FACULTÉ DE MATHÉMATIQUE ET D' INFORMATIQUE
UNIVERSITÉ DE STRASBOURG

Projet SAS

SCHITTLY Camille

Professeur : POULIN Nicolas

Strasbourg, 2023 - 2024

Table des matières

Exercice 1

Nous disposons d'un échantillon composé de garçons et de filles provenant d'un district écossais. Cet échantillon est constitué d'observations indépendantes, résumées dans un tableau. Notre objectif est de déterminer si la couleur des cheveux est liée au sexe. Nous cherchons ainsi à établir l'indépendance entre ces deux variables qualitatives. Pour ce faire, nous utilisons un test du χ^2 (khi-deux) d'indépendance.

Voici les hypothèses nulle et alernative :

$$\begin{cases} H_0 : & \text{la couleur des cheveux est indépendante du sexe} \\ H_1 : & \text{la couleur des cheveux dépend du sexe} \end{cases}$$

Nous fixons le seuil de signification à $\alpha = 0.05$ et nous appliquons la statistique de décision :

$$\chi^2_{obs} = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{(n_{ij} - t_{ij})^2}{t_{ij}} \tag{0.1}$$

Sous H_0, la statistique de test suit une loi du χ^2 avec $v = (k-1) \times (c-1)$ degrés de liberté, où k et c représentent respectivement le nombre de classes des deux variables. Dans notre cas, $v = 4$ (car $k = 4$ et $c = 2$).

Ceci est valide lorsque les conditions suivantes sont respectées :

1. Les individus de l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).
2. Les classes des variables sont exclusives.
3. Règle de Cochran : au moins 80% des effectifs théoriques sont supérieurs ou égaux à 5.
4. La taille de l'échantillon est assez grande.

Vérification des conditions d'application pour le test χ^2 :

Nous allons vérifier ces conditions afin de pouvoir appliquer le test de χ^2

1. Les individus composant l'échantillon ont été choisis de manière aléatoire, comme indiqué dans l'énoncé : "Cet échantillon est composé d'observations indépendantes".
2. Les classes des variables sont exclusives, comme précisé dans l'énoncé.
3. En ce qui concerne la règle de Cochran : La règle de Cochran stipule que pour chaque cellule de fréquence théorique, au moins 80% des effectifs théoriques doivent être supérieurs ou égaux à 5 pour que le test χ^2 soit valide.

Pour ce faire on utilise la procédure FREQ afin d'obtenir le tableau suivant :

Fréquence
Attendu
Pourcentage
Pct de ligne
Pct de col.

SEXE	CHEVEUX					
	BLOND	**ROUX**	**CHATAIN**	**BRUN**	**NOIR_DE_**	Total
GARCON	592	119	849	504	36	2100
	614.37	116.82	825.29	516.48	27.041	
	15.25	3.06	21.86	12.98	0.93	54.08
	28.19	5.67	40.43	24.00	1.71	
	52.11	55.09	55.64	52.77	72.00	
FILLE	544	97	677	451	14	1783
	521.63	99.183	700.71	438.52	22.959	
	14.01	2.50	17.43	11.61	0.36	45.92
	30.51	5.44	37.97	25.29	0.79	
	47.89	44.91	44.36	47.23	28	
Total	1136	216	1526	955	50	3883
	29.26	5.56	39.30	24.59	1.29	10.00

Table 0.1 – Table de SEXE par CHEVEUX

Chaque cellule du tableau contient 5 lignes :

1. L'effectif $n_{i,j}$
2. L'effectif théorique $t_{i,j}$ 0.2
3. Le pourcentage correspondant $f_{i,j} = \frac{n_{i,j}}{N}$
4. Le pourcentage en ligne $\frac{n_{i,j}}{n_{i,.}}$
5. Le pourcentage en colonne $\frac{n_{i,j}}{n_{.,j}}$

Grâce à ce tableau nous allons regarder si la règle Cochran est vérifiée.

Tout d'abord nous pouvons constater que l'effectif total est de $n = 3883$.

Nous avons avons besoin des effectifs théoriques dont la formule se présente comme suit

$$t_{ij} = \frac{n_{i.} \times n_{.j}}{n} \tag{0.2}$$

avec t_{ij} l'effectif théorique pour la ligne i et la colonne j.

$n_{i.}$ la somme des effectifs observés pour la ligne i.

$n_{.j}$ la somme des effectifs observés pour la colonne j.

Nous pouvons régarder si la règle de Cochran est vérifiée grace à la ligne "Attendu" du tableau.

Nous pouvons ainsi constater que tous les effectifs théoriques sont supérieurs ou égaux à 5. La règle de Cochran est donc vérifié.

4. La taille de l'échantillon est $n = 3883$ ce qui est suffisamment grand .

Test du χ^2

Maintenant que les conditions d'application sont vérifiées, nous pouvons utiliser le test du χ^2.

Statistique	**DDL**	**Valeur**	**Prob**
Khi-2	4	10.4674	0.0332
Test du rapport de vraisemblance	4	10.7555	0.0295
Khi-2 de Mantel-Haenszel	1	1.7216	0.1895
Coefficient Phi	-	0.0519	-
Coefficient de contingence	-	0.0519	-
V de Cramer	-	0.0519	-
Taille de l'échantillon = 3883			

Table 0.2 – Résultats du test de chi-deux pour la table de SEXE par CHEVEUX

On regarde la première ligne du tableau pour effectuer le test.

Nous avons $v = 4$ degrés de liberté pour notre test comme dit précédemment. La valeur de la statistique de test du χ^2 est 10.4674 avec une p-value de 0.0332, ce qui est inférieur à notre seuil de signification $\alpha = 0.05$. Par conséquent, nous rejetons H_0 et concluons en faveur de H_1 mais on prend le risque d'avoir une erreur de type 1. Ainsi, il existe une association statistiquement significative entre la couleur des cheveux et le sexe dans notre échantillon.

Choix du test utilisé :

Voici les avantage et les inconvénients du test du χ^2

Avantages du test du χ^2 :

Le test du χ^2 présente plusieurs avantages dans notre cas. Tout d'abord, il est particulièrement adapté à notre situation. Sa mise en œuvre est simple et les résultats obtenus sont facilement interprétables, ce qui facilite l'analyse. De plus, ce test est particulièrement privilégié lorsque l'échantillon est de grande taille et que les fréquences attendues dans le tableau de contingence sont élevées, ce qui est justement le cas dans notre étude.

Inconvénients du test du χ^2 :

Le test du χ^2 présente plusieurs inconvénients. Tout d'abord, il n'est pas adapté aux données continues. De plus, il présente des limitations lorsque l'échantillon est de petite taille, notamment lorsque les fréquences attendues dans le tableau

de contingence sont faibles. Dans ce cas, le test d'indépendance exact de Fisher peut être préférable.

Voici les autres tests possibles en cas d'autre situations.

Autres tests possibles : L'indépendance exacte de Fisher est utilisée lorsque les conditions d'application du test du χ^2 ne sont pas satisfaites, plus précisément quand la condition de Cochran n'est pas vérifiée ou pour de petits échantillons.

Ici ce test ne me semble pas adéquat car les conditions sont satisfaites.

Alternative

Nous pouvons utiliser le G-test, en effet le test du χ^2 de Pearson est basé sur une approximation d'un ratio de log-vraisemblance de même 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.

Les hypothèses nulle et alternative sont les mêmes que formulé précédemment avec les test du χ^2 de même on fixe toujours le seuil α à 0.05.

Comme les hypothèses et conditions d'application du G-test sont les mêmes que celles du test du χ^2 mais sans la limite de la taille d'échantillon nous pouvons appliquer le G-test.

Il s'agit de la deuxième ligne de la table 0.2.

On a $p - value = 0.0295$ ce qui est aussi plus petit que le seuil α. On rejette donc H_0 en faveur de H_1 et on en conclut le même résultat que pour le test du χ^2 .

Exercice 2

Une entreprise pharmaceutique expérimente une nouvelle molécule censée faire baisser une mesure physiologique pour des patients porteurs d'une pathologie. Pour cela, 54 patients sans lien de parenté ont été recrutés. Chaque patient s'est vu attribuer soit la nouvelle molécule (traitement A) soit un placebo (traitement B). Les groupes ont été construits de telle manière à ce que les distributions de l'âge et du sexe soient similaires dans les deux groupes.

Statistiques descriptive

Tout d'abord, examinons les statistiques descriptives des traitements *A* et *B* avec la procédure MEANS :

Traitement	N obs	N	Moyenne	Ec-type	Minimum	Maximum
A	27	27	78.836	11.141	51.040	99.270
B	27	27	65.751	12.416	41.090	91.270

Table 0.3 – Statistiques descriptives des traitements A et B pour la variable d'analyse "mesure".

On remarque que toute les observations sont utilisées, il n'y pas de valeurs manquantes. On remarque également une différence entre les moyennes des mesures des traitements A et B, avec respectivement 78,8362 pour le traitement A et 65,7514 pour le traitement B. De plus, les écarts types des mesures sont à peu près similaires pour les deux traitements, soit 11,1411 pour le traitement A et 12,4159 pour le traitement B.

La procédure UNIVARIATE nous offre plus d'information concernant les statistique descriptive.

Procédure UNIVARIATE pour le traitement A avec la variable "mesure"	
Moments	Valeurs
N	27
Somme des poids	27
Moyenne	78.8362963
Somme des observations	2128.58
Écart-type	11.1411134
Variance	124.124409
Skewness	-0.4583407
Kurtosis	0.45869431
Somme des carrés non corrigée	171036.598
Somme des carrés corrigée	3227.23463
Coeff Variation	14.1319595
Std Error Mean	2.14410828

Table 0.4 – Résultats de la procédure UNIVARIATE pour le traitement A avec la variable "mesure".

Mesures statistiques de base pour le traitement A avec la variable "mesure"	
Location	
Moyenne	78.83630
Médiane	78.22000
Mode	.
Variabilité	
Écart-type	11.14111
Variance	124.12441
Intervalle	48.23000
Écart interquartile	13.18000

Table 0.5 – Mesures statistiques de base pour le traitement A avec la variable "mesure".

Procédure UNIVARIATE pour le traitement B avec la variable "mesure"	
Moments	Valeurs
N	27
Somme des poids	27
Moyenne	65.7514815
Somme des observations	1775.29
Écart-type	12.4159766
Variance	154.156475
Skewness	0.05696282
Kurtosis	-0.0708065
Somme des carrés non corrigée	120736.016
Somme des carrés corrigée	4008.06834
Coeff Variation	18.8831891
Std Error Mean	2.38945581

Table 0.6 – Résultats de la procédure UNIVARIATE pour le traitement B avec la variable "mesure".

Mesures statistiques de base pour le traitement B avec la variable "mesure"	
Location	
Moyenne	65.75148
Médiane	64.73000
Mode	.
Variabilité	
Écart-type	12.41598
Variance	154.15647
Intervalle	50.18000
Écart interquartile	14.43000

Table 0.7 – Mesures statistiques de base pour le traitement B avec la variable "mesure".

Afin d'avoir l'aspect visuel, nous pouvons regarder les boxplots comme suit :

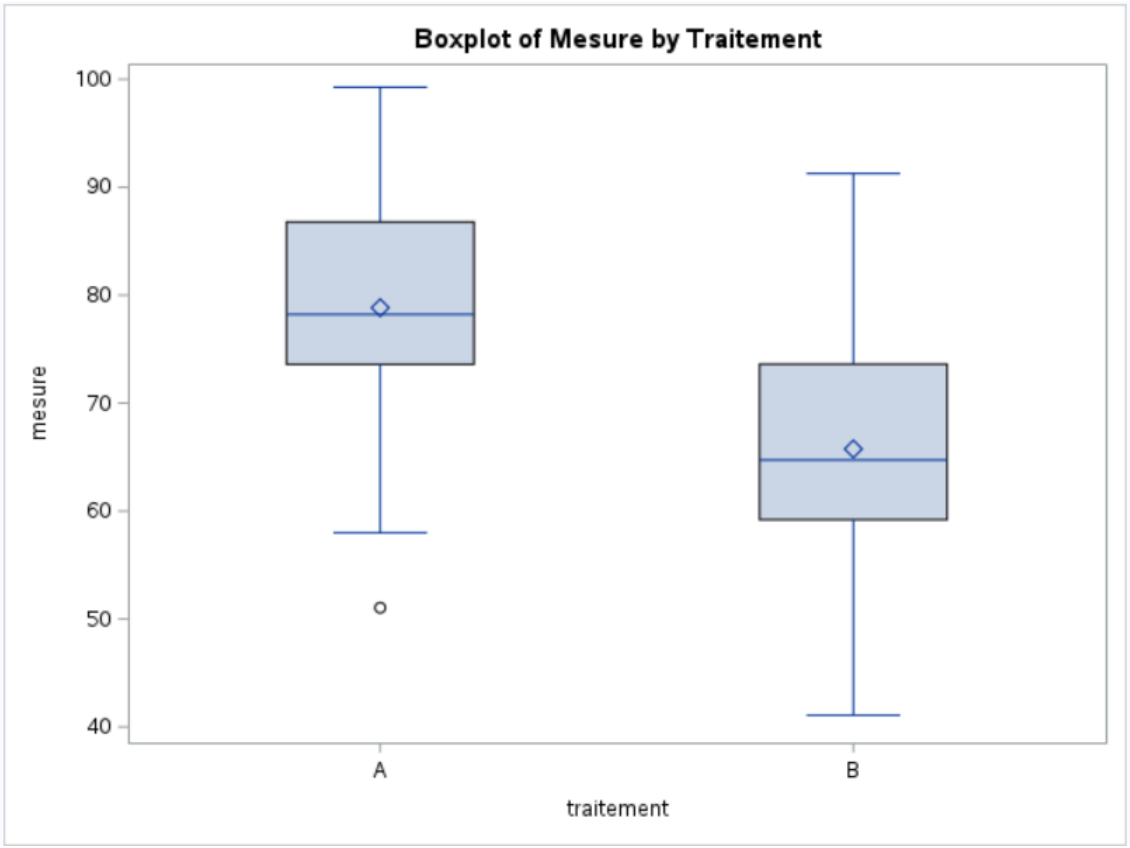


Figure 1 – Boxplot

Maintenant que nous avons effectué les statistique descriptives, nous pouvons choisir notre test.

Nous allons en effet utiliser le test de Student pour comparer les moyennes des mesures entre les groupes A et B, afin de déterminer s'il existe une différence significative.

Dans la suite nous allons utiliser les notations suivantes.

Pour notre test de Student, nous posons :

— X : variable quantitative continue (ici, la mesure).
— Y : variable qualitative à 2 niveaux (marqueur de groupe). Ici, Y est la variable traitement avec A ou B.
— X_1 : mesure pour le traitement A.
— X_2 : mesure pour le traitement B.
— n_1 : nombre d'individus du groupe A.
— n_2 : nombre d'individus du groupe B.
— μ_1 : moyenne de X_1.
— μ_2 : moyenne de X_2.
— σ_1 : écart-type de X_1.
— σ_2 : écart-type de X_2.

Voici les hypothèse nulle et alternative dans le cadre de notre test de Student :

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Nous prenons un seuil $\alpha = 0.05$.

D'après le cours la statistique de décision du test de Student est :

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_{X_1}^2}{n_1} + \frac{S_{X_2}^2}{n_2}}} \tag{0.3}$$

Sous H_0, T suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté (dans notre cas $n_1 + n_2 - 2 = 52$ avec $n_1 = n_2 = 27$).

Voici les conditions d'application du test de Student :

1. Chaque échantillon est composé d'observations indépendantes.
2. Les deux échantillons sont indépendants.
3. $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$.
4. $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$.
5. σ_1 et σ_2 sont inconnus.
6. $\sigma_1 = \sigma_2$.

Vérification des conditions d'application pour le test de Student :

Afin de pouvoir appliquer le test de Student nous devons vérifier les conditions d'application.

Pour les conditions 1. et 2., le protocole de collecte de données garantit que chaque échantillon est constitué d'observations indépendantes et que les deux échantillons sont indépendants.

Concernant les conditions 3. et 4., nous allons procéder à un test de Shapiro-Wilk sur X_1 puis sur X_2.

Les hypothèses nulle et alternative sont les suivantes :

$$\begin{cases} H_0 : \text{l'échantillon est issu d'une population normalement distribuée} \\ H_1 : \text{l'échantillon n'est pas issu d'une population normalement distribuée} \end{cases}$$

Le seuil choisi est $\alpha = 0.05$.

Une seule condition d'application doit être vérifiée : chaque échantillon doit être constitué d'observations indépendantes d'une variable quantitative continue. Cette condition est confirmée par l'énoncé, tout comme les conditions 1. et 2.

Test de Shapiro-Wilk

Nous effectuons le test de normalité avec la procédure UNIVARIATE sur SAS ce qui nous donne les tableaux suivants.

Test	Statistique	p-value
Shapiro-Wilk	0.978009	0.8149
Kolmogorov-Smirnov	0.100315	>0.1500
Cramer-von Mises	0.036949	>0.2500
Anderson-Darling	0.231747	>0.2500

Table 0.8 – Résultats des tests de normalité pour traitement = A.

Nous regardons le première ligne du tableau concernant le test de Shapiro-Wilk :

Pour X_1, la $p - valeur$ obtenue est de $0.8149 > \alpha$, donc nous conservons H_0 tout en prenant le risque de commettre une erreur de type II. Ainsi, pour X_1, l'échantillon est issu d'une population normalement distribuée. Par conséquent, nous considérons que X_1 suit une loi normale $N(\mu_1, \sigma_1)$.

Test	Statistique	p-value
Shapiro-Wilk	0.9826	0.9157
Kolmogorov-Smirnov	0.100949	>0.1500
Cramer-von Mises	0.033615	>0.2500
Anderson-Darling	0.203486	>0.2500

Table 0.9 – Résultats des tests de normalité pour traitement = B.

Pour X_2, la p-valeur obtenue est de $0.9157 > \alpha$, donc nous conservons H_0 tout en prenant également le risque de commettre une erreur de type II. Ainsi, pour X_2, l'échantillon est issu d'une population normalement distribuée. Par conséquent, nous considérons que X_2 suit une loi normale $N(\mu_2, \sigma_2)$.

Ce qui est intéressant c'est que cela rejoint les statistiques descriptive, en effet Les valeurs d'asymétrie et d'aplatissement sont très proches de 0, ce qui suggère que la distribution des données est symétrique et a une forme similaire à celle d'une distribution normale. Ces résultats indiquent que les données de la variable "mesure" pour le traitement A et B suivent statistiquement une distribution normale.

Juste pour voir ce que cela donne, je vais regarder les QQ-plot :

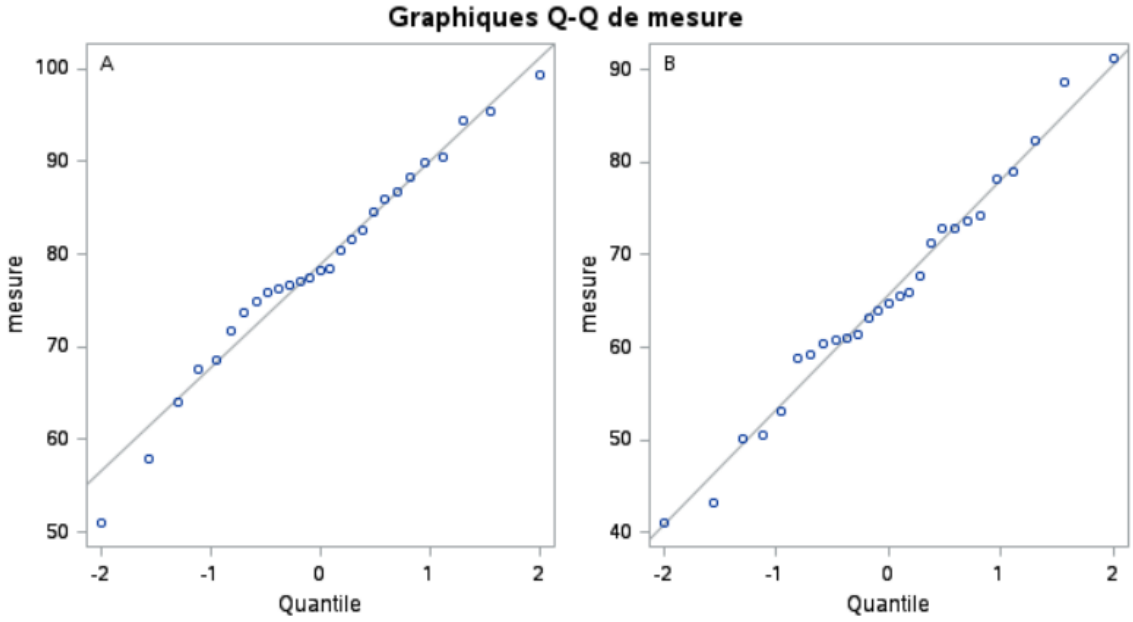


Figure 2 – QQ-plot

Comme nous pouvons le voir, le QQ-plot n'est pas très informatif sur la normalité. En effet il y a trop peu d'observation. En général on utlise le QQ-plot lorsque l'on a beaucoup d'observation et que le test de Student ne permet pas de conserver H_0 ou que l'on souhaite être sûr de pas faire une erreur de type II en conservant H_0.

Choix du test

Avantages : Le test de Shapiro-Wilk est reconnu pour sa capacité à détecter la non-normalité des données, surtout pour des échantillons de taille réduite comme ici.

Inconvénients : Cependant, sa précision diminue et ses résultats peuvent devenir imprévisibles avec de très grands échantillons. De plus, il est conçu uniquement pour évaluer la normalité d'une seule variable à la fois, limitant ainsi son utilité dans des analyses multivariées, ici il n'y avait que deux tests à faire donc ca va.

Néanmoins la procédure UNIVARIATE nous a suggéré d'autre tests dans le tableau comme le Test de Kolmogorov-Smirnov, le Test de Cramer-von Mises ou le Test de Anderson-Darling.

Ici j'ai choisi le test de Shapiro-Wilk car toute les conditions d'application sont vérifiées et que la taille des deux échantillons sont adéquat.

Pour la 5ème condition, nous avons bien σ_1 et σ_2 qui sont inconnus car c'est le cas en pratique.

Pour la dernière condition il faut effectuer un test d'égalité des variances.

Test de Fisher-Snedecor

Ce test est fait automatiquement lorsqu'on fait le test de Student.
Les hypothèses nulle et alternative sont les suivantes suivantes :

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

On prend le seuil $\alpha = 0.05$.
D'après le cours, on utilise la statistique de décision :

$$F = \frac{S_{X_1}^2}{S_{X_2}^2} \tag{0.4}$$

Sous H_0, F suit la loi de Fisher de paramètres $n_1 - 1$ et $n_2 - 1$ avec $n_1 = n_2 = 27$, lorsque les conditions suivantes sont respectées :

1. Chaque échantillon est composé d'observations indépendantes.
2. Les deux échantillons sont indépendants.
3. X_1 suit une loi normale $N(\mu_1, \sigma_1)$.
4. X_2 suit une loi normale $N(\mu_2, \sigma_2)$.

Nous avons déjà démontré que les conditions sont satisfaites. Par conséquent, d'après le test de Fisher-Snedecor, nous avons par la procedure TTEST :

Méthode	DDL num.	DDL den.	Valeur F	Pr > |t|
Folded F	26	26	1.24	0.5846

Table 0.10 – Résultats de l'égalité des variances.

Nous obtenons une $p - valeur$ de $0.5846 > \alpha$, ce qui nous conduit à conserver H_0 tout en prenant le risque d'avoir une erreur de type 2.

Toutes nos hypothèses ont été vérifiée, nous pouvons donc procéder au test de Student.

Test de Student

D'après le test de Student effectué sur SAS, nous avonspar la procédure TTEST :

Méthode	Variances	DDL	Valeur du test t	Pr > |t|
Pooled	Egal	52	4.08	0.0002
Satterthwaite	Non égal	51.401	4.08	0.0002

Table 0.11 – Résultats du test de Student.

En examinant la première ligne du tableau, nous constatons que pour un degré de liberté de 52, la valeur de la statistique de test $T = 4.08$, avec une p-value de $0.0002 < \alpha$. Ainsi, H_0 est rejetée, ce qui nous conduit à accepter statistiquement H_1 en prenant le risque d'effectuer une erreur de type 1.

Ces résultats suggèrent une différence statistiquement significative entre les moyennes des deux molécules. En outre, la moyenne des mesures de la molécule B, un placebo, est inférieure à celle de la molécule A. Ainsi, nous concluons que la molécule testée n'est pas efficace.

Ce qui est par ailleurs intéressant est que cela rejoint les statistique descriptive car la moyenne du placebo est nettement inférieur à celle initiale.

Exercice 3

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. De nombreuses études épidémiologiques ont permis de mettre en évidence l'influence sur la santé de certains composés chimiques comme l'ozone (O_3). Le jeu de données *ozone.xls* comporte 112 observations indépendantes relevées durant l'été 2001.

— La variable *maxO3* est le maximum journalier de la concentration en ozone (en $\mu g/m^3$).
— Les variables *T9*, *T12*, *T15* correspondent aux températures relevées respectivement à 9h, 12h et 15h.
— Les variables *Ne9*, *Ne12*, *Ne15* correspondent aux nébulosités relevées respectivement à 9h, 12h et 15h.
— Les variables *Vx9*, *Vx12*, *Vx15* correspondent à la composante Est-Ouest du vent relevée respectivement à 9h, 12h et 15h.

Notre objectif est d'étudier la relation potentielle entre la variable maxO3 et l'ensemble des autres variables, afin d'identifier un modèle optimal. Pour commencer, nous allons examiner les statistiques descriptives.

Statistiques descriptives :

Afin d'effectuer les statistiques descriptives nous allons utiliser la procédure MEANS.

Table 0.12 – Statistique descriptive par traitement

Variable	Libellé	N	Moyenne	Ec-type	Minimum	Maximum
maxO3		112	90.3035714	28.1872245	42.0	166
T9		112	18.3607143	3.1227257	11.3	27
T12		112	21.5267857	4.0423208	14	33.5
T15		112	22.6276786	4.5308594	14.9	35.5
Ne9		112	4.9285714	2.5949163	0	8
Ne12		112	5.0178571	2.2818601	0	8
Ne15		112	4.8303571	2.3322587	0	8
Vx9		112	−1.2143455	2.6327423	−7.8785	5.1962
Vx12		112	−1.6110036	2.7956729	−7.8785	6.5778
Vx15		112	−1.690683	2.8101977	−9	5

Le nombre d'observation lue et observé sont les mêmes ce qui signifie qu'il n'y a pas de valeurs manquante.

Pertinence du modèle

Tout d'abord, posons le modèle complet :

$$\text{maxO3} = \beta_0 + \beta_1 \cdot T9 + \beta_2 \cdot T12 + \beta_3 \cdot T15 + \beta_4 \cdot Ne9 + \beta_5 \cdot Ne12$$

$$+\beta_6 \cdot Ne15 + \beta_7 \cdot Vx9 + \beta_8 \cdot Vx12 + \beta_9 \cdot Vx15 + \epsilon$$

Il s'agit d'un modèle de régression linéaire multiple, l'équation est généralisée comme suit :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{0.5}$$

où :

— y est la variable dépendante,
— $x_1, x_2, \ldots, x_p$ sont les variables indépendantes,
— β_0 est l'ordonnée à l'origine,
— $\beta_1, \beta_2, \ldots, \beta_p$ sont les coefficients de régression,
— ϵ est le terme d'erreur.

Hypothèses pour le Test de Fisher

$$\begin{cases} H_0 : & \beta_1 = \beta_2 = \cdots = \beta_9 = 0 \\ H_1 : & \text{Au moins un } \beta_i \neq 0 \end{cases}$$

Et on choisi un seuil $\alpha = 0.05$

Hypothèses d'application de régression

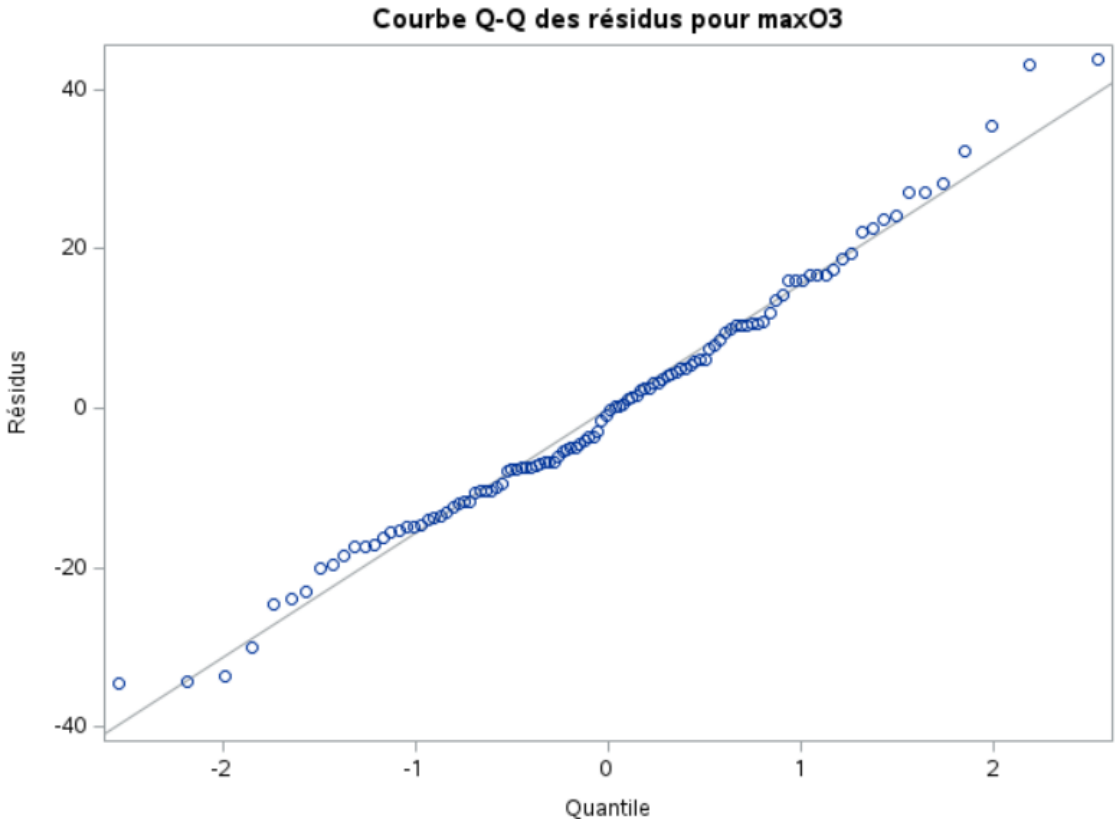
1. Normalité des résidus : Les résidus de la régression doivent suivre une distribution normale.

2. Homoscédasticité des résidus : La variance des résidus doit être constante à travers toutes les valeurs des variables indépendantes.

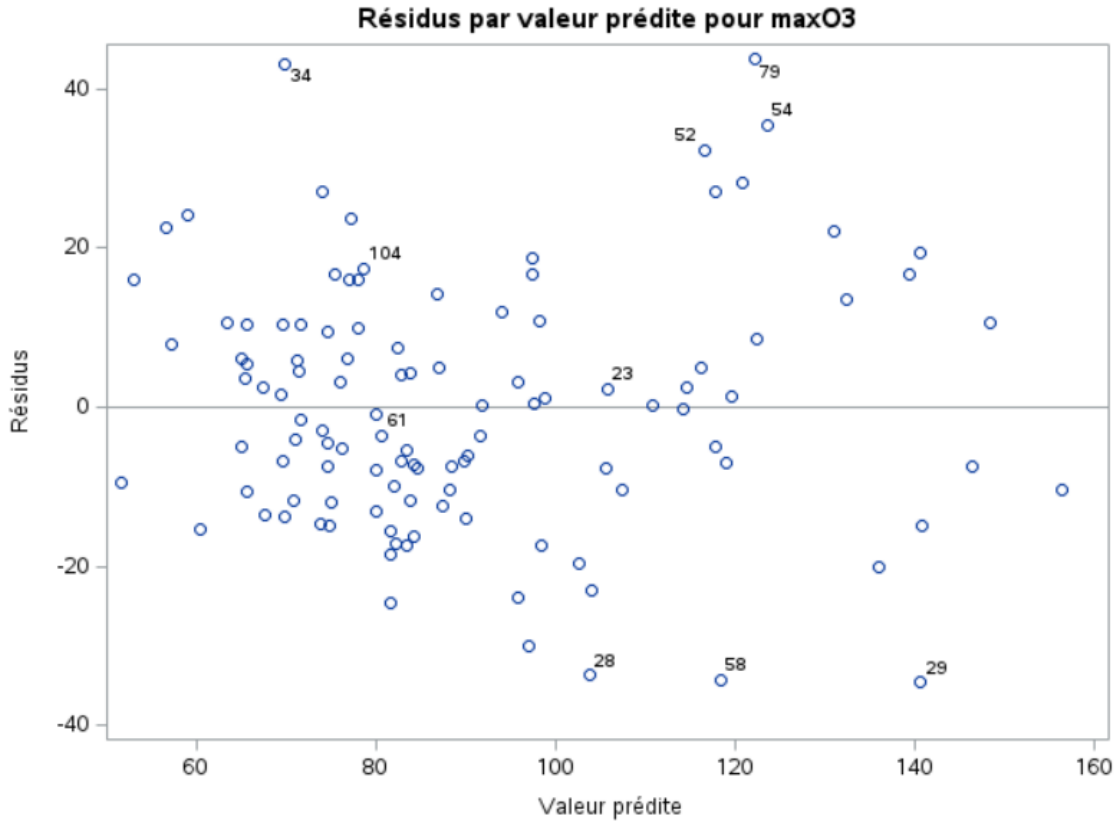
3. Indépendance des résidus : Les résidus doivent être indépendants les uns des autres.

4. Linéarité : La relation entre les variables indépendantes et la variable dépendante doit être linéaire.

Verification des hypothèses



Il y a assez de données pour se fier au QQplot, ici le nuage de points s'aligne bien le long de la droite affine, ce qui suggère que les résidus sont normalement distribués.



Le nuage de point est répartie de façon uniforme autour de 0 pour l'axe des ordonnées ce qui confirme visuellement l'hypothèse de linéarité et d'égalité des variances.

Enfin les observations sont indépendantes d'après le protocoles de récolte des données mais il n'y a pas indépendance des résidus.

Résultat du test

Comme nous avons indépendance des observations et normalité des résidus nous allons appliquer le test de Fisher.

Voici les résultat du test avec la procédure REG

Table 0.13 – Tableau d'analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	9	60905	6767.21022	25.30	< .0001
Erreur	102	27287	267.51752		
Total (sommes corrigées)	111	88192			

Les résultats montrent que le modèle de régression dans son ensemble est significatif. Avec 9 degrés de liberté (DDL) associés au modèle, la somme des carrés du modèle est de 60905. La moyenne quadratique du modèle est de 6767.21022. La

valeur F est de 25.30 avec une $p - value$ (Pr > F) inférieure à 0.0001 plus petit que le seuil $\alpha = 0.05$, ainsi on ne peut pas conservée H_0 ce qui indique que le modèle global est statistiquement significatif.

De plus le R^2 est de 0.6906 ce qui est bien.

Sélection du modèle

Il est important de ne pas trop se fier au modèle complet car plus un modèle a de paramètres, moins les estimations de ces paramètres sont précises et les tests associés efficaces. De plus, il faut un nombre suffisant d'observations par paramètre estimé : 10 observations par paramètre, il y a 112 observations soit 11 paramètres possible. Pour chaque variable quantitative, un paramètre est estimé, en comptant également l'ordonnée à l'origine β_0.

Variance Inflation Factor (VIF)

Le *Variance Inflation Factor* (VIF) est un outil qu'on utilise pour mesurer combien les variables explicatives d'un modèle de régression linéaire multiple sont corrélées entre elles (c'est ce qu'on appelle la multicolinéarité). On calcule un VIF pour chaque variable explicative X_i.

La formule pour le VIF est :

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

Ici, R_i^2 représente la qualité du modèle quand toutes les autres variables sont utilisées pour prédire X_i.

Choisir le bon modèle

Si une variable a un VIF > 10, ça veut dire qu'elle est trop corrélée avec les autres variables. Dans ce cas, il vaut mieux l'enlever du modèle parce qu'elle apporte une information redondante.

Remarque : Le seuil de 10 est un choix arbitraire. Certains préfèrent utiliser une valeur de 3 pour décider si une variable est trop corrélée. Dans notre cas nous allons garder 10.

Calcul des VIF

Pour ce faire nous allons utliser la procédure REG :

Variable	Libellé	DDL	Valeur estimée	Erreur type	Valeur du test t	Pr > |t|	VIF
Intercept	Intercept	1	13.35801	15.34273	0.87	0.3860	0
T9	T9	1	1.90596	1.22020	1.56	0.1214	6.02420
T12	T12	1	1.72075	1.62893	1.06	0.2933	17.99033
T15	T15	1	0.88243	1.30206	0.68	0.4995	14.44099
Ne9	Ne9	1	-1.23230	1.05076	-1.17	0.2436	3.08475
Ne12	Ne12	1	-1.19578	1.54976	-0.77	0.4421	5.18893
Ne15	Ne15	1	-0.05596	1.14113	-0.05	0.9610	2.93896
Vx9	Vx9	1	2.09419	1.01255	2.07	0.0411	2.94865
Vx12	Vx12	1	-0.44119	1.19804	-0.37	0.7134	4.65468
Vx15	Vx15	1	0.52199	1.04274	0.50	0.6177	3.56281

Table 0.14 – Paramètres estimés

On regarde la dernière colonne et on retire la variable avec le plus grand VIF. Ici c'est la variable T12 avec un VIF de 17.99033.

On réitère la procédure en enlevant la variable T12.

Variable	Libellé	DDL	Valeur estimée	Erreur type	Valeur du test t	Pr > |t|	VIF
Intercept	Intercept	1	16.88331	14.98382	1.13	0.2625	0
T9	T9	1	2.53861	1.06371	2.39	0.0188	4.57296
T15	T15	1	1.88751	0.88939	2.12	0.0362	6.73019
Ne9	Ne9	1	-1.08616	1.04219	-1.04	0.2998	3.03129
Ne12	Ne12	1	-1.98021	1.36103	-1.45	0.1487	3.99759
Ne15	Ne15	1	0.44205	1.03978	0.43	0.6716	2.43735
Vx9	Vx9	1	2.28446	0.99697	2.29	0.0240	2.85535
Vx12	Vx12	1	-0.69220	1.17490	-0.59	0.5570	4.47157
Vx15	Vx15	1	0.64291	1.03702	0.62	0.5367	3.51988

Table 0.15 – Tableau des paramètres estimés

Ici on constate que tout les VIF sont supérieur à 10.

Ainsi il faut faire une comparaisons des modèle, soit par l'AIC le BIC ou le test du Rapport de vraissemblance.

Sélection du Modèle avec le Test LRT et la Procédure Backward

La sélection du modèle peut être réalisée grâce au test LRT (Log Likelihood Ratio Test) en utilisant des procédures backward. Le LRT compare les vraisemblances entre deux modèles pour identifier les variables à exclure.

Hypothèse du LRT

— H_0 **(Hypothèse nulle)** : Absence d'effet ou de différence significative entre les modèles avec et sans une variable spécifique.

— Si la p-valeur < 0,05, on rejette H_0, indiquant que la variable a un effet significatif et doit être incluse dans le modèle.

Procédure Backward sur SAS

1. On commence avec le modèle complet incluant toutes les variables.
2. Ensuite SAS génère un tableau avec une ligne par variable, affichant les p-valeurs des tests LRT.
3. — Enfin si toutes les p-valeurs des tests LRT sont < 0,05, on arrête le processus.
   — Sinon, on retire la variable avec la p-valeur la plus grande.
4. : On répète les étapes jusqu'à ce que toutes les p-valeurs restantes soient < 0,05.

Voici le résumé de la procédure

Étape	**Variable supprimée**	**Libellé**	**Nombre de variables dans le modèle**	**R carré partiel**	**R carré du modèle**	**Critère C(p)**	**Valeur F**	**Pr > F**
1	Ne15	Ne15	7	0.0005	0.6867	7.1807	0.18	0.6716
2	Vx15	Vx15	6	0.0011	0.6856	5.5394	0.36	0.5490
3	Vx12	Vx12	5	0.0003	0.6853	3.6449	0.11	0.7443
4	Ne9	Ne9	4	0.0039	0.6814	2.9198	1.30	0.2561

Table 0.16 – Synthèse de l'élimination descendante

Voici donc le modèle optimal.

Analyse de variance

Table 0.17 – Tableau d'analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	4	60092	15023	57.21	< .0001
Erreur	107	28099	262.61187		
Total (sommes corrigées)	111	88192			

Estimation des paramètres

Variable	Valeur estimée des paramètres	Erreur type	SC Type II	Valeur F	Pr > F
Intercept	15.02388	13.80973	310.81911	1.18	0.2791
T9	2.84132	0.97039	2251.45209	8.57	0.0042
T15	1.70757	0.79107	1223.60196	4.66	0.0331
Ne12	-2.51378	0.93344	1904.54727	7.25	0.0082
Vx9	2.39924	0.70578	3034.77818	11.56	0.0009

Table 0.18 – Estimation des paramètres

Interprétation des résultats

Le R^2 est de 0.6814 ce qui est bien.

D'après le tableau de l'analyse de la variance une valeur F de 57.21 ce qui est élevée et une p-value < 0.0001 qui est plus petit que le seuil α indiquent que le modèle est globalement significatif .

Comme les pvalue de chaque variable est plus petit que 0.05, cela signifie que chaque variable est significatives.

Le modèle optimal retenu est donc

$$\text{maxO3} = \beta_0 + \beta_1 \cdot T9 + \beta_3 \cdot T15 + \beta_5 \cdot Ne12 + \beta_7 \cdot Vx9$$

Limite des deux tests utilisés

Limite du test LRT

L'ordre dans lequel les variables sont ajoutées ou supprimées peut affecter les résultats.

Limite du VIF

Les Variance Inflation Factors (VIF) présentent certains inconvénients. Dans les échantillons de petite taille, les estimations des VIF peuvent être moins précises et moins fiables. De plus, le seuil de tolérance pour les VIF est subjectif. Par exemple, certaines personnes utilisent un seuil de 10, tandis que d'autres préfèrent un seuil plus strict de 3. Ce qui peut conduire à des résultats différents.

Conclusion

Nos observations, en particulier l'estimation des paramètres suggèrent que la température, à 9h et 15h mais un peu plus à 9h, influence positivement la concentration d'ozone (car estimation positive), tandis que la nébulosité à 12h est associée à une diminution du taux d'ozone (car estimation négative). De plus, un vent fort d'est en ouest à 9h entraîne une augmentation significative du taux d'ozone (car estimation positive).