Projet de l'UE Statistique avec Python

Léo GAUTHIER

On travaillera au seuil de validité $\alpha = 5\%$ pour les tests d'intérêt et techniques.

Exercice 1 : Indépendance

Pour déterminer si la couleur des cheveux est indépendante du sexe dans ce district écossais, il faut utiliser le test du Khi-2 d'indépendance de Pearson. L'hypothèse nulle est H_0 : la couleur des cheveux et le sexe sont des variables indépendantes, et l'hypothèse alternative H_1 : les 2 variables ne sont pas indépendantes. Ce test possède des conditions d'application, qu'on doit absolument vérifier avant de faire le test :

1. Les observations sont indépendantes.
2. Les classes de variables sont exclusives.
3. La règle de Cochran est vérifiée (au moins 80% des effectifs théoriques sont supérieurs à 5).
4. La taille de l'échantillon est suffisante.

La première condition relève du procédé d'échantillonnage, on ne peut pas vraiment la vérifier ici.
La seconde est vérifiée, en effet il n'y pas de répétitions de même couleur de cheveux ou de sexe dans les différentes colonnes ou lignes du tableau.
On vérifie la troisième condition soit avec une boucle dans le code, soit en calculant immédiatement les effectifs théoriques :

	BLOND	ROUX	CHATAIN	BRUN	NOIR DE JAIS
GARCON	614.37033222	116.81689415	825.28972444	516.48210141	27.04094772
FILLE	521.62966778	99.18310585	700.71027556	438.51789853	22.95905228

Il n'y a aucun effectif théorique inférieur ou égal à 5, donc la règle de Cochran est vérifiée.
Avec un échantillon de 3 883 individus, et comme la règle de Cochran est vérifiée, on peut estimer que la taille de l'échantillon est suffisante.
On peut donc faire le test du Khi-2. La statistique χ^2_{obs} du test vaut 10,47 et suit, sous H_0, une loi du χ^2. On détermine le nombre de degrés de liberté avec le nombre de classe des variables. Théoriquement, sous H_0, la statistique χ^2_{obs} va suivre une loi du χ^2 à $(k-1)(n-1)$ degrés de liberté, avec k et n le nombre de classe des variables. Ici, on a $k = 2$ et $n = 5$.

Donc dans ce cas, la statistique du test va suivre une loi du χ^2 , sous H_0, à 4 degrés de liberté.
On a une p-valeur $p = 0,0332$. Puisque $p < \alpha$, on accepte H_1 : la couleur des cheveux et le sexe ne sont pas des variables indépendantes.
La p-valeur étant inférieure mais relativement proche du seuil α, on peut avoir peur d'un risque de 1ère espèce, accepter H_1 alors que H_0 est vraie.
On sait également que le test du Khi-2 de Pearson implique une approximation du ratio de vraisemblance, ce qui rend moins précis le résultat. Il existe un test alternatif (pour les mêmes hypothèses H_0 et H_1) à ce test : le G-Test, qui utilise la valeur exacte du ratio de vraisemblance.
Le G-test a les mêmes conditions d'application que le test du Khi-2, et sous H_0, la statistique G du test va suivre une loi du χ^2 avec le même nombre de degrés de liberté.
Dans l'exercice, $G = 10,76$ et la p-valeur $p_g = 0,0295$. On aboutit à la même conclusion, en réduisant le risque de 1ère espèce car $p_g < p < \alpha$.
On peut donc conclure que la couleur de cheveux et le sexe ne sont pas des variables indépendantes.

Exercice 2 : Test de Student

Pour pouvoir répondre à la question d'Alice, on se demande si il y a une différence significative, qui ne peut être expliquée par l'aléa d'échantillonnage, entre le poids moyen des Maneles et Manalas. On travaille avec 2 groupes, donc on utilise le test de Student classique.
Il y a 3 données manquantes dans l'échantillon des Manalas. On a alors $n_1 = 27$ et $n_2 = 24$. Avant de passer au test, on va calculer les statistiques descriptives du jeu de données :

	Nb d'obs	Moyenne	Ecart-type	Min	Max	1er quartile	Médiane	3ème quartile
Manele	27	110,39	6,70	96,62	124,27	106,43	110,40	114,87
Manala	24	112,38	8,12	94,41	131,22	106,74	111,22	117,19

Le test de Student possède évidemment des conditions d'application :

1. Les observations au sein des échantillons sont indépendantes.
2. Les échantillons sont indépendants.
3. Les échantillons 1 et 2 sont distribués selon des lois normales.
4. Les écarts types sont inconnus.
5. Les écarts types sont égaux.

Les 2 premières conditions dépendent encore du procédé de récolte des données, donc on ne peut pas vraiment les vérifier ici.
La troisième condition est importante car on a un petit échantillon, uniquement 27 observations. Plus la taille de l'échantillon est grande, moins la

normalité des échantillons est un problème.
Pour pouvoir vérifier cette condition, on utilise le test de Shapiro-Wilk, qui est le test de normalité le plus puissant pour les petits échantillons. Ce test a 2 conditions d'application : l'indépendance des observations et l'indépendance entre échantillons. On a supposé que ces deux conditions étaient vérifiées, donc on peut faire le test pour chaque échantillon. On teste H_0 : l'échantillon est distribué selon une loi normale contre H_1 : l'échantillon n'est pas distribué selon une loi normale.
Pour les Maneles, on a la statistique de test $SW_e = 0,99$, et la p-valeur $p_e = 0,9971$. Puisque $p_e > \alpha$ et est proche de 1, on accepte H_0 avec un petit risque de 2ème espèce (accepter H_0 alors que H_1 est vraie). L'échantillon concernant les Maneles est donc distribué selon une loi normale.
On fait la même chose avec l'échantillon des Manalas : on a $SW_a = 0,96$ et $p_a = 0,5003$. Comme $p_a > \alpha$, on accepte H_0 avec un petit risque de 2ème espèce : l'échantillon concernant les Manalas est distribué selon une loi normale. La troisième condition d'application est ainsi vérifiée.
La quatrième est évidemment vérifiée : la moyenne étant inconnue, on ne peut pas connaître l'écart type.
Pour vérifier la cinquième condition, il faut utiliser un test d'égalité des variances de Fisher-Snedecor. Cette condition est très importante pour le test de Student. Le test de Fisher-Snedecor vient tester $H_0 : \sigma_e^2 = \sigma_a^2$ contre $H_1 : \sigma_e^2 \neq \sigma_a^2$. Il a 4 conditions d'application : l'indépendance des observations, des échantillons, et les 2 échantillons doivent être distribués selon des lois normales. Ces 4 conditions ont été vérifié pour le test de Student, donc on peut réaliser le test de Fisher : sous H_0, la statistique de test F va suivre une loi de Fisher à $(n_1 - 1), (n_2 - 2)$ degrés de liberté, où n_1 et n_2 sont les tailles d'échantillon des deux groupes. Dans le cadre de notre jeu de données, on a $n_1 = 27$, $n_2 = 24$, et $F = 0,68$. Sous H_0, $F \sim \mathcal{F}(26, 23)$ et on a une p-valeur $p = 0,8293 > \alpha$, donc on accepte $H_0 : \sigma_e^2 = \sigma_a^2$, avec un petit risque de 2ème espèce car p est proche de 1.
Les conditions d'application pour le test de Student étant vérifiées, on peut faire le test. On teste $H_0 : \mu_a = \mu_e$ contre $H_1 : \mu_a \neq \mu_e$.
On décide de faire un test bilatéral car Alice se demande si il ya une différence de poids entre les Maneles et Manalas, non pas si l'un est supérieur à l'autre. Sous H_0, la statistique de test T va suivre une loi de Student à $(n_1 + n_2 - 2)$ degrés de liberté.
Dans l'exercice, on a $n_1 = 27$, $n_2 = 24$, $T = -0,96$ et T va suivre sous H_0 une loi de Student à 49 degrés de liberté. La p-valeur est $p = 0,3402 > \alpha$ donc on accepte $H_0 : \mu_a = \mu_e$, il n'y a donc pas de différence significative entre le poids moyen des Maneles et des Manalas.

Exercice 3 : ANOVA

La question de Rémi nous fait utiliser une méthode proche de celle qu'on a utilisé pour répondre à la question d'Alice, mais on a, cette fois-ci, non pas 2 mais 4 groupes. On va donc utiliser l'ANOVA (Analysis of Variance) à 1 facteur. On a $n = 89$ observations (et aucune manquante !), une variable quantitative (l'épaisseur des timbres) et une variable de groupe (le pays d'origine) à $k = 4$ niveaux.
Avant de commencer le modèle et de vérifier ses conditions d'application, on regarde les statistiques descriptives de l'épaisseur, qui est la seule variable quantitative de ce jeu de données :

	Nb d'obs	Moyenne	Ecart-type	Min	Max	1er quartile	Médiane	3ème quartile
Allemagne	19	251,63	6,05	238	261	246,5	252	256
Autriche	25	251,84	6,28	242	265	246	252	258
Belgique	23	253,22	6,83	237	265	247,5	254	258
France	22	210,77	8,52	196	230	205,25	211	216

On remarque immédiatement une différence importante de la moyenne observée de l'épaisseur des timbres entre les timbres français et les timbres des 3 autres pays. L'objectif de cette analyse va donc être d'identifier si cette différence est significative ou non.
L'ANOVA est, contrairement au test de Student, un modèle. On va modéliser, $\forall 1 \leq i \leq 4$, les variables d'échantillons X_i sous la forme $X_i = \mu_i + \epsilon_i$, où μ_i est la moyenne du groupe i et les ϵ_i, résidus du modèle, suivent une $\mathcal{N}(0, \sigma_i)$.
L'objectif est de tester $H_0 : \mu_{All} = \mu_{Au} = \mu_B = \mu_F$ contre H_1 : au moins une moyenne est différente des autres.
L'ANOVA a 4 conditions d'application :

1. Les observations au sein des échantillons sont indépendantes.
2. Les échantillons sont indépendants.
3. Les écarts types σ_i sont tous égaux.
4. Chaque échantillon est distribué selon une loi normale.

La quatrième condition implique de faire 4 tests de normalité, ce qui est assez long. On peut donc vérifier les conditions d'application sur les résidus du modèle :

1. L'échantillon d'observation des résidus est indépendant.
2. Les écarts-types σ_i des résidus sont les mêmes (homoscédasticité).
3. Les ϵ_i sont distribués selon une même loi normale $\mathcal{N}(0, \sigma^2)$, σ^2 étant la variance commune des ϵ_i.

Pour la première condition, cela relève encore du procédé de récolte des données qui n'est pas vérifiable.

On doit vérifier la troisième condition avant la deuxième car le test de Bartlett, qu'on utilise pour cette dernière, a besoin entre autre de la normalité des échantillons. On sait également que le test de l'ANOVA n'est absolument pas robuste face à l'hétéroscédasticité, donc cette condition est primordiale.
Donc pour la troisième condition, on réalise un test de Shapiro-Wilk sur les résidus. La condition d'application du test de Shapiro-Wilk est l'indépendance de l'échantillon des observations, qu'on a supposé vérifié. On va donc tester H_0 : l'échantillon est distribué selon une loi normale contre H_1 l'échantillon n'est pas distribué selon une loi normale.
Dans le cadre de notre modèle, on a une statistique de test $SW_r = 0,99$ et une p-valeur $p_r = 0,7197 > \alpha$, relativement proche de 1 donc on est pas trop sujet à un gros risque de 2ème espèce. On ne rejette donc pas H_0 : l'échantillon d'observation des résidus du modèle est distribué selon une loi normale.
On peut désormais vérifier l'égalité des variances avec le test de Bartlett. Ses conditions sont l'indépendance des observations et entre échantillon, et la normalité des résidus.
La première condition a été supposée vérifiée au début, et la deuxième vient d'être vérifiée avec le test de Shapiro-Wilk. Le test de Bartlett teste donc $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ contre H_1 : au moins une des variances est différente. Sous H_0, la statistique de test B suit asymptotiquement une loi du χ^2 à $(k-1)$ degrés de liberté.
Dans le cadre de la question de Rémi, on a $k = 4$ niveaux, la statistique de décision vaut $B = 3,05$ et, sous H_0, suit asymptotiquement une χ^2 à 3 degrés de liberté. La p-valeur vaut $p_b = 0,3838 > \alpha$, donc on accepte H_0 : toutes les variances σ_k^2 sont égales à une variance commune σ^2.
On a montré que les ϵ_k suivent tous une $\mathcal{N}(0, \sigma)$, ils sont donc égaux en loi à un certain $\epsilon \sim \mathcal{N}(0, \sigma)$
Maintenant que toutes les conditions d'application pour l'ANOVA sont vérifiées, on peut passer au test de $H_0 : \mu_{All} = \mu_{Au} = \mu_B = \mu_F$ contre H_1 : au moins une moyenne est différente des autres. La statistique de décision du test de l'ANOVA, notée F, suit sous H_0 une loi de Fisher à $(k-1, n-k)$ degrés de liberté, avec n la taille de l'effectif.
Dans notre cas, $k = 4, n = 89$, on a $F = 194,62$ et la p-valeur $p = 5,82e-38 << \alpha$, donc on accepte H_1 avec quasiment aucun risque de se tromper puisque p est très très proche de 0 : il y a au moins une moyenne significativement différente des autres.
La question est maintenant : quel(s) groupe(s) a une moyenne significativement différente des autres ?
Il est assez clair dans l'énoncé qu'on a un groupe de référence : celui des timbres français. Rémi soupçonne déjà les timbres français d'être moins épais

que ceux de nos voisins, et les statistiques descriptives qu'on avait faites au début semblent aller dans son sens. On a donc 3 comparaisons avec des tests de Student à réaliser.
On a l'indépendance au sein des observations et entre échantillons, chaque échantillon est distribué selon une loi normale, on ne connaît pas les écarts-types et ceux-ci sont tous égaux entre eux, donc les conditions d'application du test de Student sont vérifiées.
Il faut cependant prendre en compte le fait que chaque échantillon est utilisé plusieurs fois, et donc corriger la p-valeur. Pour se faire, on devrait utiliser théoriquement la correction de Dunnett mais l'implémentation en Python est trop complexe. On utilise donc la correction de Bonferroni même si on sait que celle-ci est trop conservatrice.
On va se placer dans le cadre de comparaisons multiples mais sans groupe de référence, et on va récupérer les p-valeurs qui nous intéressent (pas les p-valeurs corrigées), les multiplier par le nombre de tests réellement fait (ici 3) et on gardera $min(1, (3p - valeur))$.
On teste donc $H_0 : \mu_F = \mu_{All}$ contre $H_1 : \mu_F \neq \mu_{All}$, puis pareil entre la France et l'Autriche et entre la France et la Belgique.
Les 3 p-valeurs obtenues avec la correction de Bonferroni sont arrondies à 0, donc les 3 p-valeurs corrigées qu'on a sont nulles (ou très proches de 0). En les comparant toutes les 3 au seuil $\alpha = 0,05$ fixé au début du projet, on accepte H_1 pour nos trois tests : la moyenne de l'épaisseur des timbres français est significativement différente de la moyenne de l'épaisseur des timbres allemands, autrichiens ou belges.
On peut donc conclure que l'impression initiale de Rémi concernant la différence d'épaisseur des timbres français était correcte.
Finalement, on va faire des comparaisons sans groupe de référence afin de savoir quels sont les origines des timbres dont l'épaisseur moyenne est significativement différente, sans se baser sur le groupe de référence français.
On va avoir 6 comparaisons à faire.
On a 6 p-valeurs corrigées à l'aide de la correction de Tukey (qui est moins conservatrice que la correction de Bonferroni) : pour la comparaison Allemagne / Autriche, $p_{allau} = 0,997 > \alpha$, donc il n'y a pas de différence significative entre la moyenne de l'épaisseur des timbres autrichiens et allemands. Pour la comparaison Allemagne / Belgique, $p_{allb} = 0,8841 > \alpha$, donc il n'y a pas de différence significative entre l'épaisseur moyenne des timbres belges et allemands. Pour celle Allemagne / France, on retrouvve $p_{allf} = 0,0 < \alpha$, donc il y a une différence significative entre les deux moyennes d'épaisseurs de timbres. Pour la comparaison Autriche / Belgique, on a $p_{aub} = 0,9037 > \alpha$, donc il n'y a pas de différence significative entre les deux moyennes d'épaisseurs. Pour Autriche / France, on retrouve aussi $p_{auf} = 0,0 < \alpha$,

donc il y a une différence significative d'épaisseurs moyennes. Et enfin, pour la dernière comparaison Belgique / France, on retrouve $p_{bf} = 0,0 < \alpha$, donc il y a bien une différence significative entre les deux épaisseurs moyennes.
On retrouve les mêmes résultats que pour la comparaison avec le groupe français pour référence, et on a des résultats plus détaillés.
Cela nous permet donc d'apporter une réponse complète à l'interrogation de Rémi : il y a une différence significative d'épaisseurs entre les timbres français et ceux autrichiens, allemands et belges. Il n'y a, en revanche, pas de différence significative d'épaisseurs entre les timbres autrichiens, belges et allemands !