Exercice 1

Dans cette exercice le but est de procéder à l'analyse statistique adéquate pour pouvoir déterminer si la couleur des cheveux est indépendante du sexe

1) Hypotheses à tester:

H_0 : les variables sont indépendantes. La couleur des cheveux est indépendante du sexe

H_1 : les variables ne sont pas indépendantes. La couleur des cheveux n'est pas indépendante du sexe

2) Choix de α

on va prendre $\alpha = 5$ car il représente un bon compromis entre le risque de rejeter à tort l'hypothèse nulle (erreur de type I) et le risque de ne pas rejeter une hypothèse nulle fausse (erreur de type II).

3) Choix du test Correspondant

a) Test du χ^2

Nous allons regarder dans un premier temps le test du χ^2 de Pearson d'indépendance qui permet de tester si 2 variables qualitatives observées sur 1 échantillon sont indépendantes. En l'occurence ici, de répondre à la problématique si la couleur de cheveux est indépendente du sexe. Il suffit de vérifier que les conditions d'applications soient satisfaites :

Conditions d'applications:

- Les individus composant l'échantillon ont été choisis aléatoirement car il est mentionné « Cet échantillon est composé d'observations indépendantes » OK par le protocole de récolte des données.
- Les classes des variables sont exclusives. OK
- Règle de Cochran : au moins 80% des effectifs théoriques sont au moins égaux à 5. ?
- La taille de l'échantillon doit lettre assez grande. Assez subjectif mais OK

Vérification des conditions d'application

Il suffit de vérifier si la règle de Cochran est vérifiée

On importe les packages nécessaires dont $\mathbf{pandas, numpy}$ et $\mathbf{stats}$

On crée un objet contenant les effectifs observés $(n_{i,j})$ que j'ai nommé $\mathbf{obs}$

Voici le tableau des effectifs observés grace à python

	BLOND	ROUX	CHÂTAIN	BRUN	NOIR DE JAIS
GARÇON	592	119	849	504	36
FILLE	544	97	677	451	14

On va utiliser la fonction $\mathbf{chi2\_contingency}$ pour obtenir différentes informations dont le tableau des effectifs théoriques et visualiser ce tableau :

Voici le tableau des effectifs théoriques $(t_{i,j})$ obtenu grace à python

	BLOND	ROUX	CHÂTAIN	BRUN	NOIR DE JAIS
GARÇON	614,3703322	116,81689415	825,28972444	516,48210147	24,04094772
FILLE	521,62966778	99,18310585	700,71027556	438,51789853	22,95905228

Maintenant que nous avons les effectifs théoriques nous pouvons vérifier si la règle de Cochran est vérifiée

Visuellement nous pouvons affirmer que cette règle est vérifier mais nous pouvons aller plus loin grâce à un code Python. Il s'agit d'une fonction que j'ai nommé **coch**

On obtient qu'il y a 0.0 % des effectifs théoriques strictement inférieur à 5

Le test de Cochran se révèle donc satisfaisant.

Statistique de décision du test de χ^2

Maintenant comme les conditions d'applications du test sont remplies, nous pouvons regarder les valeurs de la statistique de décision et de la p-valeur

Mon code python renvoie ceci:

La statistique de décision vaut 10.467448536455917 . Sous H0 elle suit une loi du Chi2 à 4 degrés de liberté. Ainsi la p-value est égale à 0.03324834398196967

Comparaison avec α

Nous pouvons procéder à la dernière étape et comparer cette valeur de référence à la valeur critique au seuil α. On a p-value < α

Ainsi nous pouvons rejeter l'hypothèse nulle

C'est à dire qu'il n'y a pas indépendance entre couleur de cheuveux et sexe.

b) Test de Fisher

Il s'agit d'un test qui permet de tester l'indépendance dans le cas où la règle de Cochran n'est pas vérifiée. Ici je considère donc qu'il n'est pas nécessaire de regarder ce test comme le test de cochran est vraiment bien vérifiée.

De plus le test du χ^2 est plus puissant que le test de Fisher lorsque les échantillons sont grands comme dans le contexte.

Néanmoins le test du χ^2 est basé sur une approximation d'un ratio de log-vraisemblance pour calculer la valeur de p. Par conséquent, la précision des estimations dépend des hypothèses et peut ne pas être très précise.

c) G-test

Le G-test n'utilise pas l'approximation mais calcule le vrai rapport de log-vraisemblance ce qui permet d'obtenir des résultats plus fiables. Il peut donc être intérressant de regarder ce test et de utiliser le résultat qu'il en découle.

De plus le G-test est souvent plus puissant que le test du χ^2 et il produit généralement moins d'erreurs de type 1 (rejeter à tort l'hypothèse nulle) et d'erreurs de type 2 (accepter à tort l'hypothèse nulle) que le test du χ^2, aussi il peut donc être utilisé dans des cas où les effectifs sont faibles ce qui fait moins de restrictions.

Conditions d'applications

- Les individus composant l'échantillon ont été choisis aléatoirement car il est mentionné « Cet échantillon est composé d'observations indépendantes » OK
- Les classes des variables sont exclusives. OK
- Règle de Cochran : au moins 80% des effectifs théoriques sont au moins égaux à 5. OK par ce qui précede

Statistique de décision du G-Test

On va utiliser la fonction **chi2_contingency** pour obtenir différentes informations dont le tableau des effectifs théoriques et visualiser ce tableau :

Voici le tableau des effectifs théoriques obtenu grace à python

	BLOND	ROUX	CHÂTAIN	BRUN	NOIR DE JAIS
GARÇON	592	119	849	504	36
FILLE	544	97	677	451	14

qui est le même tableau que pour le test de χ^2 (ce qui est normale)

Maintenant on va regarder la p-value

Grace au code python j'obtiens que la statistique de décision vaut 10.76. Sous H0 elle suit une loi du Chi2 à 4 degrés de liberté. Ainsi la p-value est égale à 0.0295.

Comparaison de la p-value avec α

On a bien p-value < α donc on peut rejeter l'hypothèse nulle.

Conclusion

D'après le G-test nous pouvons conclure qu'il n'y a pas indépendance entre la couleur de cheveux et le sexe.

Exercice 2

Partie 1 : Statistique descriptive

Dans un premier temps nous allons faire les statistiques descriptives liées au poids des Manala et Manele.

Tout d'abord nous allons importer les données dans l'élement **man**

Nous pouvons donc passer aux statistiques descriptives :

On importe les packages adéquat comme **pandas, numpy** et **stats**

Pour Manele j'obtient les données suivante:

Mediane = 110.4

1er Quartile = 106.43

3eme Quartile = 114.87

Nombre de données = 27

Minimum = 96.62

Maximum = 124.27

Moyenne = 110.38740740740741

Variance = 44.82071994301993

Ecart-type = 6.694828

Assymetrie = 0.0429306467729167

Applatissement = -0.4446029513183656

Pour Manala j'obtient les données suivante:

Mediane = 111.215

1er Quartile = 106.7425

3eme Quartile = 117.19

Nombre de données = 24

Minimum = 94.41

Maximum = 131.22

Moyenne = 112.38625

Variance = 65.89455489130434

Ecart-type = 8.117546

Assymetrie = 0.3458209247607714

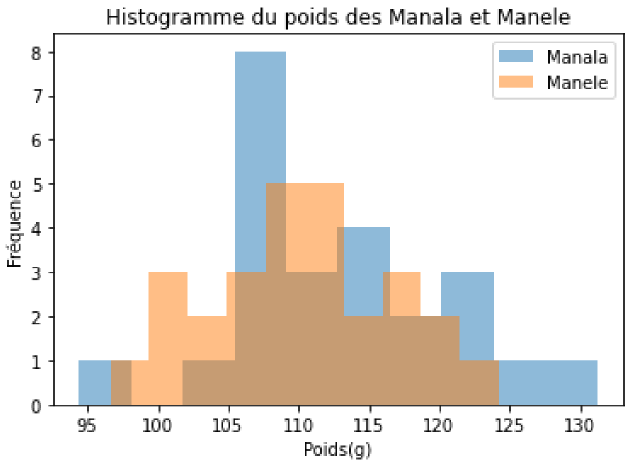
Applatissement = 0.18728434878658273

Pour aller un peu plus loin je me suis dit que cela peut être plus parlant de regarder cela à travers des graphiques

On importe le package nécessaire **matplotlib.pyplot**

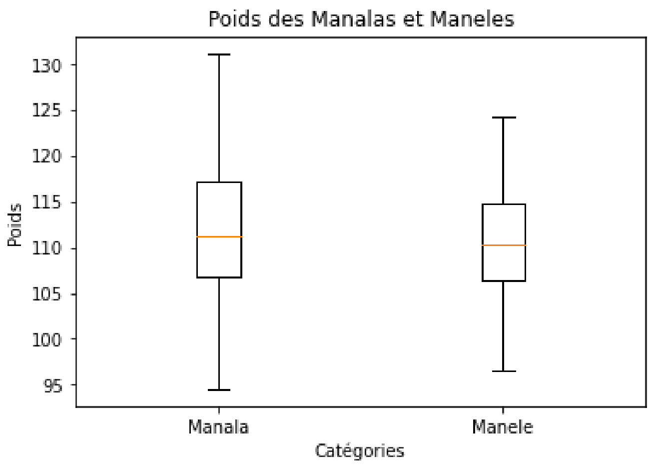
Ensuite on créer les élements nécessaires qui seront aussi importante dans la suite **donneesMe** et **donneesMa** qui regroupe les données du poids des Manele et le poids des Manala

Histogramme



Ce qui est assez parlant ici c'est que l'on peut penser que les échantillons suivent une loi normale d'un point de vue strictement visuel.

Boite à moustache



On peut constater que les moyennes sont à la même hauteur ce qui peut donner une idée concernant la réponse à la problématique

Partie 2 : Traitement statistique

Le but de l'exercice est de répondre à la question de d'alice afin de savoir si il y a une différence de poids entre Manele et Manala.

Ici nous avons deux groupes : « Manele » et « Manala » On peut donc d'après le contexte s'orienter vers un test de Student.

Dans la suite les Me et Ma en indice représentent les données en liens avec Manele et Manala.

1) Hypothèses à tester :

$H_0 : \mu_{Me} = \mu_{Ma}$. Pas de différence significative entre Manele et Manala

H_1 : $\mu_{Me} \neq \mu_{Ma}$. Différence significative entre Manele et Manala

2) Selection du test correspondant : Test de Student

L'utilisation du test statistique de Student présente des avantages en termes de puissance et de robustesse pour détecter de petites différences entre les moyennes des groupes.

Cependant, il est important que les conditions d'applications soient vérifiées pour éviter les erreurs dans le résultat.

Des erreurs de type 1 peuvent se produire en concluant à une différence entre les groupes alors qu'il n'y en a pas réellement, tandis que des erreurs de type 2 peuvent se produire en concluant à l'absence de différence alors qu'il en existe une réellement.

Malgré ces inconvénients, il est possible de minimiser ces erreurs en vérifiant les conditions.

3) Choix du seuil α

On va prendre α = 5%

4) Conditions d'applications:

- Chaque échantillon est composé d'observations indépendantes. OK car "elle sélectionne au hasard un certain nombre" (Protocole de récolte des données.)
- Les 2 échantillons sont indépendants. OK (Protocole de récolte des données).
- $X_1 \sim N(\mu_{Me}, \sigma_{Me})$. ?
- $X_2 \sim N(\mu_{Ma}, \sigma_{Ma})$. ?
- σ_{Me} et σ_{Ma} inconnus. OK
- $\sigma_{Me} = \sigma_{Ma}$. ?

5) Importation des données

On importe les packages nécessaire comme **numpy** et **ttest_ind**

Afin d'éviter d'alourdir le code, on va créer des objets qui seront récurrents dans le code nommé **donneesMe** et **donneesMa**

6) Conditions d'applications : Normalité (Test de shapiro)

On souhaite tester si la distribution de la variable Manele et manala du jeu de données man est normale

Hypotèses à tester

Les hypothèses testées sont :

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

H_1 : L'échantillon n'est pas issu d'une population normalement distribuée.

Choix du seuil α

on va prendre α = 5%

Choix du test correspondant

On peut faire un test de normalité avec un test de Kolmogorov-Smirnov ou un test d'Anderson-Darling ou même d'autres mais le plus adapté à la normalité est le test de Shapiro-Wilk car il ne permet de tester que la normalité.

Conditions d'applications

L'échantillon doit être composé d'observations indépendantes d'une variable quantitative continue. OK par le protocole de récolte des données.

Statistique de décision

J'utlise la commande **stats.shapiro** sur python

Pour les données de Manele, j'obtiens une p-value de 0.9971219897270203 et pour Manala une p-value de 0.5003083348274231

Comparaison avec α

p-value > 5 %

Donc on peut considerer que les données de Menele et Manala suivent une loi normale d'après le test de Shapiro.

Néanmons, il est important de noter qu'il existe un risque d'erreur associé à ce test.

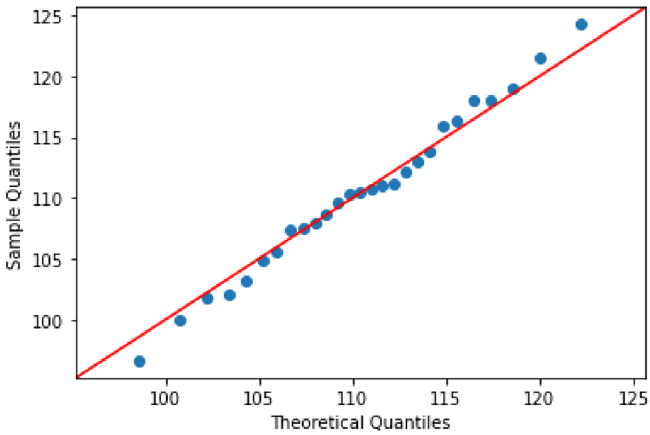
En effet, le principal risque d'erreur du test de Shapiro-Wilk est le risque d'erreur de type I. Cela peut se produire lorsque les données ne sont pas suffisamment nombreuses ou lorsque la distribution des données est très éloignée d'une distribution normale. Ici nous avons 27 et 24 données ce qui est suffisant pour se referer au test de shapiro pour voir si les échantillons suivent une loi normale.

Essayons de voir cela graphiquement compléter ce test par un graphique quantile-quantile (QQ-plot).

QQ-plot

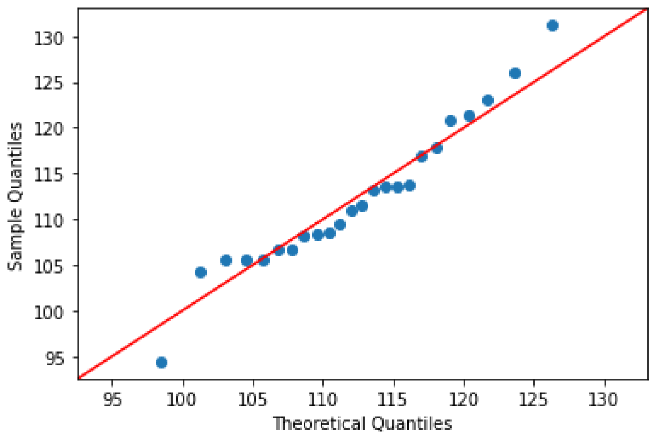
Le Q-Q plot est une méthode graphique pour comparer 2 distributions de probabilité en traçant les quantiles de ces distributions.

Ce qui renvoie pour Menele:



Les points semblent s'aligner le long de la droite. on peut donc considérer que cela suit une loi normale.

Ce qui renvoie pour manala:



Cela veut dire que comparé à la distribution normale il y a un beaucoup plus de données dans les extrémités que dans le centre de la distribution. En effet, ici les points correspondant aux valeurs les plus élevées sont éloignées de la droite pour un certain nombre d'entre eux d'affilée. Ceci indique que les effectifs observés dans l'échantillon sont plus nombreux dans les queux de distribution que ce que la distribution normale prédit en théorie.

J'en conclu que comme il y a assez d'effectif observé, l'hypothèse de normalité est satisfaite.

7) Condition d'application 2 : variance égale (Test de Fischer-Snedecor)

Hypothèses

$H_0 : \sigma^2_{Me} = \sigma^2_{Ma}$

$H_1 : \sigma^2_{Me} \neq \sigma^2_{Ma}$

Seuil α

je vais prendre α = 5 %

Choix du test correspondant

Je vais utliser le test de Fisher-Snedecor qui permet de tester l'égalité des variances.

Conditions d’applications :

- Chaque éhantillon est composé d’observations indépendantes. OK par le protocole de récolte des données.
- Les 2 échantillons sont indépendants. OK par le protocole de récolte des données.
- $X_1 \sim N(\mu_{Me}, \sigma_{Me})$. OK
- $X_2 \sim N(\mu_{Ma}, \sigma_{Ma})$. OK

Statistique de décision du test de Fischer-Snedecor :

Je créer la fonction $\mathbf{f\_test}$ afin de pouvoir avoir les valeurs de la statistique de décision

Ainsi j'obtiens la statistique de décision suivante

La statistique de décision vaut 0.6801885226625094 Sous H_0 elle suit une lois de Fisher à 26 et 23 ddl.la p_value est égale à 0.8293076476081731

Comparaison avec le seuil α

Nous avons que la p-value > α

Ainsi on peut considèrer l'égalité des variances donc le test de Welch n’est pas nécessaire

De plus l'avantage d'utiliser le test de Student plutôt que le test de Welch réside dans la situation où les échantillons comparés présentent des variances égales ou similaires. Dans ce cas, le test de Student est plus puissant que le test de Welch.

Cependant, l'inconvénient d'utiliser le test de Student est qu'il suppose des échantillons indépendants et suit la distribution normale. Si ces hypothèses ne sont pas respectées, le test de Student peut générer des erreurs de type 1 ou des erreurs de type 2. Dans notre cas il y a donc aucun problème.

En termes de puissance, le test de Welch peut être moins puissant que le test de Student lorsque les variances des échantillons sont égales. Cependant, lorsque les variances sont différentes, le test de Welch devient plus puissant que le test de Student. Ici le test de student est donc plus puissant.

Maintenant nous pouvons appliquer le test de Student car les conditions sont vérifiées

8)Statistique de décision du test de Student

Hypotèses :

$H_0 : \mu_{Me} = \mu_{Ma}$

$H_1 : \mu_{Me} \neq \mu_{Ma}$

Grace au code python j'obtiens que la statistique de décision vaut -0.9632469680891367 . Sous H0, elle suit une loi de Student à 49 . La p-value est égale à 0.34015339113128873

9) Comparaison avec le seuil α

p-value >5 % On conserve alors H_0

On ne peut pas rejeter le fait que les moyennes soient égales et donc le fait qu'il n'y ai pas de différence significative entre Manele et Manala . On conclut qu'il n'y a pas suffisamment de preuves pour affirmer que Manele et Manala soient différent (du moins leurs moyennes).

Il y avait encore une possibilité dans le cas où les données ne sont pas indépendantes mais qu'il s'agit d'une dépendance de type avant/aprés on utilise le test de Student apparié. Ici ce n'est vraiment pas nécessaire donc je ne considère pas ce test.

Exercice 3

Partie 1 : Statistique descriptive

Tout d'abord j'importe les données nécessaire sous le nom $\mathbf{timbres}$ et les données des épaisseur en fonction du pays sous les noms $\mathbf{tbrAl}$,$\mathbf{tbrAu}$,$\mathbf{tbrBe}$ et $\mathbf{tbrFr}$

J'importe également les différent package : $\mathbf{pandas}$, $\mathbf{numpy}$ et $\mathbf{stats}$

Allemagne

Mediane = 252.0

1er Quartile = 246.5

3eme Quartile = 256.0

Nombre de données = 19

Minimum = 238

Maximum = 261

Moyenne = 251.6315789473684

Variance = 36.578947368421055

Ecart-type = 6.048053

Assymetrie = -0.31071599868696925

Applatissement = -0.42427667458359686

Autriche

Mediane = 252.0

1er Quartile = 246.0

3eme Quartile = 258.0

Nombre de données = 25

Minimum = 242

Maximum = 265

Moyenne = 251.84

Variance = 39.39000000000001

Ecart-type = 6.276145

Assymetrie = 0.05571239963690025

Applatissement = -0.9365670792868732

Belgique

Mediane = 254.0

1er Quartile = 247.5

3eme Quartile = 258.0

nombre de données = 23

Minimum = 237

Maximum = 265

Moyenne = 253.2173913043478

Variance = 46.63241106719368

Ecart-type = 6.828793

Assymetrie = -0.33689318625492265

Applatissement = -0.30250847594104346

France

Mediane = 211.0

1er Quartile = 205.25

3eme Quartile = 216.0

Nombre de données = 22

Minimum = 196

Maximum = 230

Moyenne = 210.77272727272728

Variance = 72.56493506493506

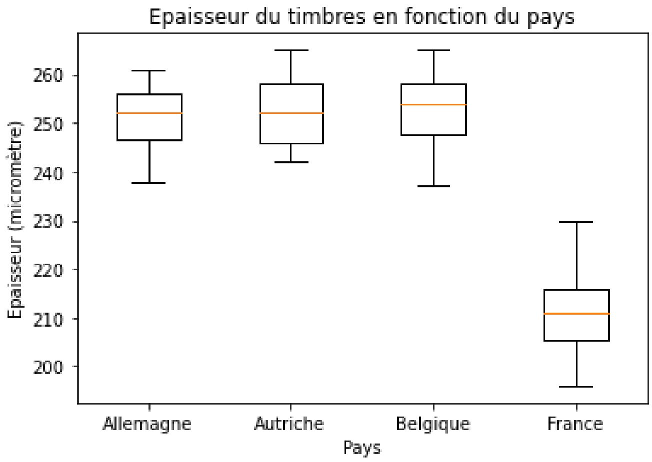
Ecart-type = 8.518505

Assymetrie = 0.08580712110327479

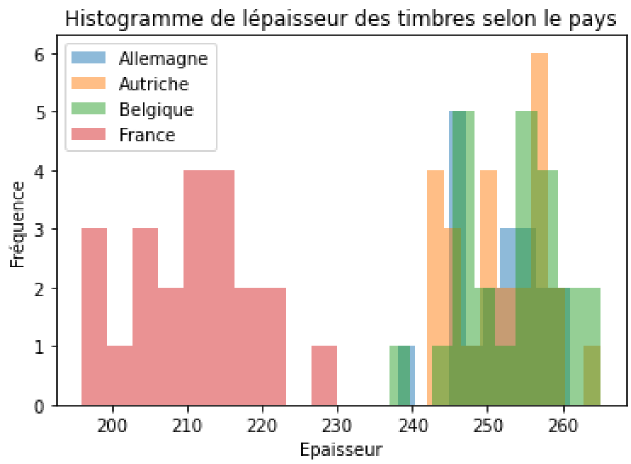
Applatissement =-0.3354933187861082

Boite à Moustache

On utlise le package **matplotlib.pyplot**



Histogramme



On peut commencer à faire des supposition concernant la problématique, ici on peut constater que les moyennes des épaisseurs des timbres sont assez proche à part pour la France. La boite à moustache est assez parlante pour cela ainsi que les statistques sur les moyennes des timbres par pays. Pour affirmer cette supposition nous allons procéder à la seconde partie qui consiste à effectuer les tests statistiques nécessaire pour répondre à la problématique.

Partir 2 : Traitement statistique

On souhaite comparer les moyennes d'une variable aléatoire quantitative (épaisseur) dans différents groupes (pays)

1) Recolte des données

J'ai nommé **tbrAl, tbrAu, tbrBe** et **tbrFr** les données qui vont être récurrente dans le code, ces données regroupe les données des épaisseur des timbre en fonction du pays

Dans la suite les élements avec en indice Al, Au, Be et Fr représentent tout ce qui est liée à l'Allemagne, l'Autriche, La Belgique et la France.

2) Hypothèses

H_0 : $\mu_{Al} = \mu_{Au} = \mu_{Be} = \mu_{Fr}$

H_1 : au moins une moyenne est différente des autres

3) Seuil α

Nous allons prendre α = 5%

4) Choix du test

Nous allons effectuer le test de ANOVA car on souhaite comparer les moyennes d'une variable aléatoire quantitative dans 4 groupes (variable qualitative), il suffit de vérifier que les conditions d'applications soient satisfaites :

5) Conditions du test

- L'échantillon des résidus est composé d'observations indépendantes. OK par le protocole de récolte des données car "Il selectionne de ùanière aléatoire"
- Homoscédasticité : la variance des résidus est la même pour tous les groupes. ?
- ϵ ~ N (0,σ) où σ^2 est la variance commune des résidus.?

En effet il est plus simple de faire un test sur les résidus pour éviter de faire 4 tests de normalité

6) Vérification des conditions du test

Conditions d'applications : test de normalité

Hypothèses à tester

H_0 : l'échantillon est issu d'une population normalement distribuée.

H_1 : l'échantillon n'est pas issu d'une population normalement distribuée.

Seuil α

je prends $\alpha = 5$ %

a) Alternative 1 : test de shapiro-wilk

Condition du test

L'échantillon doit être composé d'observations indépendantes d'une variable quantitative continue. OK par le protocole de récolte des données.

Statistique de décision

Pour l'épaisseur des timbres en Allemagne la p-value est : 0.6912437677383423

Pour l'épaisseur des timbres en Autriche la p-value est : 0.27231159806251526

Pour l'épaisseur des timbres en Belgique la p-value est : 0.8086344599723816

Pour l'épaisseur des timbres en France la p-value est : 0.912662148475647

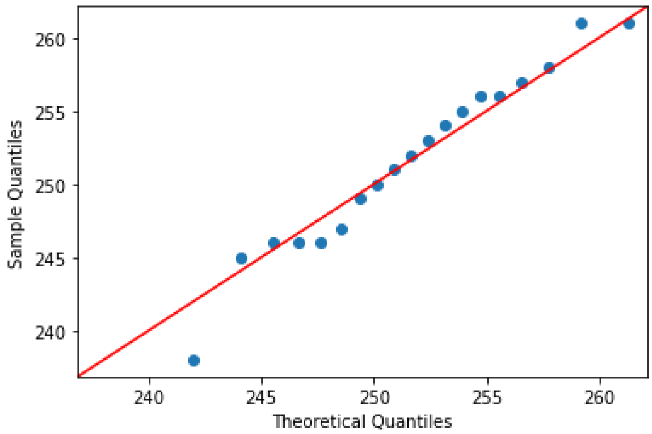
Comparaison avec α

Dans tout les cas on a p-value > α. Ce qui implique qque l'on peut considérer qu'il y a normalité

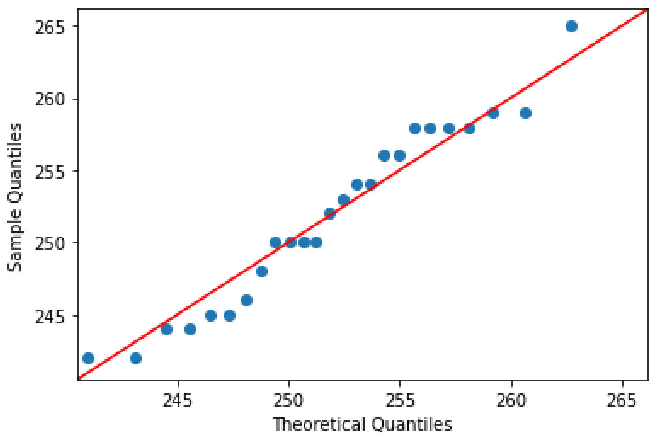
L’ANOVA est relativement robuste à la non-normalité des résidus donc un Q-Q plot peut être intéressant.

b) Alternative 2 : QQ-plot

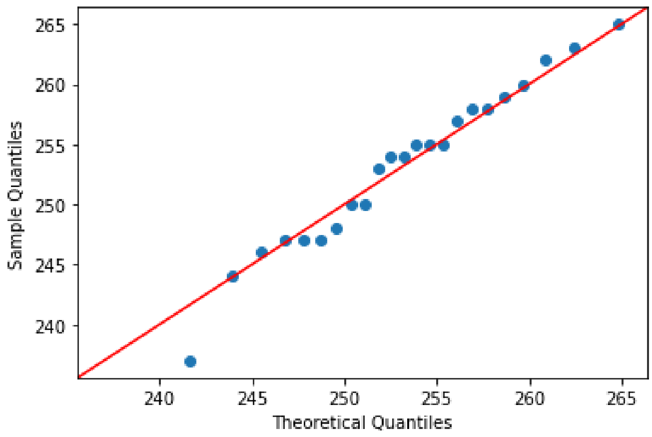
Voici le qq-plot pour l'Allemagne :



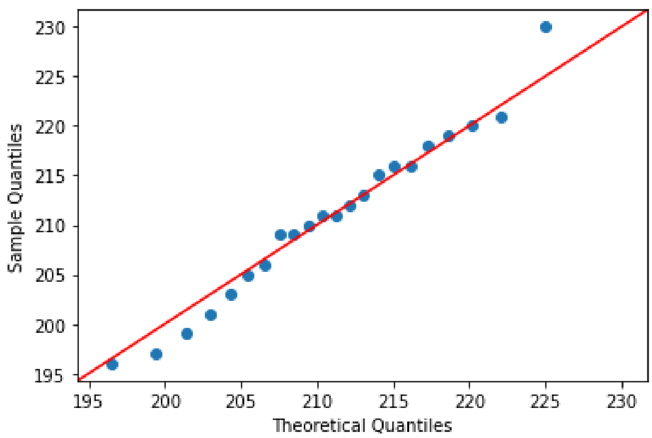
Voici le qq-plot pour l'Autriche :



Voici le qq-plot pour la Belgique:



Voici le QQ-plot pour la France:



Dans ces quatres cas les point suivent le long de la droite sauf peut-être aus extremités, mais ceci n'est pas compromettant étant donné que les échantillons ont des donnnées assez suffisante.

c) alternative 3: test sur les residus

Pour minimiser le nombre de tests et le risque d'erreur de type 2, il vaut mieux faire un seul test sur l'ensemble des résidus. C'est pour ca que je préfere utiliser le test de normalitée de Shapiro sur les résidus plutôt que sur tout l'échantillon car il permet de s'assurer spécifiquement que les erreurs d'un modèle statistique sont normalement distribuées, de plus le test de normalité de Shapiro sur les résidus peut potentiellement réduire le risque de ces erreurs car il évalue spécifiquement la distribution des erreurs d'un modèle. Néanmoins cela réduit la puissance du test. Mais pour ce que l'on souhaite faire cela me semble adéquat.

Récolte des données

Je crée l'objet **model** qui permet de récolter les données de l'épaisseur des timbre en fonction du pays

Condition d'application

L'échantillon doit être composé d'observations indépendantes d'une variable quantitative continue. OK par le protocole de récolte des données.

Statistique de décision

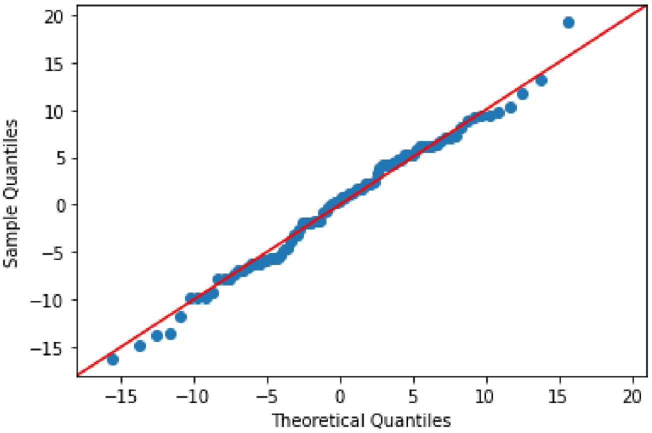
J'obtiens une p-value de 0.7196378111839294

Comparaison avec α

Comme p-value > α j'en conclu l'on peut considerer l'hypothèse de normalité

QQ-plot

Afin d'être sur que les résidus suivent une loi normale j'effectue un QQ-plot



On a bien que les point suivent le long de la droite je considère qu'il y a normalité

Conditions d’applications : Variance, Test de Bartlett :

Hypothèses

$H_0 : \sigma^2_{Al} = \sigma^2_{Au} = \sigma^2_{Be} = \sigma^2_{Fr}$.

H_1 : au moins une variance est différente des autres.

Choix du seuil α

Je prends α = 5%

Choix du test correspondant

On peut dire que le test de Bartlett n'est pas fiable lorsque les données ne sont pas normalement distribuées. Par contre, le test de Levene est moins affecté par la non-normalité des données, mais il est moins puissant que le test de Bartlett. Étant donné que l'homoscédasticité est essentielle pour l'ANOVA comme l'ANOVA n'est pas robuste à l'héteroscédasticité, je préfere utiliser le test de Bartlett autant que possible.

Conditions d'application:

- Observations indépendantes. OK par le protocole de récolte des données.
- Normalité dans chaque groupe (résidus). OK par ce qui precede

Statistique de décision

J'utlise la commande **stats.barlett** Et j'obtiens que la statistique de décision est égale à 3.0509422741943237 . La p-value est 0.38383737450147365

Comparaison avec le seuil α

On a p-value > α, je peux considerer qu'il y a l'égalité des variance et donc on a bien $\sigma_{Al} = \sigma_{Au} = \sigma_{Be} = \sigma_{Fr}$

7) Statistique de décision

Maintenant que les conditions sont vérifiées je peux appliquer ANOVA

J'utlise la commande **sm.stats.anova_lm** pour faire ma statistque de décision

J'obtiens que la p-value vaut 5.824759e-38

8) Comparaison avec α

On a p-value < α donc on rejette H0 c'est à dire que les moyennes ne sont pas toutes égales. Cela revient à dire qu'il y a au moins un pays où l'épaisseur des timbres est différente des autres pays.

Partie 3 : L'après ANOVA

On souhaite comparer les groupes afin de comprendre dans quel groupe il y a une différence.

Nous allons donc faire 6 comparaisons

Bonferroni

On utlise la commande **statsmodels.stats.multicomp**

On obtient le tableau que j'ai retranscrie comme suit :

Group 1	Group 2	stat	pval	pval_corr	reject
Allemagne	Autriche	-0.1108	0.9123	1	False
Allemagne	Belgique	-0.7883	0.4352	1	False
Allemagne	France	17.4405	0	0	True
Autriche	Belgique	-0,7282	0,4702	1	False
Autriche	France	18,9651	0	0	True
Belgique	France	18,4831	0	0	True

Il est assez évident qu'il y a toujours un reject quand l'un des deux pays est la France, j'en conclu qu'il y a la différence pour ce pays

Comme ce test est trop conservateur, je vais me fier plûtot au test de Tukey. En effet le test de Tukey permet un contrôle plus précis des erreurs

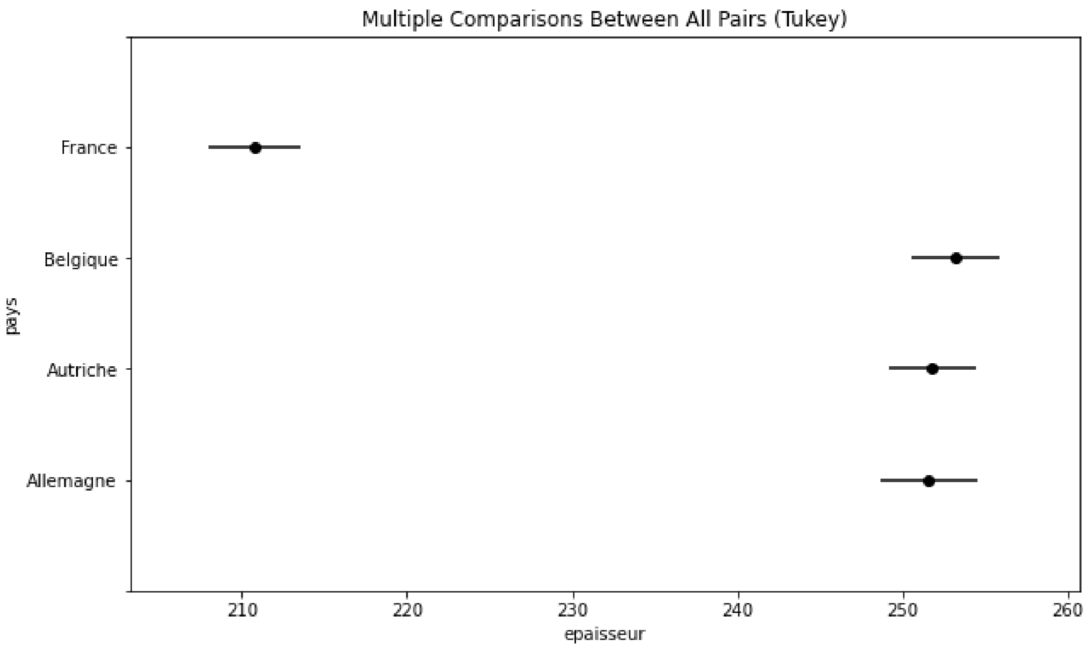
Tukey

On obtient un tableau comme suit que j'ai retrancrie:

Group 1	Group 2	meandiff	p-adj	lower	upper	reject
Allemagne	Autriche	0,2084	0,9997	-5,367	5,7839	False
Allemagne	Belgique	1,5858	0,8841	-4,0933	7,265	False
Allemagne	France	-40,8589	0,0	-46,5961	,35-1216	True
Autriche	Belgique	1,3774	0,9037	-3,9154	6,6702	False
Autriche	France	-41,0673	0,0	-46,4224	-35,7122	True
Belgique	France	-42,4447	0,0	-47,9077	-36,9817	True

J'en conclu la même chose que pour Bonferroni

Graphiquement



Il est aussi évident qu'il s'agit de la France qui est différente.

En conclusion grace aux tests effectués je considère que l'épaisseur des timbres en France est différente de celle de l'Autriche, l'Allemagne et la Belgique.

Vérification

Je vais refaire le test d'ANOVA en enlevant toute les données concerant l'épaisseur des timbres en France.

1) Recolte des données

J'ai crée le jeu de données **timbreswfrance** pour avoir les données de l'épaisseur des timbres sans celles de la France

2) Hypothèses

H_0 : $\mu_{Al} = \mu_{Au} = \mu_{Be}$

H_1 : au moins une moyenne est différente des autres

3) Seuil α

Nous allons prendre α = 5%

4) Choix du test

Nous allons effectuer le test de ANOVA car on souhaite comparer les moyennes d’une variable aléatoire quantitative dans 3 groupes (variable qualitative), il suffit de vérifier que les conditions d’applications soient satisfaites :

5) Conditions du test

- L’échantillon des résidus est composé d’observations indépendantes. OK par le protocole de récolte des données.
- Homoscédasticité : la variance des résidus est la même pour tous les groupes. ?
- $\epsilon \sim$ N (0,σ) où σ^2 est la variance commune des résidus.?

6) Vérification des conditions du test

Conditions d’applications : test de normalité

Hypothèses à tester

H_0 : l’échantillon est issu d’une population normalement distribuée.

H_1 : l’échantillon n’est pas issu d’une population normalement distribuée.

Seuil α

je prends α=5 %

test sur les residus

conditions d'applications

L’échantillon doit être composé d’observations indépendantes d’une variable quantitative continue. OK par le protocole de récolte des données.

Récolte des données

Je crée l'objet **model2** qui permet de récolter les données de l'épaisseur des timbre en fonction du pays

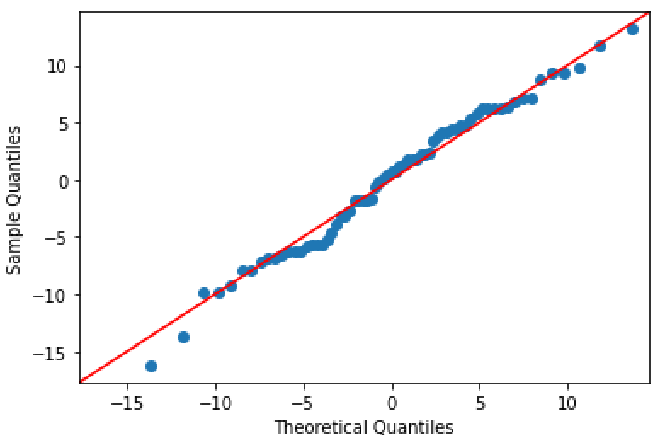
Statistique de décision

j'obtient une p-value de 0.5049103498458862

Comparaison avec α

p-value > α

QQ-plot



Les points sont alignée le long de la droite

Je peux donc considerer l'hypothèse de normalité

Conditions d’applications : Variance, Test de Bartlett :

Hypothèses

$H_0 : \sigma^2_{Al} = \sigma^2_{Au} = \sigma^2_{Be}$

H_1 : au moins une variance est différente des autres.

Choix du seuil α

Je prends α = 5%

test de Barlett

Conditions d'application:

observations indépendantes. OK par le protocole de récolte des données.

normalité dans chaque groupe (résidus). OK

Statistique de décision

j'utlise la commande stats.barlett Et j'obtiens que la statistique de décision est égale à 0.3153574166397785 . La p-value est 0.8541241608749621

Comparaison avec le seuil α

On a p-value > α je peux considerer qu'il y a l'égalité des variances

7) Statistique de décision

J'utlise la commande **sm.stats.anova_lm** Pour faire ma statistque de décision

J'obtiens que la p-value vaut 0.671005

8) Comparaison avec α

On a p-value > α donc on ne peut pas rejeter H_0.

Ainsi cela confirme l'hypotèse de départ