\alpha = 5\%$ (soit $\alpha = 0.05$).

3.4 Vérification des conditions d'applications du test

Avant d'utiliser l'ANOVA, nous devons vérifier les conditions nécessaires à son application.

1. Chaque échantillon doit être constitué d'observations indépendantes : c'est bien le cas grâce au protocole de récolte des données, car les épaisseurs des timbres ont été mesurées individuellement et les timbres ont été sélectionnés de manière aléatoire.

2. Les quatre échantillons doivent être indépendants : c'est également le cas, car les timbres de chaque pays ont été sélectionnés séparément et de manière aléatoire.

3. Les quatres échantillons doivent suivre une distribution normale : $X_0 \sim \mathcal{N}(\mu_0, \sigma_0)$, $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, $X_3 \sim \mathcal{N}(\mu_3, \sigma_3)$.
   Pour vérifier cette condition, nous devons utiliser quatre tests de Shapiro-Wilk, un pour chacun des quatre échantillons.

   - Hypothèses :
     - H_0 : L'échantillon 0 (respectivement 1, 2 ou 3) est issu d'une population normalement distribuée.
     - H_1 : L'échantillon 0 (respectivement 1, 2 ou 3) n'est pas issu d'une population normalement distribuée.
       On fixe le seuil $\alpha = 5\%$ (soit $\alpha = 0.05$).
   - Vérification : Afin d'utiliser ce test, l'échantillon 0 (respectivement 1, 2, ou 3) doit être composé d'observations indépendantes, ce qui est le cas ici.
   - Résultats : La statistique du test calculée pour :
     - l'échantillon 0 est $\approx \mathbf{0.97}$ et la p-valeur associée est de $\mathbf{p \approx 0.69}$.
     - l'échantillon 1 est $\approx \mathbf{0.95}$ et la p-valeur associée est de $\mathbf{p \approx 0.27}$.
     - l'échantillon 2 est $\approx \mathbf{0.98}$ et la p-valeur associée est de $\mathbf{p \approx 0.81}$.
     - l'échantillon 3 est $\approx \mathbf{0.98}$ et la p-valeur associée est de $\mathbf{p \approx 0.91}$.
   - Conclusion : Au seuil de significativité $\alpha = 0.05$, les p-valeur des quatres échantillons étant supérieur à α, nous ne rejetons pas H_0 pour les quatres tests. Nous concluons donc que **les quatres échantillons suivent une distribution normale**.

4. σ_0, σ_1 σ_2 et σ_3 doivent être égaux. Pour vérifier cette condition, nous utiliserons un test de Bartlett qui permet de comparer les variances de plusieurs échantillons afin de déterminer si elles sont statistiquement similaires.

   - Hypothèses :
     - H_0 : Les variances des quatre échantillons sont égales.
     - H_1 : Au moins une variance diffère des autres.
       On fixe le seuil $\alpha = 5\%$ (soit $\alpha = 0.05$).
   - Vérification : Afin d'utiliser ce test, nous devons vérifier ces conditions d'applications :
     (a) Chaque échantillon doit être composé d'observations indépendantes.
     (b) Les quatres échantillons suivent une distribution normale. Ces deux conditions ont été vérifiées précédemment, nous pouvons donc utiliser ce test.

     Ces deux conditions ayant été vérifiées précédemment, nous pouvons donc utiliser ce test.

- Résultats : La statistique du test calculée est $\approx \mathbf{5.53}$. La p-valeur associée est de $\mathbf{p \approx 0.14}$.
- Conclusion : Au seuil de significativité $\alpha = 0.05$, la p-valeur ($p \approx 0.14$) est supérieur à α. Nous ne rejetons pas H_0, nous concluons donc que $\sigma_0 = \sigma_1 = \sigma_2 = \sigma_3$.

Les conditions étant vérifiées, nous pouvons maintenant calculer la statistique de décision, ainsi que la p-valeur, pour déterminer si les différences observées sont suffisamment significatives pour rejeter l'hypothèse nulle.

3.5 Résultats du test

- La statistique calculée est $\approx \mathbf{194.62}$.
- La p-valeur associée est de $\mathbf{p \approx 5.82 * 10^{-38}}$.

3.6 Conclusion

Au seuil de significativité $\alpha = 0.05$, la p-valeur ($p \approx 5.82 * 10^{-38}$) étant largement inférieur à α, nous rejetons l'hypothèse H_0. Nous concluons donc qu'**il existe une différence statistiquement significative entre les épaisseurs des timbres des différents pays**.

3.7 Test post-hoc de Tukey

Après avoir rejeté l'hypothèse nulle, il est possible d'identifier spécifiquement quels groupes diffèrent entre eux. Pour cela, nous pouvons utiliser un test post-hoc comme le test de Tukey. Ce test permet de comparer toutes les paires possibles de groupes, tout en contrôlant le taux d'erreur de type I. Ce test est applicable lorsque les groupes sont indépendants et suivent une distribution normale, avec des variances entres les groupes égaux. Ces conditions ayant déjà été vérifiées précédemment nous pouvons l'appliquer. Voici le graphique et le tableau obtenues après avoir réalisé le test.

Comparaison	Différence de Moyenne	Intervalle de Confiance	p-valeur
Groupe 0 vs Groupe 1	0.2084	[-5.367, 5.7839]	0.9997
Groupe 0 vs Groupe 2	1.5858	[-4.0933, 7.265]	0.8841
Groupe 0 vs Groupe 3	-40.8589	[-46.5961, -35.1216]	0.0
Groupe 1 vs Groupe 2	1.3774	[-3.9154, 6.6702]	0.9037
Groupe 1 vs Groupe 3	-41.0673	[-46.4224, -35.7122]	0.0
Groupe 2 vs Groupe 3	-42.4447	[-47.9077, -36.9817]	0.0

Table 9: Résultats du test de Tukey pour la comparaison des épaisseurs des timbres entre pays

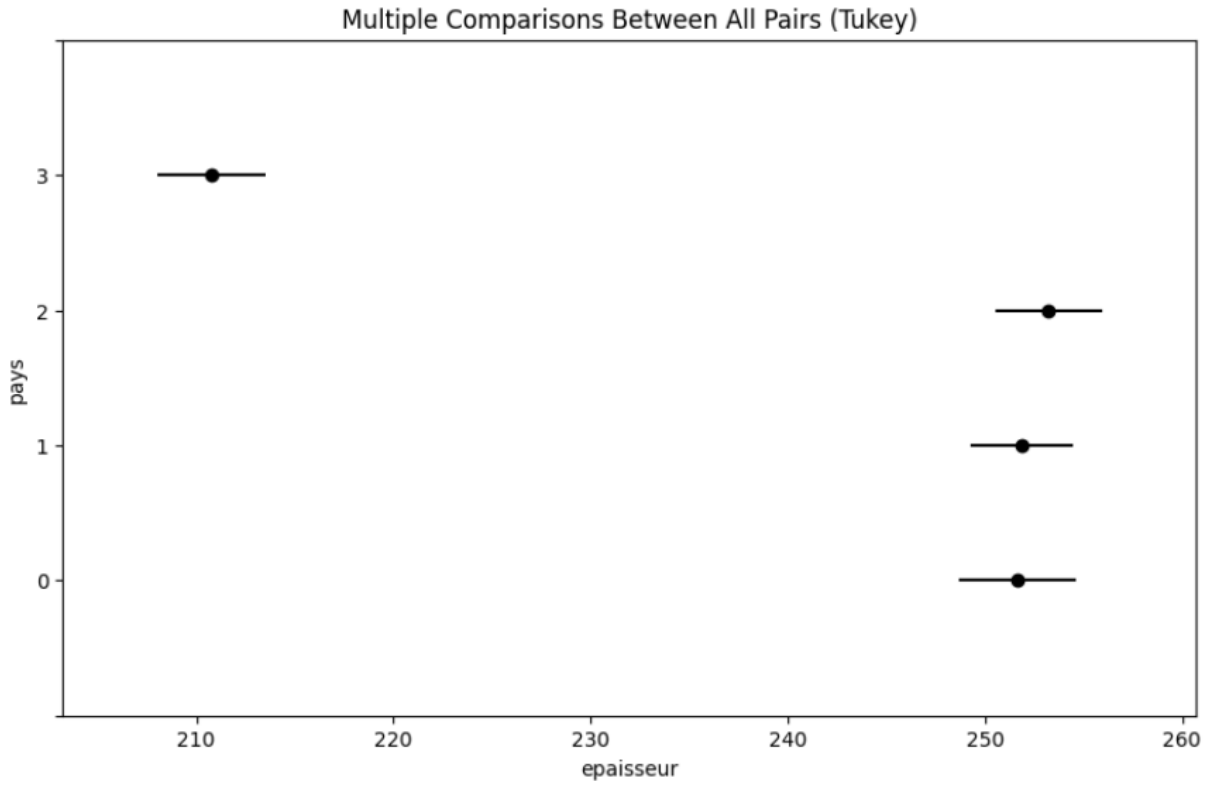

Figure 1: Graphique des résultats du test de Tukey

Les points représentent les moyennes des groupes, et les traits correspondent aux intervalles de confiance pour chaque groupe. Les intervalles de confiance indiquent la plage dans laquelle se situe la véritable moyenne de chaque groupe avec un certain niveau de confiance, ici à 95%. Le groupe 3 se trouve à gauche en raison de son épaisseur de timbres significativement différente de celle des groupes 0, 1 et 2 qui se trouvent à droite. Les groupes à droite montrent des épaisseurs similaires entre eux. Les p-valeurs ajustées présentes dans le tableau nous permettent de donner des conclusions au seuil de significativité $\alpha = 0.05$. Les valeurs supérieures au seuil $\alpha = 0.05$ indiquent qu'il n'y a pas de différence statistiquement significative entre les groupes comparés, tandis que les p-valeurs ajustées inférieures à $\alpha = 0.05$ montrent des différences significatives, permettant de rejeter l'hypothèse nulle.

Les résultats du test de Tukey permettent ainsi d'identifier que **les timbres français se distinguent des autres pays en termes d'épaisseur**, tandis que **les timbres des autres pays semblent avoir des épaisseurs similaires**.