République Algérienne Démocratique et Populaire

Ministère de l'Enseignement supérieur et de la Recherche scientifique

Université de Médéa

Faculté des Sciences

Polycopié de cours de Biostatistique

Selon le programme pédagogique officiel de la licence-L3-
(Immunologie, Sciences Biologiques, Science de la nature et de la vie)

Dr. KHAMES Gholame Allah Yacine

Maître de conférences classe B

Département SNV

Validé par le conseil scientifique de la faculté des sciences à .....

Numéro de série: .....

Année universitaire 2021-2022

Préambule

Ce polycopié fait suite au document intitulé (Biostatistique), réalisé en 2022.

Une approche étape par étape a été rédigée comme une aide dans le cours de statistique pour les étudiants dont les connaissances mathématiques se limitent à l'algèbre de base. Il suit une approche non théorique sans preuves formelles, expliquant les concepts de manière intuitive et les étayant d'exemples abondants.

Le cours est constitué d'un support pédagogique très utile, aux étudiants de Licence etde Master des sciences de la nature et de la vie et des sciences environnementales. C'est est unoutil nécessaire pour approfondir les raisonnements et les méthodes de base de l'analysestatistique, ainsi que ses applications à la résolution de problèmes statistiques rencontrés dansle domaine.

Le polycopié couvre une partie importante des éléments de base de l'analyse statistique pour les classesde Licence pour les étudiants des sciences de la nature et de la vie. Il repose sur des exercicespratiques avec une augmentation progressive des degrés de difficulté ;

Les applications couvrent un large éventail de sujets susceptibles de plaire aux étudiants d'horizons divers, et elles incluent des problèmes dans les domaines des sports, de la santé, de l'architecture, de l'éducation, du divertissement, de l'environnement, transports, habitudes alimentaires.

Il constitue un support pédagogique, aux étudiants de Licence et de Master intéressésdans le domaine de l'immunologie, de microbiologie et de l'écologie. Ce document traite les statistiques paramétriques et statistiques non paramétriques. Les tests paramétriquesse basent sur des distributions statistiques supposées dans les données. Par conséquent, certaines conditions de validité doivent être vérifiées pour que le résultat d'un test paramétrique soit fiable.

Les tests non-paramétriquesne se basent pas sur des distributions statistiques. Ils peuvent donc être utilisés même si les conditions de validité des tests paramétriques ne sont pas vérifiées.

Sommaire

Chapitre 1: Corrélation **Page1**

Chapitre 2: Régression linéaire simple **Page 8**

Chapitre 3: Régression multiple **Page 17**

Chapitre 4: Analyse de variance à un seul facteur et test de Scheffé **Page 26**

Chapitre 5: Test de Khi 2 (Chi carré) **Page 46**

Chapitre 6: Test non paramétriques **Page 55**

Bibliographie: **Page 65**

Chapitre 1:Corrélation

Objectifs :

- **Dessiner un diagramme de dispersion pour un ensemble de paires ordonnées**
- **Calculer le coefficient de corrélation de Pearson (r)**
- **Tester l'hypothèse H0**

Corrélation

Il y a une corrélation entre deux variables quand l'une est liée ou associée à l'autre. Le coefficient de corrélation linéaire (*r*) est calculé à partir des données de l'échantillon. Il mesure la force (ou l'intensité) et la direction d'une relation linéaire entre deux variables quantitatives. Cette valeur est calculée en utilisant la formule ci-dessous ou d'autres formules équivalentes.

$$\boldsymbol{r = \frac{\frac{\sum[(x_i-\bar{x})\times(y_i-\bar{y})]}{(n-1)}}{\sqrt{\frac{(x_i-\bar{x})^2}{(n-1)}}\times\sqrt{\frac{(y_i-\bar{y})^2}{(n-1)}}}} \qquad \textbf{(1)}$$

n : nombre de paires

$\sum(x_i - \bar{x})$: somme des écarts par rapport à la moyenne $\bar{x}$

$\sum(y_i - \bar{y})$: somme des écarts par rapport à la moyenne $\bar{y}$

$\frac{\sum[(x_i-\bar{x})\times(y_i-\bar{y})]}{(n-1)}$: la covariance

$\sum(x_i - \bar{x})^2$: carrés des écarts à la moyenne $\bar{x}$

$\sum(y_i - \bar{y})^2$: carrés des écarts à la moyenne $\bar{y}$

$\sqrt{\frac{(x_i-\bar{x})^2}{(n-1)}}$: écart-type (*x*)

$\sqrt{\frac{(y_i-\bar{y})^2}{(n-1)}}$: écart-type (*y*)

Le coefficient de corrélation linéaire est compris entre -1 et +1 (Figure 1, A et B) :

1. S'il existe une forte relation linéaire positive entre les variables, la valeur de *r* sera proche de +1.
2. S'il existe une forte relation linéaire négative entre les variables, la valeur de *r* sera proche de –1.
3. S'il n'y a pas de relation linéaire entre les variables ou faible, la valeur de *r* sera proche de 0.

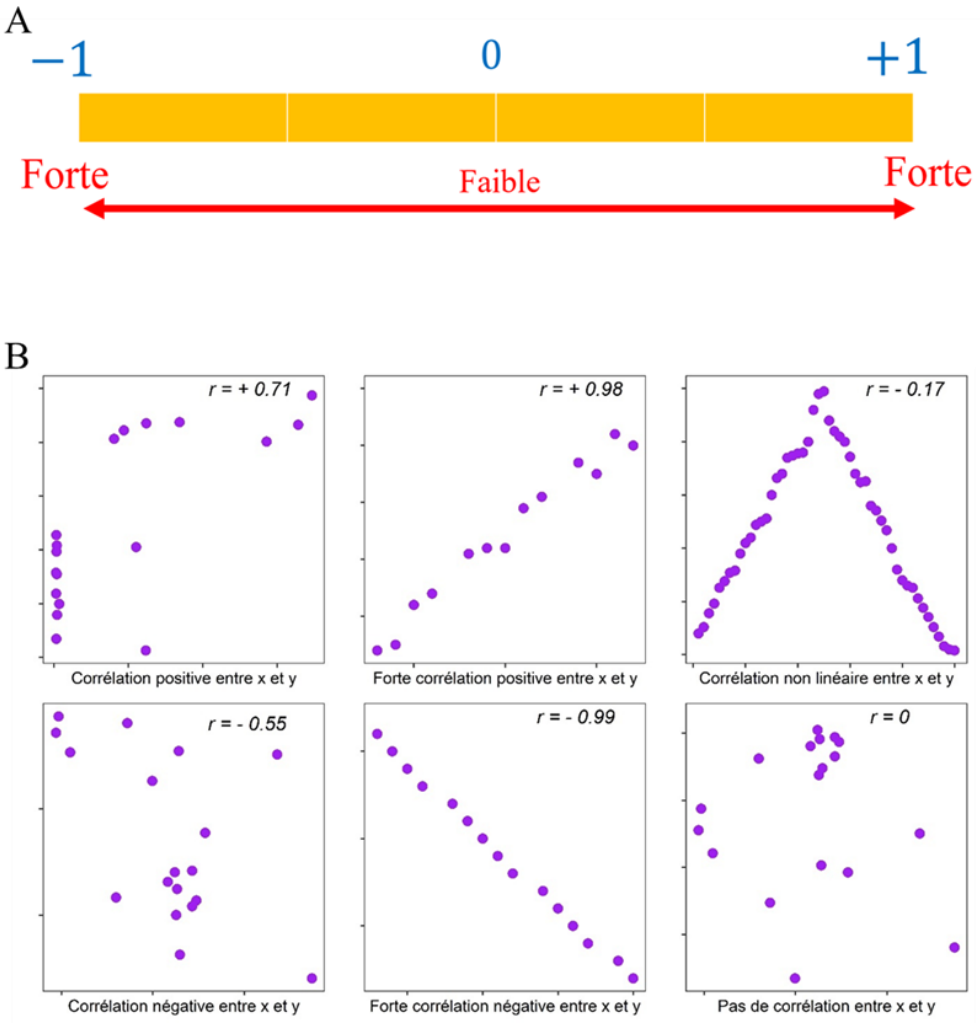


Figure 1. La gamme de valeurs du coefficient de corrélation de Pearson (r)

Formule équivalente :

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_1^2) - (\sum x_i)^2} \times \sqrt{n(\sum y_i{}^2) - (\sum y_i)^2}} \qquad (2)$$

Cette formule simplifie le calcule et aisée à utiliser avec un tableau.

Le Tableau 1 montre la répartition de 13 paires de 13 moineaux ayant des longueurs et d'âges (croissance) différents. Le but est de déterminer le coefficient de corrélation linéaire de Pearson (*r*).

Détermination de la valeur du coefficient de corrélation linéaire

Faites un tableau comme indiqué :

Etape 1 :

Tableau 1. La répartition de 13 paires de 13 moineaux ayant des longueurs et d'âges (croissance) différents

Age (jours) (x_i)	**Longueur des ailes (cm)** (y_i)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
3	1.4	-7	-2.02	14.11	49	4.0804
4	1.5	-6	-1.92	11.49	36	3.6864
5	2.2	-5	-1.22	6.08	25	1.4884
6	2.4	-4	-1.02	4.06	16	1.0404
8	3.1	-2	-0.32	0.63	4	0.1024
9	3.2	-1	-0.22	0.22	1	0.0484
10	3.2	0	-0.22	0	0	0.0484
11	3.9	1	0.48	0.48	1	0.2304
12	4.1	2	0.68	1.37	4	0.4624
14	4.7	4	1.28	5.14	16	1.6384
15	4.5	5	1.08	5.42	25	1.1664
16	5.2	6	1.78	10.71	36	3.1684
17	5	7	1.58	11.09	49	2.4964
$\bar{x} = 10$	$\bar{y} = 3.41$			$\sum = 70.8$	$\sum = 262$	$\sum = 19.65$

Etape 2 :

Remplacer dans la formule ci-après et résoudre (*r*).

$$r = \frac{\frac{\sum[(x_i-\bar{x})\times(y_i-\bar{y})]}{(n-1)}}{\sqrt{\frac{(x_i-\bar{x})^2}{(n-1)}} \times \sqrt{\frac{(y_i-\bar{y})^2}{(n-1)}}}$$

$$= \frac{\frac{70.8}{(13-1)}}{\sqrt{\frac{262}{(13-1)}} \times \sqrt{\frac{19.65}{(13-1)}}} = 0.98$$

Le coefficient de corrélation linéaire de Pearson (*r*) égale à 0.98, suggère une forte liaison entre l'âge et la longueur des ailes chez les moineaux.

Le coefficient de corrélation comme indiqué précédemment, la zone du coefficient de corrélation (*r*) se situe entre –1 et +1, montre une relation forte relation linéaire.

Si la valeur de r est proche de 0, la relation linéaire est faible ou inexistante. Ainsi, il y a deux possibilités lorsque r n'est pas égal à zéro : soit la valeur de r est suffisamment élevée pour conclure qu'il y a une relation linéaire significative entre les variables, ou la valeur de r est due au hasard.

P-value

Corrélation n'implique pas une causalité. A partir des données du tableau 1 et l'équation de corrélation de Pearson (*r*), nous pouvons conclure qu'il y a une forte liaison entre la longueur des ailes et l'âge (croissance), mais nous ne pouvons pas affirmer que la croissance (âge) cause des longueurs plus importantes. La croissance peut être affectées par d'autres variables autres que la longueur des ailes. Ainsi, pour déterminer l'existence d'une corrélation linéaire **significative** entre deux variables, la p-value est la meilleure solution.

Trouver la p-value :

Etape 1 :

Posons

H_0: il n'y a pas de corrélation significative

H_1: il a une corrélation significative

Etape 2 :

Utiliser la formule ci-dessous :

$$\boldsymbol{t = \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}}} \qquad \textbf{(3)}$$

r : le coefficient de corrélation de Pearson

n : nombre des paires

Pour une valeur de *t* calculer, la P-value est lue sur la table de la loi de Student avec un $\boldsymbol{ddl = n - 2}$. Si H_0est rejetée, on conclut à l'existence d'une corrélation linéaire significative. Et si H_0 n'est pas rejetée, il n'y a pas suffisamment de preuves pour affirmer une corrélation linéaire significative.

L'exemple précèdent montre une forte corrélation linéaire. Affirmer qu'il y'a une corrélation significative est équivalent à dire que le coefficient de corrélation *r* de toute la population est différent de 0.

Solution

Etape 1 :

H_0 : il n'y a pas de corrélation significative entre l'âge (croissance) et la longueur des ailes

H_1 : il a une corrélation significative entre l'âge (croissance) et la longueur des ailes

Etape 2 :

Utiliser la formule ci-dessous :

$$t = \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}}$$

$$= \frac{0.98}{\sqrt{\frac{(1-0.98^2)}{(13-2)}}}$$

$$= \frac{0.98}{\sqrt{\frac{(1-0.98^2)}{(13-2)}}}$$

=16.33

d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437

16.33

Figure 2. Estimation de la valeur de p (p-value)

Etape 3 :

ddl = 13 – 2 = 11

Pour t =16.33, la valeur de la p-value est largement inférieur à 0,001 ; H_0 rejetée. Il y a suffisamment d'éléments pour confirmer une corrélation entre l'âge ou la croissance et la longueur des ailes.

Corrélation – Régression

La relation entre deux variables peut être une dépendance fonctionnelle de l'une par rapport à l'autre, C'est-à-dire que l'on suppose que l'ampleur de l'une des variables (**la variable dépendante** ou **variable réponse**) est déterminée par l'ampleur de la deuxième variable (**la variable indépendante**, **explicative** ou **variable prédictive**), Alors que l'inverse n'est pas vrai.

Exemple :

Dans la relation entre la pression artérielle et l'âge chez l'homme, La pression artérielle peut être considérée comme la variable dépendante et l'âge comme la variable indépendante. La variable indépendante est la variable en régression qui peut être contrôlée ou manipulée. Dans ce cas, la pression artérielle est la variable indépendante désignée comme la variable *x*. La variable dépendante est la variable en régression qui ne peut être contrôlée ou manipulée. L'âge est la variable dépendante, désignée comme la variable *y*. La raison de cette distinction entre les variables est que vous supposez que l'âge obtenue par les personnes dépend de la pression artérielle enregistrée. De plus, vous supposez que, dans une certaine mesure, la personne peut réguler ou contrôler la pression artérielle (Exemple : moins de stress, sport, médicament...). La variable indépendante est également connue sous le nom de variable explicative, et la variable dépendante est également appelée variable de réponse.

Le terme "**dépendant**" n'implique pas nécessairement une relation de cause à effet entre les deux variables. Une telle relation de dépendance est appelée une **régression**. Le terme de régression fait référence au type de régression le plus simple dans laquelle seules deux variables sont prises en compte.

Les variables indépendantes et dépendantes peuvent être tracées sur un graphique appelé "**Diagramme de dispersion**". La variable indépendante *x* est tracée sur l'axe horizontal, et la variable dépendante *y* est tracée sur l'axe vertical.

Chapitre 2: Régression linéaire simple

Objectifs :

- **Déterminer l'équation de la droite de régression**
- **Validation par visualisation (diagramme des résidus)**
- **Validation par le calculer du coefficient de détermination (R^2)**
- **Familiariser avec le concept de régression multiple**

Régression

Le but de la régression est de trouvé une relation par rapport à une ou plusieurs variables indépendantes (explicatives). Pour étudier les relations entre deux variables, il faut collecter les données puis construire un nuage de point. Le but du diagramme de dispersion (nuage de point) est de déterminer la nature de la relation entre les variables. Les possibilités comprennent une relation linéaire, une relation linéaire négative, une relation curviligne (non linéaire) ou aucune relation (Figure 1, B). Après avoir tracé le nuage de points et déterminé une relation linéaire, les prochaines étapes consistent à calculer la valeur du coefficient de corrélation de Pearson (*r*) et à tester la significativité de cette relation. Si la valeur du coefficient de corrélation est significative, l'étape suivante consiste à déterminer **l'équation de la droite de régression**.

Il est pratique et instructif de représenter graphiquement des données de régression simple en utilisant l'ordonnée (axe des y) pour la variable dépendante et l'abscisse (axe des *x*) pour la variable indépendante.

Tableau 2. La répartition de 13 paires de 13 moineaux ayant des longueurs et d'âges (croissance) différents

Age (jours) (X)	Longueur des ailes (cm) (Y)
3	**1.4**
4	**1.5**
5	**2.2**
6	**2.4**
8	**3.1**
9	**3.2**
10	**3.2**
11	**3.9**
12	**4.1**
14	**4.7**
15	**4.5**
16	**5.2**
17	**2**

Un tel graphique est présenté dans la Figure 2 pour Tableau 2 où les données apparaissent sous la forme d'une dispersion de 13 points (en rouge), chaque point représente une paire de valeurs de *x* et de *y*. Une paire de données *x* et *y* peut être désignée comme (x_1, y_1), une autre comme (x_2, y_2), une autre comme (x_3, y_3), et ainsi de suite, ce qui donne ce que l'on appelle un diagramme de dispersion. De l'ensemble des n données (x_i, y_i), les lignes bleus passant par les données de cette figure est représentée par la Figure 3.

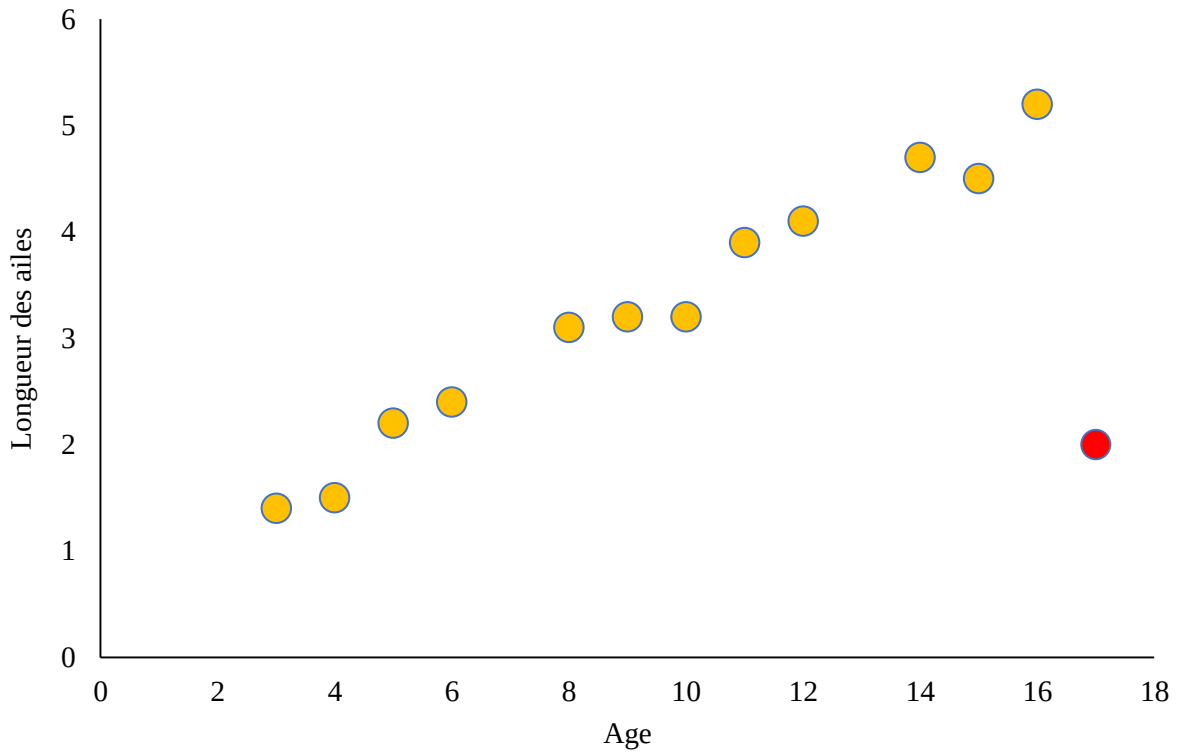


Figure 3. La répatition de la longueur des ailes en fonction de l'age ches les moineaux

Dans de nombreux types de données biologiques, la relation entre deux variables n'est pas unequestion de dépendance. Dans ce cas, l'ampleur de l'une des variables change comme l'ampleur de la deuxième variable. Dans de telles situations, il y a corrélation, plutôt que de régression.

Equation de régression

Plusieurs lignes peuvent être dessinées sur le graphique à proximité des points. Avec un nuage de points, vous devez être capable de dessiner la ligne du meilleur ajustement. Autrement dit, **la meilleure équation de régression (**Figure 4A).

Le meilleur ajustement signifie que la somme des carrés des distances verticales de chaque point de la ligne est au minimum**(**Figure 4B). La différence entre la valeur réelle *y* et la valeur prédite $\hat{y}$(la distance) est appelée une **erreur résiduelle** ou une **erreur prédite**. La méthode utilisée pour rendre les résidus aussi petits que possible est appelée **la méthode des moindres carrés**. Les résidus sont utilisés pour déterminer au mieux la relation entre les deux variables (*x* et *y*).

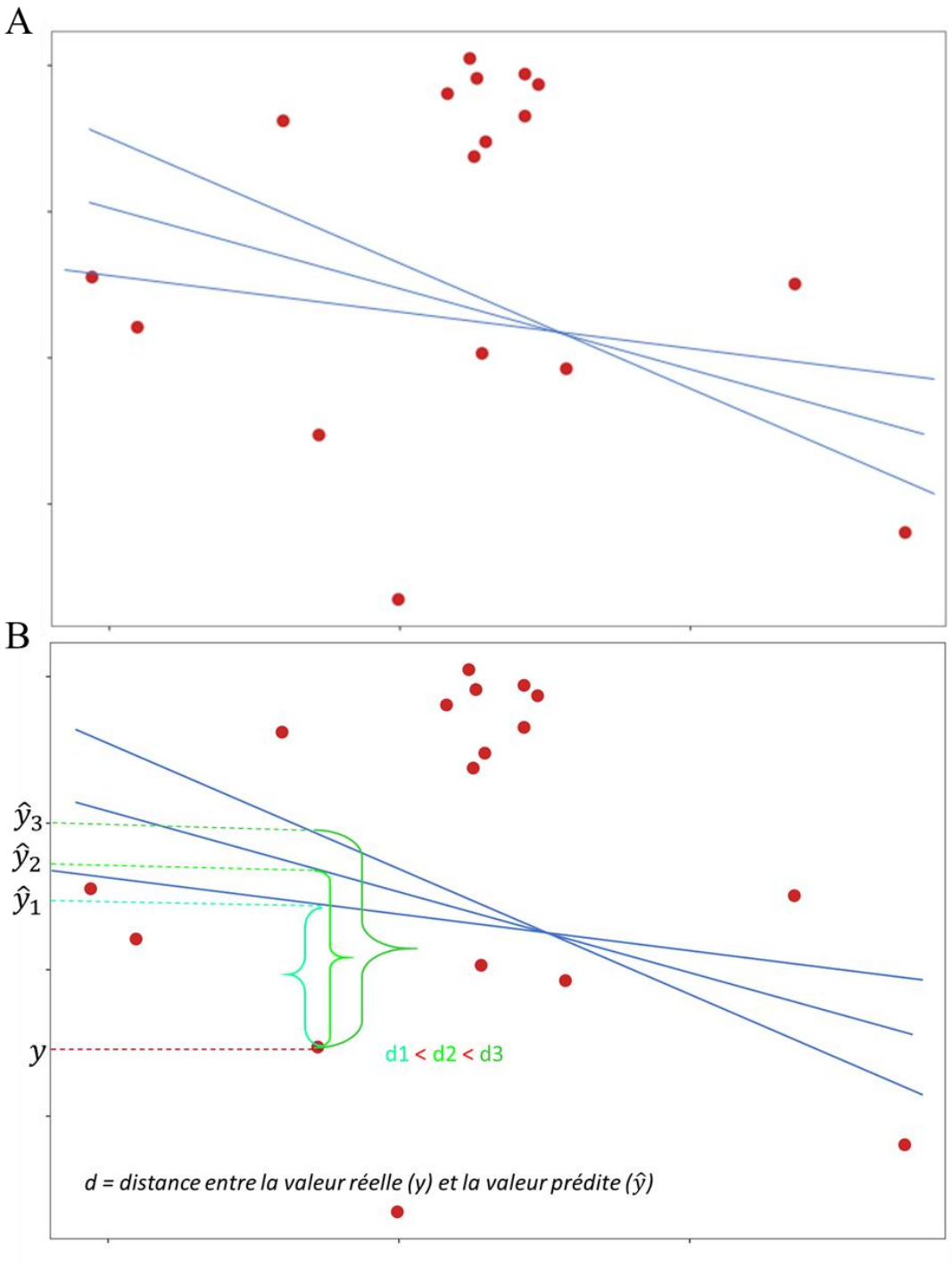


Figure 4. La meilleure droite de régression par la méthode des moindres carrés

Il y a plusieurs méthodes pour trouver l'équation de la droite de régression. Une seule formule pour trouver la meilleure équation de régression est donnée ici. Cette équation minimise la somme des carrés des distances des résidus.

L'équation de régression qui représente au mieux l'ensemble du nuage de points

$$\boldsymbol{y = a\,x + b} \qquad (4)$$

La pente*a*

$$\boldsymbol{a = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}} \qquad (5)$$

L'ordonnée à l'origine (intercept *b*)

$$\boldsymbol{b = \bar{y} - a\,\bar{x}} \qquad \textbf{(6)}$$

Détermination de l'équation de régression

Etape 1 :

Faite un tableau comme indiqué ci-dessous

Age (jours) (x_i)	**Longueur des ailes (cm) (y_i)**	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
3	1.4	-7	-2.02	14.11	49	4.0804
4	1.5	-6	-1.92	11.49	36	3.6864
5	2.2	-5	-1.22	6.08	25	1.4884
6	2.4	-4	-1.02	4.06	16	1.0404
8	3.1	-2	-0.32	0.63	4	0.1024
9	3.2	-1	-0.22	0.22	1	0.0484
10	3.2	0	-0.22	0	0	0.0484
11	3.9	1	0.48	0.48	1	0.2304
12	4.1	2	0.68	1.37	4	0.4624
14	4.7	4	1.28	5.14	16	1.6384

15	4.5	5	1.08	5.42	25	1.1664
16	5.2	6	1.78	10.71	36	3.1684
17	5	7	1.58	11.09	49	2.4964
$\bar{x} = 10$	$\bar{y} = 3.41$			$\sum = 70.8$	$\sum = 262$	$\sum = 19.65$

Etape 2 :

Remplacer dans la formule de la pente (a) et de l'intercepte (b) :

$$a = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{70.8}{262} = 0.27$$

$$b = \bar{y} - a\,\bar{x} = 3.41 - 0.27(10) = 0.71$$

Alors, **l'équation de régression** est :

$$\hat{y} = 0.27\,x + 0.71$$

L'utilisation de l'équation de régression ($\hat{y}$)

L'équation de régression sert à prédire les valeurs d'une variable à partir des données d'une autre variable. Il est possible d'utiliser son équation pour faire des prédictions à condition que l'équation de régression reflet plus au moins la réalité. Cela peut être vérifié en testant la significativité du coefficient de corrélation de Pearson (r), le calcul du coefficient de détermination (R^2) ou bien par visualisation en traçant un diagramme des résidus.

Validation par visualisation : diagramme des résidus

Le diagramme des résidus est un nuage de points des valeurs dans lequel chaque point de y est représenté par le résidu $d = y - \hat{y}$. Ainsi, les valeurs de d sont en fonction des valeurs de x.

1. Si le diagramme des résidus ne montre aucune structure particulaire, l'équation de régression est une bonne représentation de l'association entre les deux variables.
2. Si le diagramme des résidus possède une forme donnée, l'équation de régression n'est pas une bonne représentation de l'association entre les deux variables.

Il est recommandé d'utiliser un logiciel pour construire un diagramme des résidus pour valider le model de régression entre l'âge et la longueur des ailes. Utiliser le logiciel Excel ou le logiciel R pour analyser la forme du nuage de points.

Le diagramme des résidus (Figure 5A) montre une répartition non homogène des résidus (*Residuals*) en fonction des valeurs de la variable indépendante (*Fitted values*). Pour corriger ce problème, une analyse de régression devrait considérer l'effet des valeurs extrêmes. *Une valeur extrême correspond à un point qui se situe très à l'écart des autres*, définies dans cet exemple par le point rouge :x = 17, y = 2 (Tableau 2 & Figure 3). La valeur extrême est facile à identifier dans le diagramme de dispersion : établissez la droite de régression avec et sans ce point. Si le diagramme des résidus tend vers une répartition homogène, le point est influent.

Ce point est souvent loin de la droite de régression (Figure 3). Il est préférable d'éliminer ce point avec une nouvelle droite de régression et un nouveau graphe des résidus. Cette fois ci le diagramme des résidus est plus ou moins homogène (Figure 5B).

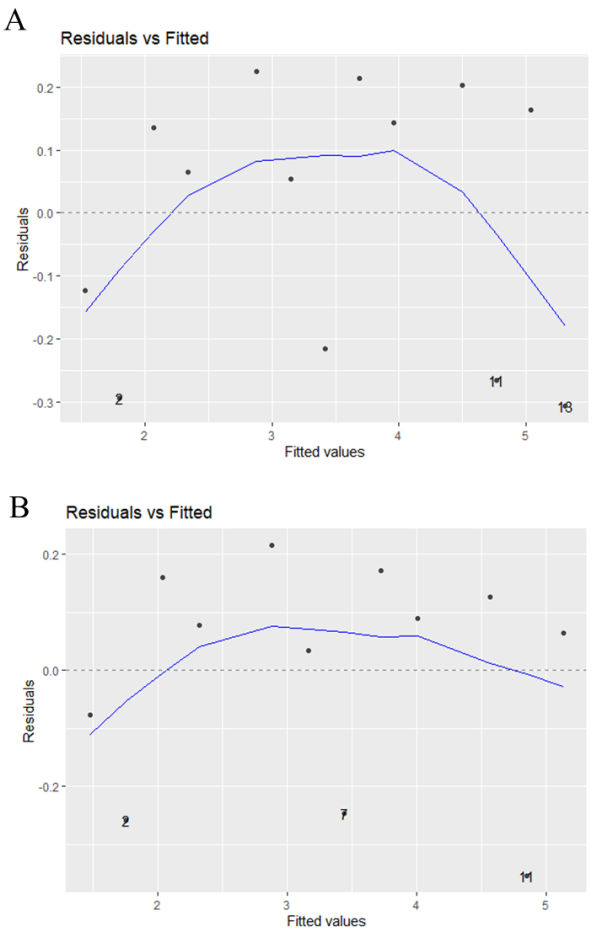


Figure 5. Diagramme des résidus avec les valeurs extrêmes (A) et sans les valeurs extrêmes (B) (tracé à l'aide du logiciel R)

En l'absence de corrélation linéaire significative, nous ne pouvons pas utiliser **l'équation de la droite de régression** pour réaliser des prédictions. Il n'est donc pas nécessaire de collecter des paires de données pour un échantillon d'adultes aléatoirement sélectionnés. La meilleure prédiction pour la valeur de $y_i$est tout simplement le y moyen ($\bar{y}$) (Figure 6).

En l'absence de corrélation linéaire significative, n'utilisez pas l'équation de la droite de régression pour réaliser des prédictions.

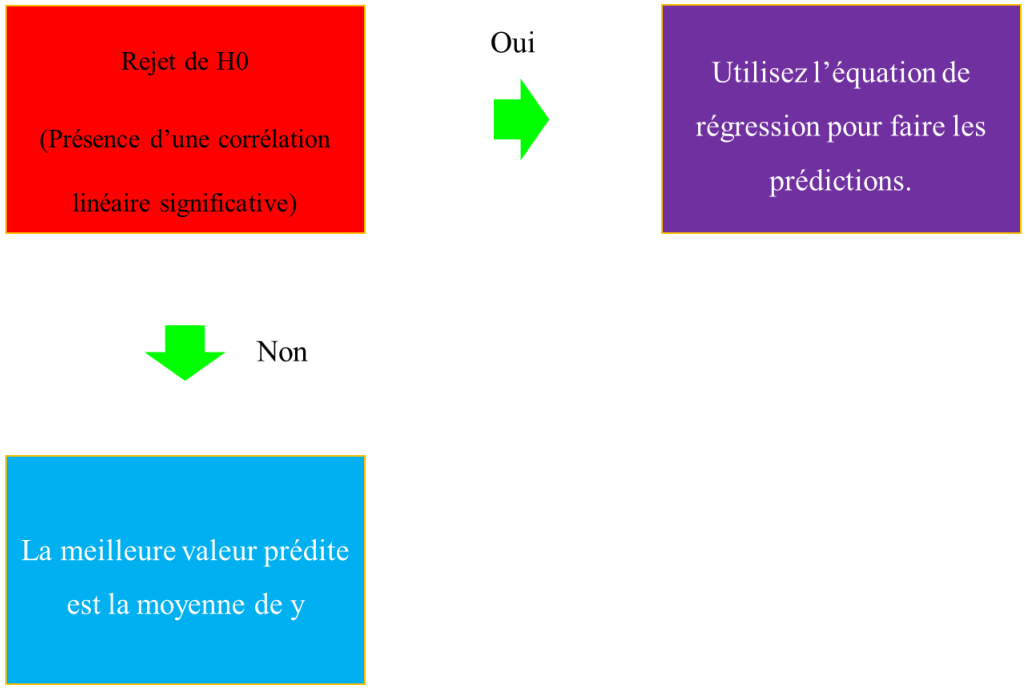


Figure 6. La meilleure prédiction lors d'une corrélation significative et une corrélation non significative

Validation par le coefficient de détermination (R^2)

Comme pour le coefficient de corrélation (r), le coefficient de détermination quantifie la force de liaison entre deux variables quantitatives. Le coefficient de détermination, varie entre 0 et +1, mesure la variation de la variable dépendante (y) qui s'explique par la droite de régression et la variable indépendante (x).

Il est généralement plus facile de trouver le coefficient de détermination en mettant au carré le coefficient de corrélation de Pearson (r) (méthode 1). Par conséquent, si r = 0,98, **alors R^2 = 0,96, ce qui équivaut en termes de pourcentage à 96%**.

Le coefficient de détermination désigné par R^2 est le rapport entre la variation expliquée et la variation totale.

$$R^2 = \frac{Variance\ \ expliquée\ (SCE)}{Variance\ totale (SCT)} \quad = \quad \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y_i-\bar{y})^2} \qquad (7)$$

Trouver le coefficient de détermination (R^2) (méthode 2)

Construire un tableau comme indiqué ci-dessous :

Age (jours) (x_i)	**Longueur des ailes (cm) (y_i)**	$\hat{y}$**= 0.27 (x)+0.71**	$(\hat{y}-\bar{y})^2$	$(y_i-\bar{y})^2$
3	**1.4**	1.52	3.5721	4.0401
4	**1.5**	1.79	2.6244	3.6481
5	**2.2**	2.06	1.8225	1.4641
6	**2.4**	2.33	1.1664	1.0201
8	**3.1**	2.87	0.2916	0.0961
9	**3.2**	3.14	0.0729	0.0441
10	**3.2**	3.41	0	0.0441
11	**3.9**	3.68	0.0729	0.2401
12	**4.1**	3.95	0.2916	0.4761
14	**4.7**	4.49	1.1664	1.6641
15	**4.5**	4.76	1.8225	1.1881
16	**5.2**	5.03	2.6244	3.2041
17	**5**	5.3	3.5721	2.5281
$\bar{x} = 10$	$\bar{y} = 3.41$		$\sum = 19.09$	$\sum = 19.65$

En remplaçant dans la formule $R^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y_i-\bar{y})^2}$,vous trouverez un coefficient approximativement égal à 0,96.

R^2 = 0,96, ce qui signifie que 96% de la variation totale de *y* est expliquée par l'association linéaire entre *x* et *y*. Cela implique que 4% de la variation totale de *y* demeure non expliquée.

Chapitre 3: Régression multiple

Objectifs :

- **Familiariser avec le concept de régression multiple**

Nous avons abordé les analyses des relations de régression et de corrélation entre deux variables (régression linéaire simple). La présente section déploie ce type d'analyse aux régressions et aux corrélations en examinant les interrelations entre trois variables ou plus (régression multiples). Dans **la régression multiple**, une des variables est considérée comme dépendante d'au moins une des autres.

Dans la régression multiple, il y a plusieurs variables indépendantes et une variable dépendante.

L'équation de régression multiple est :

$$\hat{y} = b + a_1 x_1 + a_2 x_2 + \cdots + a_k x_k$$

k : nombre de variables indépendantes

a_1, a_2, a_n: Les pentes

b : l'intercepte ou l'ordonnée à l'origine

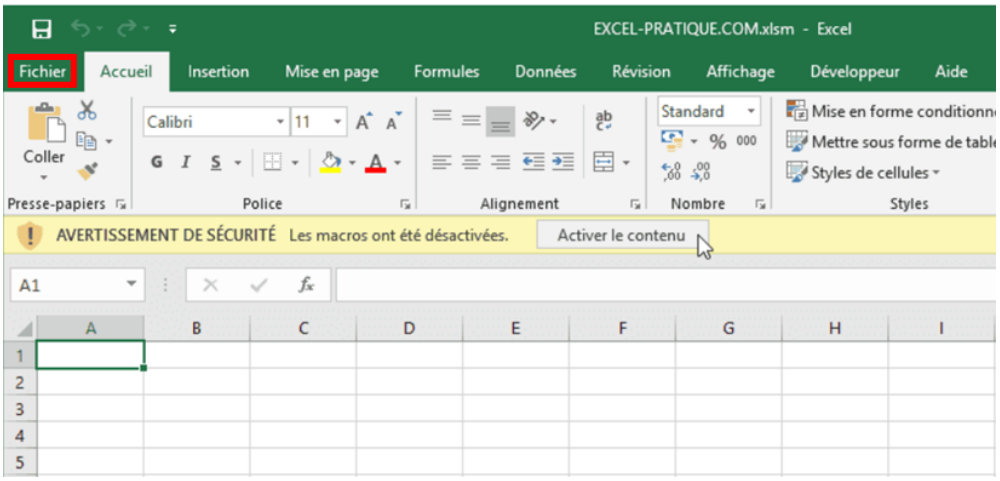
$\hat{y}$: valeurs prédites

Pour établir une équation de régression multiple, il est nécessaire d'utiliser des logiciels tels que Excel ou bien Logiciel R. L'accent sera mis sur l'interprétation de sorties des logiciels.

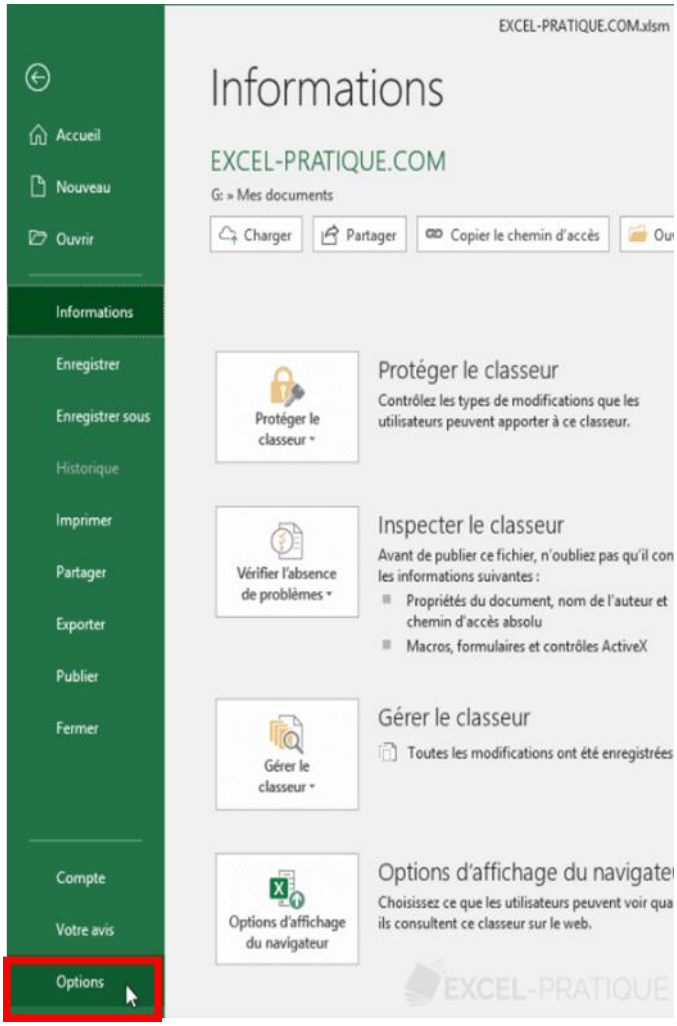
Trouver les résultats de la régression multiple à l'aide du logiciel Excel

1. Allez dans le menu d'Excel.
2. Cliquez sur Options.
3. Cliquez sur Compléments.
4. Dans la partie gérer, sélectionnez compléments d'Excel et cliquez sur Ok.
5. Dans la liste, cochez Utilitaire d'analyse.
6. Cliquez sur Ok.
7. L'utilitaire d'analyse est maintenant actif.

Etape 1 :

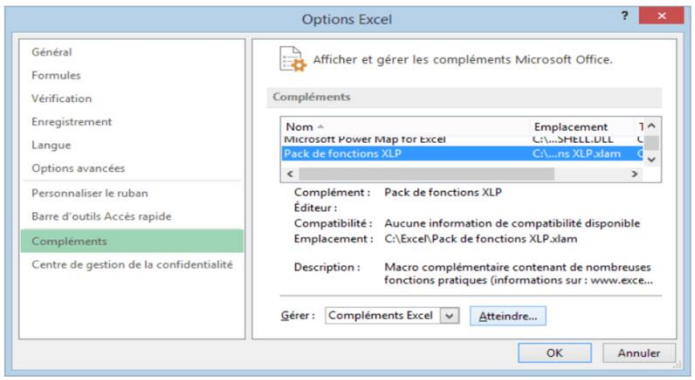


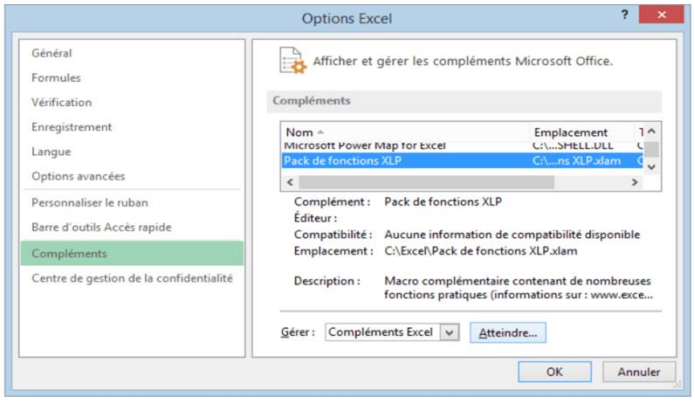
Etape 2 :

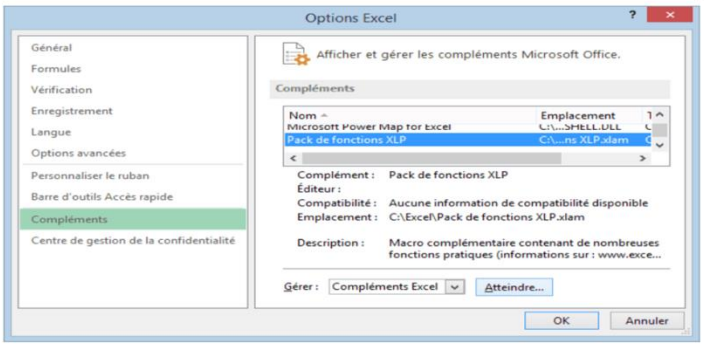


Etape 3 :

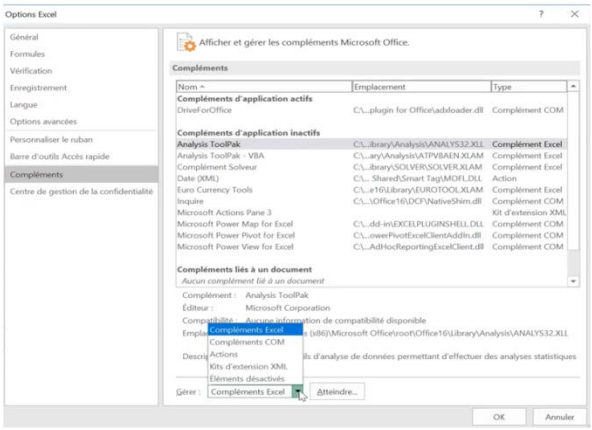
Version 2010

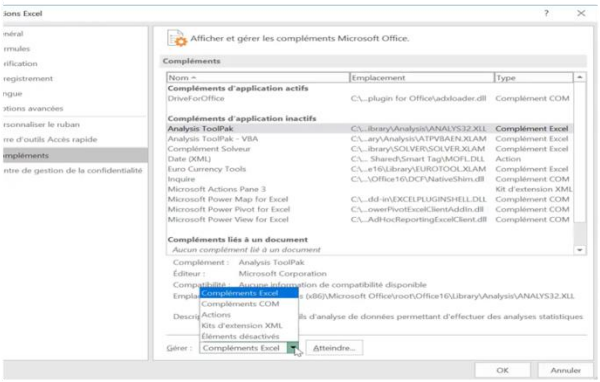


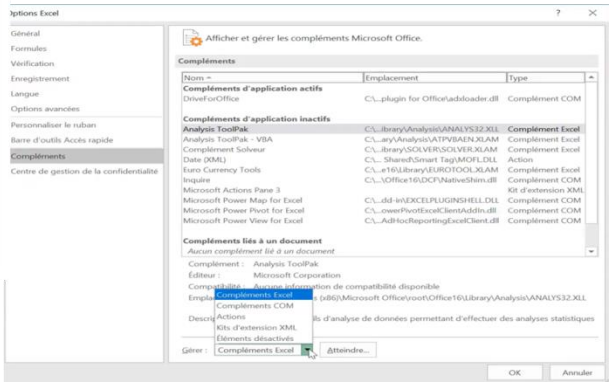




Version 2013

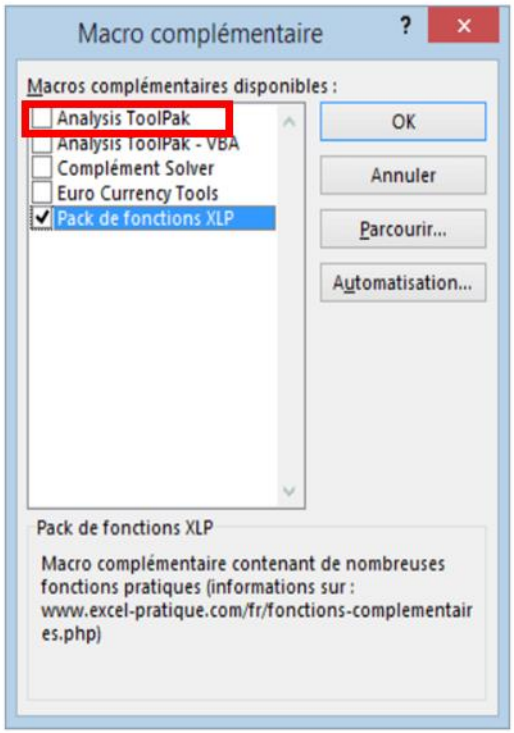


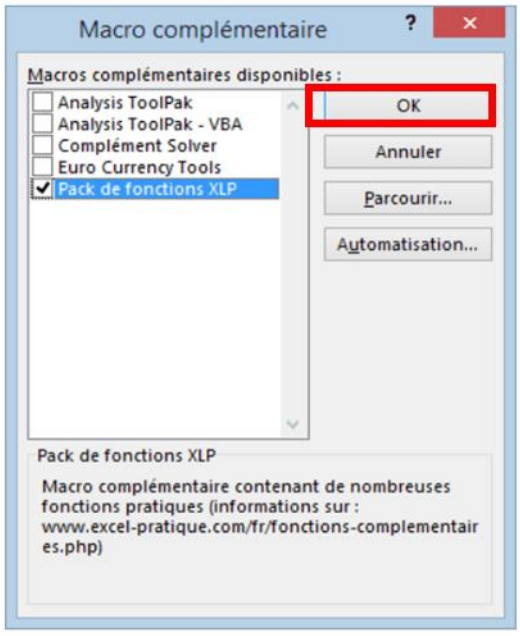




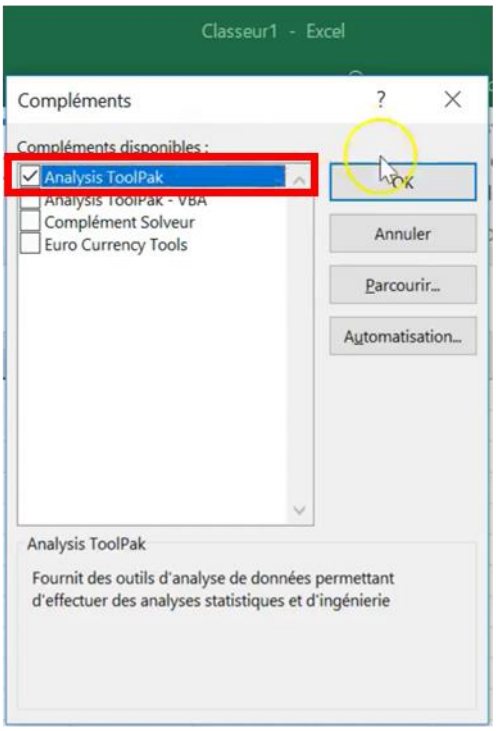
Etape 5 et 6:

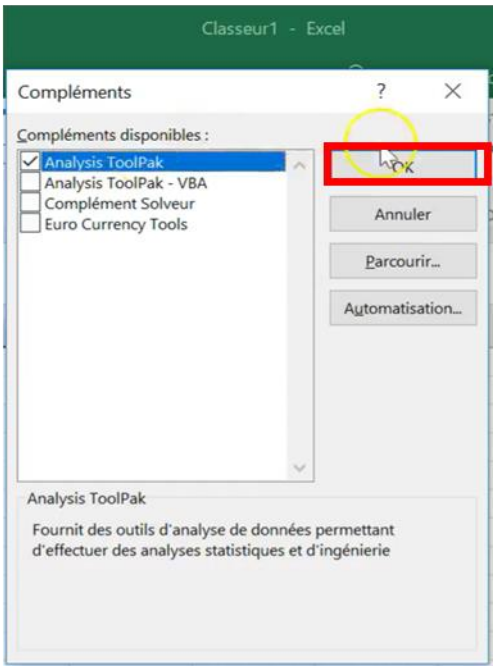
Version 2010



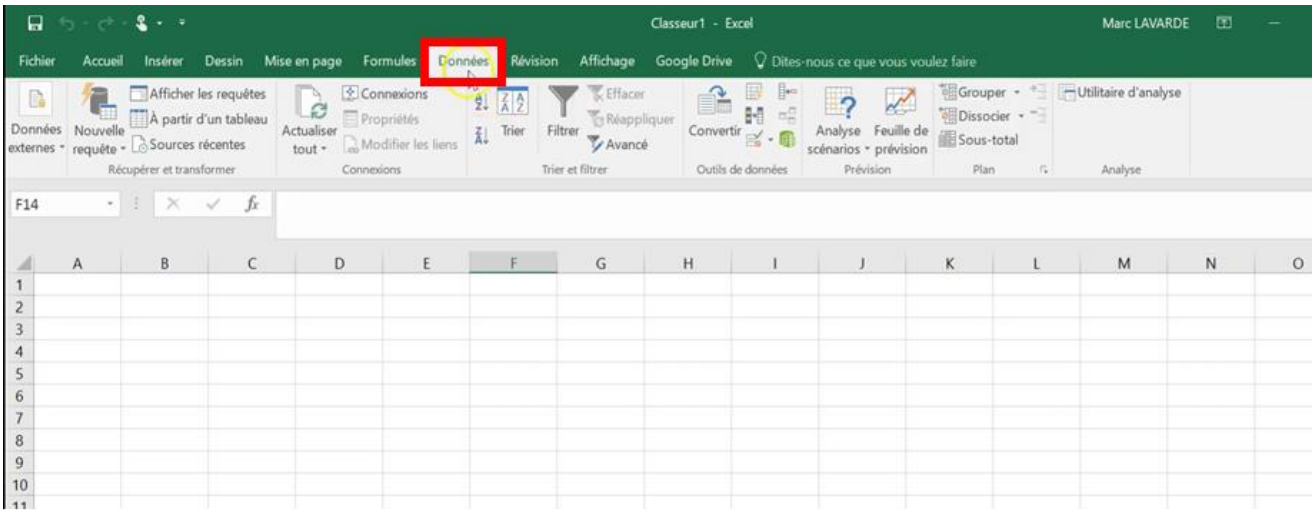


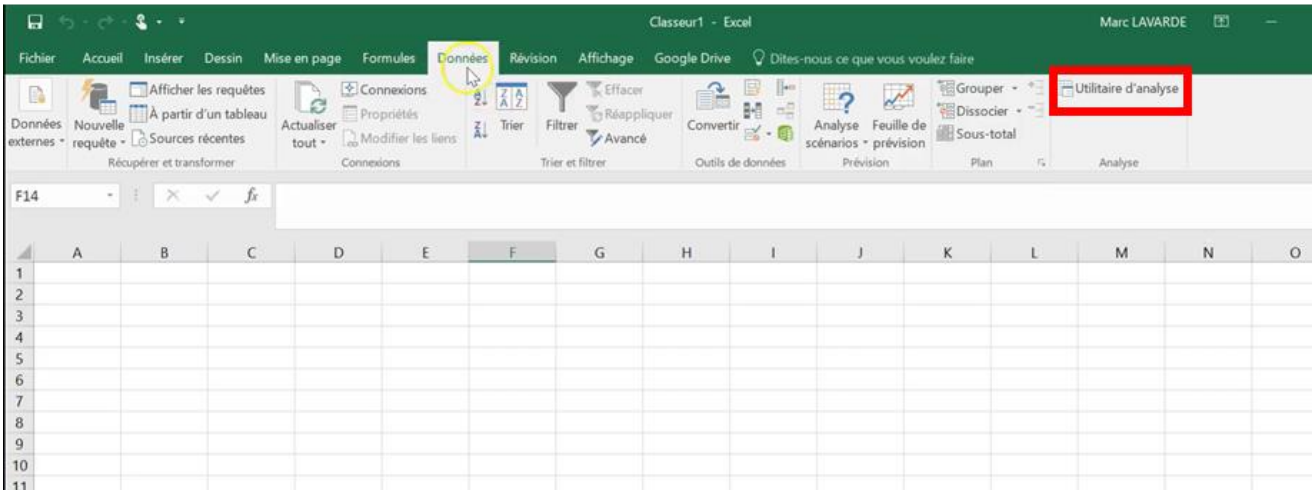
Version 2013





Etape 7 :





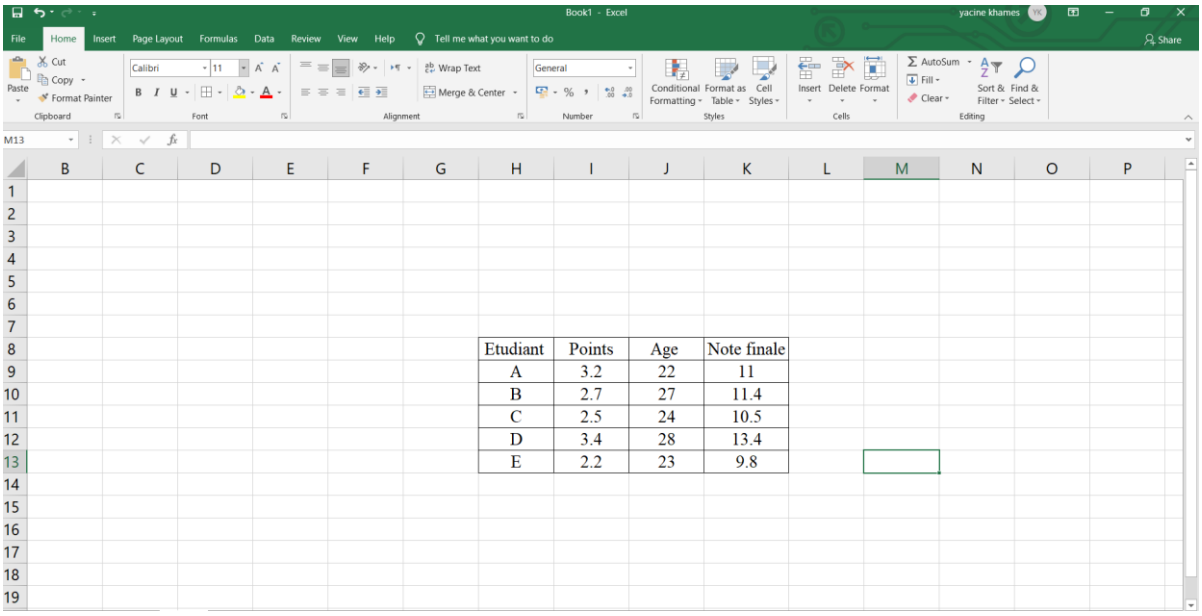
Exemple

Un exemple d'analyse de régression multiple est le cas d'un instructeur en soins souhaite savoir si les points gagnés et l'âge d'un étudiant sont liés à la note obtenue à l'examen final d'infirmier. Pour ce faire, il sélectionne cinq étudiants et obtient les données suivantes.

Etudiant	Points	Age	Note finale
A	3.2	22	11
B	2.7	27	11.4
C	2.5	24	10.5
D	3.4	28	13.4
E	2.2	23	9.8

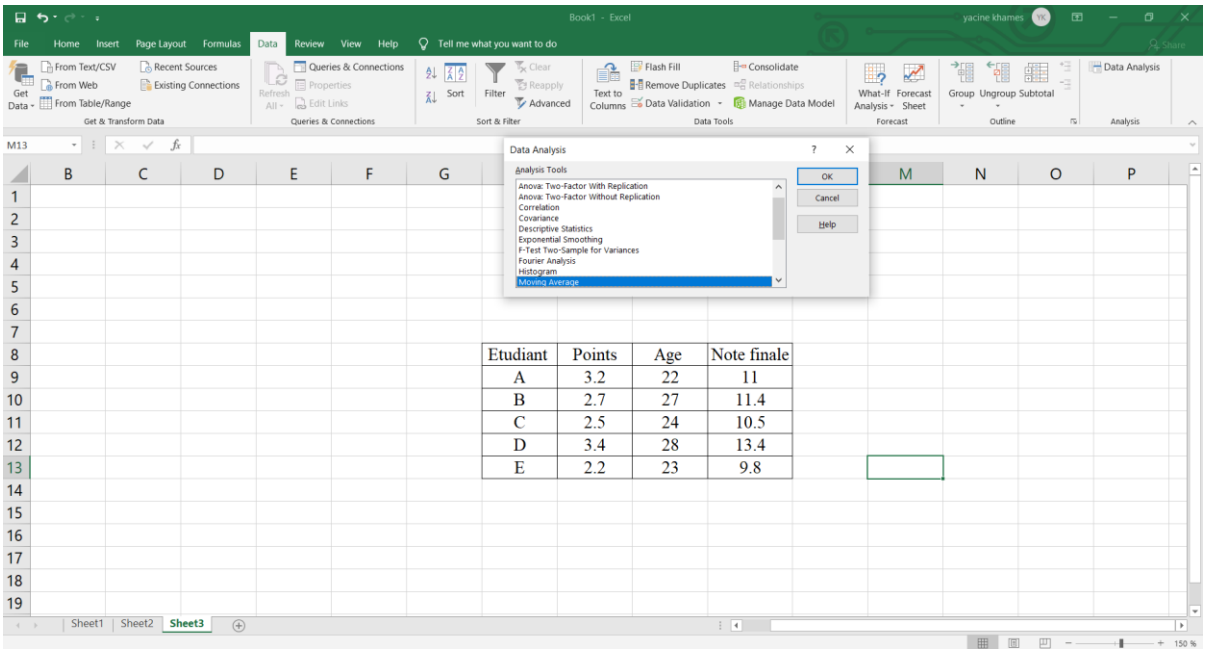
Les calculs nécessaires à la régression multiple sont si compliqués qu'un logiciel doit être utilisé, ici c'est le logiciel Excel :

Etape 1 : Copiez les données sur la feuille Excel



Etudiant	Points	Age	Note finale
A	3.2	22	11
B	2.7	27	11.4
C	2.5	24	10.5
D	3.4	28	13.4
E	2.2	23	9.8

Etape 2 : Ouvrez l'utilitaire d'analyse et choisissez la *Régression Linéaire*



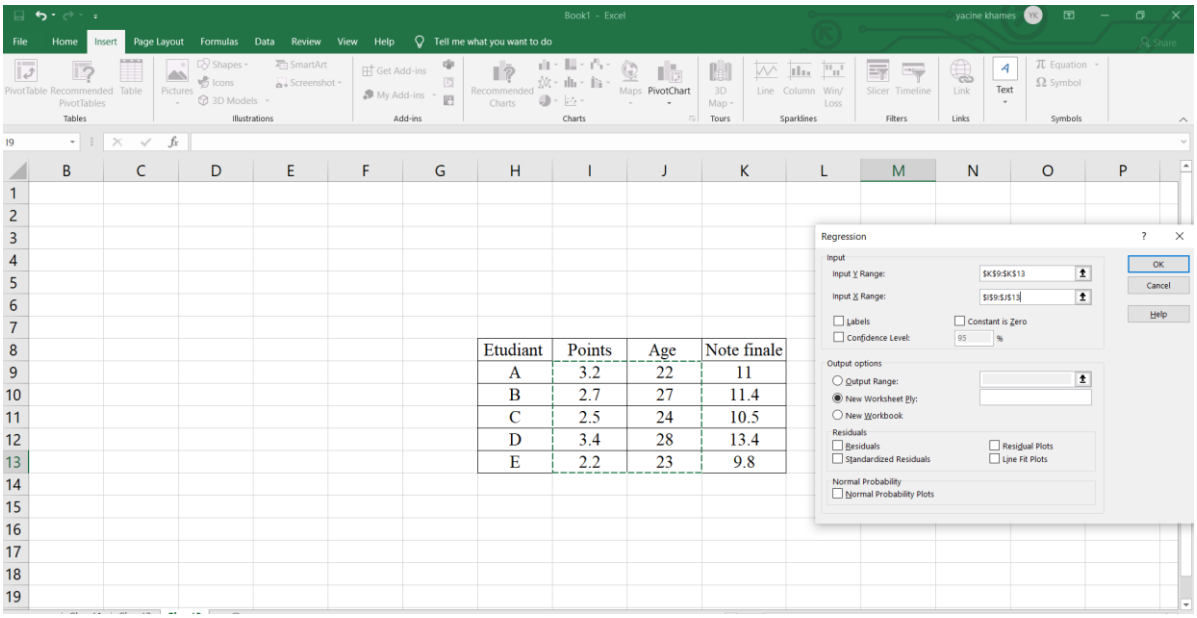
Etudiant	Points	Age	Note finale
A	3.2	22	11
B	2.7	27	11.4
C	2.5	24	10.5
D	3.4	28	13.4
E	2.2	23	9.8

Etape 3 : Lancez l'analyse en choisissant la variable dépendante (note finale), la case *Input y range* s'active. Pareil pour les variables indépendantes, la case *Input x range* s'active pour les deux variables points et âge.

Remarque :

Quand vous sélectionnez les colonnes, choisissez uniquement les valeurs (sans les caractères).

Cochez également l'option *Residuals plots* pour visualiser les diagrammes résiduels.



Etudiant	Points	Age	Note finale
A	3.2	22	11
B	2.7	27	11.4
C	2.5	24	10.5
D	3.4	28	13.4
E	2.2	23	9.8

Etape 4 : Les résultats ci-dessous, calculés par le logiciel Excel, contient tous les éléments pour identifier les coefficients (l'ordonnée à l'origine, les pentes).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.989288
R Square	0.978691
Adjusted F	0.957382
Standard E	0.280182
Observatic	5

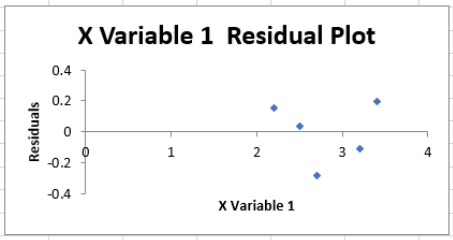
ANOVA

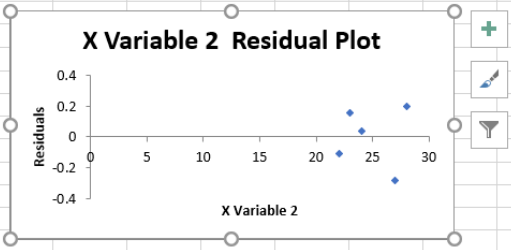
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	2	7.210996	3.605498	45.92885	0.021309
Residual	2	0.157004	0.078502		
Total	4	7.368			

	Coefficients	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-0.8962	1.384937	-0.64711	0.583916	-6.85511	5.062701	-6.85511	5.062701
X Variable	1.752803	0.304744	5.751728	0.028923	0.441597	3.064009	0.441597	3.064009
X Variable	0.290659	0.058275	4.987743	0.037925	0.039923	0.541396	0.039923	0.541396

RESIDUAL OUTPUT

)bservatior	*redicted Y*	*Residuals*
1	11.10727	-0.10727
2	11.68417	-0.28417
3	10.46163	0.038369
4	13.20179	0.198208
5	9.645131	0.154869





Alors, l'équation de régression multiple est $\hat{y} = -0.8962 + 1.75\,x_1 + 0.29\,x_2$

Le coefficient R^2 ajustée mesure l'ajustement de l'équation de régression multiple. R^2 = 0.97 proche de 1 et/ou p-value équivaut à 0.021 (*Anova*), ce qui signifie que nous avons un ajustement parfait et un modèle significatif.

Chapitre 4: Analyse de variance à un seul facteur et test de Scheffé

Objectifs :

- **Utiliser l'Anova à un facteur pour déterminer s'il existe une différence significative entre trois moyennes ou plus.**

- **Déterminer quelles moyennes diffère, en utilisant le test de Scheffé si l'hypothèse nulle est rejetée dans l'Anova.**
- **Utiliser l'Anova à deux facteurs pour déterminer s'il existe une différence significative dans les principaux effets**

Analyse de variance à un seul facteur et test de Scheffé

Lorsque les mesures d'une variable sont obtenues pour chacun de deux échantillons collectés indépendamment, des hypothèses H_0 : $\mu 1 = \mu 2$; H_1 : $\mu 1 \neq \mu 2$ sont appropriées.

Cependant, les biologistes obtiennent souvent des données sous la forme de trois échantillons ou plus, qui proviennent de trois populations ou plus, une situation qui nécessite des analyses multi-échantillons, comme présenté dans cette section.

Il semblerait que certains cherchent à tester des hypothèses multi-échantillons en appliquant des tests à deux échantillons à toutes les paires d'échantillons possibles. Mais une telle procédure, qui consiste à utiliser une série de deux tests d'échantillons pour traiter une hypothèse à plusieurs échantillons, n'est pas valable.

Exemple

Pour tester la comparaison de 3 moyennes deux à deux à la fois, 3 tests sont nécessaires, avec 3 risques d'erreurs α et trois taux de confiance $(1–\alpha)$. Et par conséquent, le taux de confiance total égaleà $(1-\alpha)^3$. Pour la comparaison de 5 moyennes deux à deux signifie que le taux de confiance total égale à $(1–\alpha)^5$. Et pour la comparaison de 10, le taux de confiance baisse considérablement et sera égale à $(1-\alpha)^{10}$.

Au fur et à mesure que le nombre de populations à comparer augmente, la probabilité de commettre une erreur type α en utilisant plusieurs tests de deux moyennes deux à deux pour un niveau de significativité donné α augmente également. Pour traiter ce problème, la technique d'analyse de la variance est utilisée.

La procédure utilisée pour tester $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

1-Utilisez la formule pour obtenir les résultats d'analyse :

$$\boldsymbol{F = \frac{Variance\ entre\ les\ groupes}{Variance\ intra\ groupes} = \frac{S_B^2}{S_W^2}} \qquad \textbf{(8)}$$

2-Identifier le *F* calculé et la p-value.

3- Concluez à partir de ces critères soit par la méthode a ou b:

Méthode a (la plus simple) :

Si $F_{Calculé} \geq F_{Critique}$, rejetez l'hypothèse nulle d'égalité des moyennes et concluez que la moyenne d'au moins une population est différente des autres.

Si $F_{Calculé} < F_{Critique,}$ ne pas rejetez l'hypothèse nulle d'égalité des moyennes.

Méthode b (**facultatif**)

Si la p-value ≤ risque (α), rejeter l'hypothèse nulle d'égalité des moyennes et concluez que

la moyenne d'au moins une population est différente des autres.

Si la p-value > risque (α), ne pas rejeter l'hypothèse nulle d'égalité des moyennes.

La variance entre (inter) les groupes (S_B^2) mesure les différences de moyennes qui résultent des différents traitements donnés (facteurs) de chaque groupe. Pour calculer cette valeur, il est nécessaire de trouver la moyenne générale.

Leddl_{Inter} entre les groupes égale à $k-1$, k est le nombre de groupes ou facteurs.

La formule de la variance entre-groupe est le rapport entre la Somme des Carrés Intergroupe (SCB) et le degré de liberté intergroupe ($ddl_{Inter} = k-1$)

$$\boldsymbol{S_B^2(Var_{Inter}) = \frac{SCB}{ddl_{inter}} = \frac{\sum n_i(\bar{X}_i-\bar{X}_G)^2}{k-1}} \quad \textbf{(9)}$$

k : nombre de groupes (facteurs ou traitements)

n_i : la taille de l'échantillon i

$\bar{X}_i$: la moyenne de l'échantillon i

$\bar{X}_G$: la moyenne générale de toutes les échantillons (la moyenne des moyennes $\bar{X}_i$)

La variance intra groupe (résiduelle ou erreur), désignée par S_W^2, permet de trouver la variance globale en calculant une moyenne pondérée des variances individuelle. Le ddl_{Intra} pour une variance intra groupe égale à $n-k$, n est le nombre d'observation (nombre de case dans le tableau).

La formule pour la variance intra-groupe est le rapport entre la Somme des Carrés Intragroupe (SCW) et le degré de liberté intragroupe ($ddl_{Intra} = n-k$)

$$\boldsymbol{S_W^2 = (Var_{Intra}) = \frac{SCW}{ddl_{Intra}} = \frac{\sum(n_i-1)\times S_i^2}{\sum(n_i-1)} = \frac{(n_1-1)\times S_1^2+(n_2-1)\times S_2^2+\cdots+(n_k-1)\times S_k^2}{(n_1-1)+(n_2-1)+\cdots+(n_k-1)}} \quad \textbf{(10)}$$

k : nombre de groupes (facteurs ou traitements)

n_i : la taille de l'échantillon i

S_i^2 : Variance de l'échantillon i

Remarque

La somme des carrés intergroupe + la somme des carrés intragroupe : $SCB + SCW = SCT$

La somme de ddl intergroupe + le ddl intragroupe est le ddl total : $k-1+n-k=n-1$

Déterminer la valeur du test *F* et conclusion

Etape 1
Trouver la moyenne de chaque échantillon

$$\bar{X}=\frac{\sum nix_i}{n}$$

Etape 2
Trouver la moyenne générale

$$\bar{X}_G=\frac{\sum x_i}{n} \text{ Ou bien } \bar{X}=\frac{\sum \bar{X}_i}{k}$$

Etape 3
Trouver la variance intergroupe

$$S_B^2(Var_{Inter})=\frac{SC_{Inter}}{k-1}=\frac{\sum n_i(\bar{X}_i-\bar{X}_G)^2}{k-1}$$

Etape 4

Trouver la variance intragroupe

$$S_W^2(Var_{Intra})=\frac{SC_{Intra}}{k-1}=\frac{\sum(x_i-\bar{X}_i)^2}{n-k}$$

Etape 5

Trouver la valeur $F_{Calculé}$

$$F=\frac{S_B^2}{S_W^2}$$

Etape 6

Trouver la valeur $F_{Critique}$ ont lisant la table de la loi de Fisher (Tableau 6) avec :

ddl entre groupes $ddl_{Inter}=k-1$ *ddl* inter groupes $ddl_{Intra}=n-k$

La $F_{Critique}$ est lue sur la table de la loi de Fisher

Etape 7

Si $F_{Calculé} \geq F_{Critique}$, rejetez l'hypothèse nulle d'égalité des moyennes et concluez que la moyenne d'au moins une population est différente des autres.

Si $F_{Calculé}<F_{Critique,}$ ne pas rejetez l'hypothèse nulle d'égalité des moyennes.

Toutes ces formules sont utilisées pour résumer l'analyse de la variance et sont placés dans un tableau récapitulatif, comme indiqué dans le tableau 3 :

Tableau 3. Résumé Anova

Type	***Somme des Carrés***	***ddl***	***Variance***	***F***
***Entre (Inter)* (Between)**	SCB	$ddl_B = k - 1$	S_B^2	$F = \frac{S_B^2}{S_W^2}$
Intra (erreur ou résiduel) (Within)	SCW	$ddl_W = n - k$	S_W^2	
Total	SCT	$ddl_T = n - 1$		

Remarque

L'hypothèse alternative signifie qu'au moins une moyenne diffère d'une ou plusieurs des autres moyennes et, comme indiqué précédemment, une valeur d'essai significative signifie qu'il est fort probable que cette différence de moyennes ne soit pas due au hasard, **mais elle n'indique pas où se trouve la différence**.

Facteur

Un facteur est une variable qualitative pour laquelle nous allons faire une comparaison entre les moyennes calculées à partir de variables quantitatives(Tableau 4).

Tableau 4. (a) variable indépendante (facteur) avec 4 modalités

(b) quatre échantillons ou modalités (facteur)

(a)

Facteur	Variable quantitative
Groupe A	Valeur
Groupe A	Valeur
Groupe A	Valeur
Groupe B	Valeur
Groupe B	Valeur
Groupe B	Valeur
Groupe C	Valeur
Groupe C	Valeur
Groupe C	Valeur
Groupe D	Valeur
Groupe D	Valeur
Groupe D	Valeur

(b)

Facteur			
Groupe A	Groupe B	Groupe C	Groupe D
Valeur	Valeur	Valeur	Valeur
Valeur	Valeur	Valeur	Valeur
Valeur	Valeur	Valeur	Valeur
$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$

Exemple

Le tableau ci-dessous montre l'adhésivité de quatre pages web, c'est-à-dire le nombre de secondes qu'un visiteur a passé sur la page web.

	Page 1	**Page 2**	**Page 3**	**Page 4**
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Moyenne ($\bar{X}$)	172.8	182.6	175.6	164.6
Moyenne générale ($\bar{X}_G$)	173.9			

Lorsque vous comparez seulement deux groupes, c'était une question simple ; il suffit juste d'examiner la différence entre les moyennes de chaque groupe. Avec quatre moyennes, il y a six comparaisons possibles entre les groupes.

Plus nous faisons de telles comparaisons par paires, plus le risque d'être trompé par le hasard est grand (voir l'exemple ci-dessus). Au lieu de nous préoccuper de toutes les différentes comparaisons entre les pages individuelles que nous pourrions éventuellement faire, nous pouvons faire un seul test global qui répond à la question suivante : Les pages pourraient-elles toutes avoir la même adhésivité sous-jacente ?

Méthode a :

Etape 1

Énoncer les hypothèses

H_0 : $\mu 1 = \mu 2 = \mu 3 = \mu 4$

H_1 : Au moins une moyenne est différente des autres

Etape 2

$$\bar{X}_G = \frac{\sum x_i}{n} = \frac{164+172+\cdots+170+164}{20} = \frac{172.8+182.6+175.6+164.6}{4} = 173{,}90$$

Etape 3

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Moyenne ($\bar{X}$)	172.8	182.6	175.6	164.6
Moyenne générale ($\bar{X}_G$)	173.90			

$\sum n_i = 5$ C'est le nombre de ligne de chaque groupe

$$S_B^2 = (\text{VarInter}) = \frac{\sum n_i(\bar{X}_i - \bar{X}_G)^2}{k-1}$$

$$= \frac{5 \times (172.8 - 173.90)^2 + 5 \times (182.6 - 173.90)^2 + 5 \times (175.6 - 173.90)^2 + 5 \times (164.6 - 173.90)^2}{4-1}$$

$$= \frac{831}{3} = 277{,}13$$

Etape 4

$$S_W^2 = (\text{VarIntra}) = \frac{\sum(n_i - 1) \times S_i^2}{\sum(n_i - 1)}$$

$$= \frac{\begin{array}{c}(164-172.8)^2+(172-172.8)^2+(177-172.8)^2+(156-172.8)^2+(195-172.8)^2+\\(178-182.6)^2+(191-182.6)^2+(182-182.6)^2+(185-182.6)^2+(177-182.6)^2+\\(175-175.6)^2+(193-175.6)^2+(171-175.6)^2+(163-175.6)^2+(176-175.6)^2+\\(155-164.6)^2+(166-164.6)^2+(164-164.6)^2+(170-164.6)^2+(168-164.6)^2\end{array}}{20-4}$$

$$= \frac{1618}{16} = 101.15$$

Etape 5

La statistique de F

$$F = \frac{S_B^2}{S_W^2} = \frac{277{,}13}{101.15} = 2{,}73$$

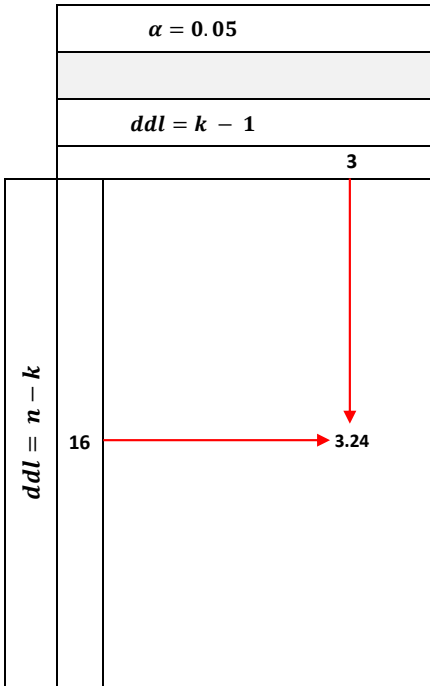
Etape 6

$$ddl_{Inter} = k - 1 = 4 - 1 = 3$$
$$ddl_{Intra} = n - k = 20 - 3 = 16$$

$F_{Critique}$ = 3,24 (voir la table de Fisher)

Type	*Somme des Carrés*	*ddl*	*Variance*	*F*	*FCritique*
Entre (Inter) (Between groups)	831	3	277,13	**2,73**	**3,24**
Intra (erreur ou résiduel) (Within groups)	1618	16	101,15		
Total	2450	19			

Schéma de la table de Fisher pour $\alpha = 0.05$



Remarque

La table de la loi de Fisher change en fonction du risque α. Ici nous ne nous pouvons pas présenter toutes les tables, nous allons travailler uniquement sur la tales de la loi de Fisher pour α 0,05.

Conclusion

Pour α = 0,05

La valeur du test étant 2,73 < 3,24, la décision est donc il n'y a pas suffisamment de preuves pour rejeter l'hypothèse nulle.

Méthode b (facultatif) :

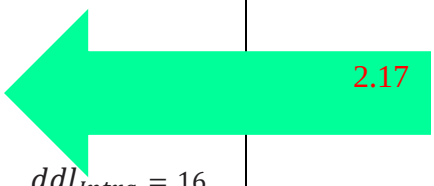
Cette méthode consiste à projeter la statistique calculée dans toutes les tables de la loi de Fisher ($\alpha = 0.05, \alpha = 0.025, \alpha = 0.1$). Le tableau ci-dessous montre les valeurs F lues sur toutes les tables de la loi de Fisher.

Les cinq premières étapes sont les mêmes que pour la méthode a, on reprend à l'étape 6.

La statistique calculée F est 2,71. A l'aide du tableau ci-dessous (résumé de toutes les tables de loi de Fisher), la p-value est comprise entre 0.1 et 0.05 ; soit une p-value supérieure à 0.05.

Comme pour la méthode a, on conclut qu'il n'y a pas suffisamment de preuve pour rejeter l'hypothèse nulle.

α	$ddl_{Inter} = 3$	
0.1	2.67	
0.05	3.24	$ddl_{Intra} = 16$
0.025	4.08	

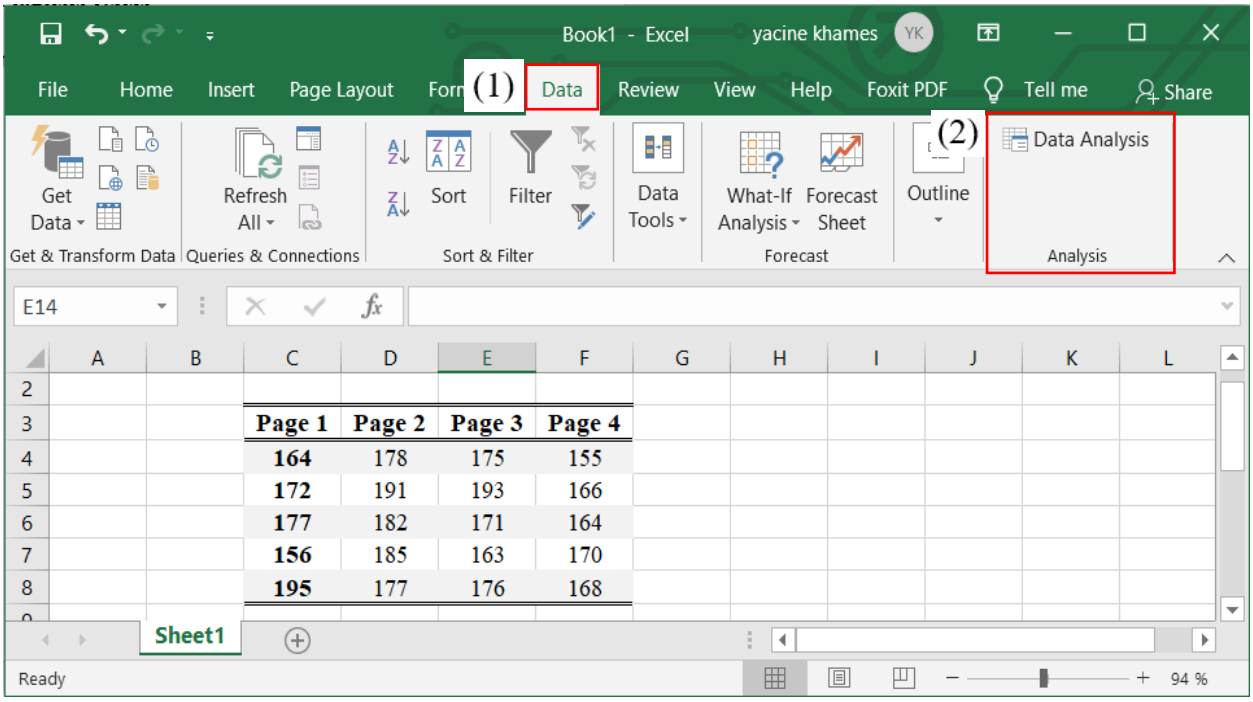


Analyse de variance à un facteur à l'aide du logiciel Excel

L'utilitaire d'analyse comprend les outils décrits dans les sections suivantes. Pour accéder à ces outils, cliquez sur (2) **Analyse** (***Analysis***) dans l'onglet (1) **Données** (***Data***).

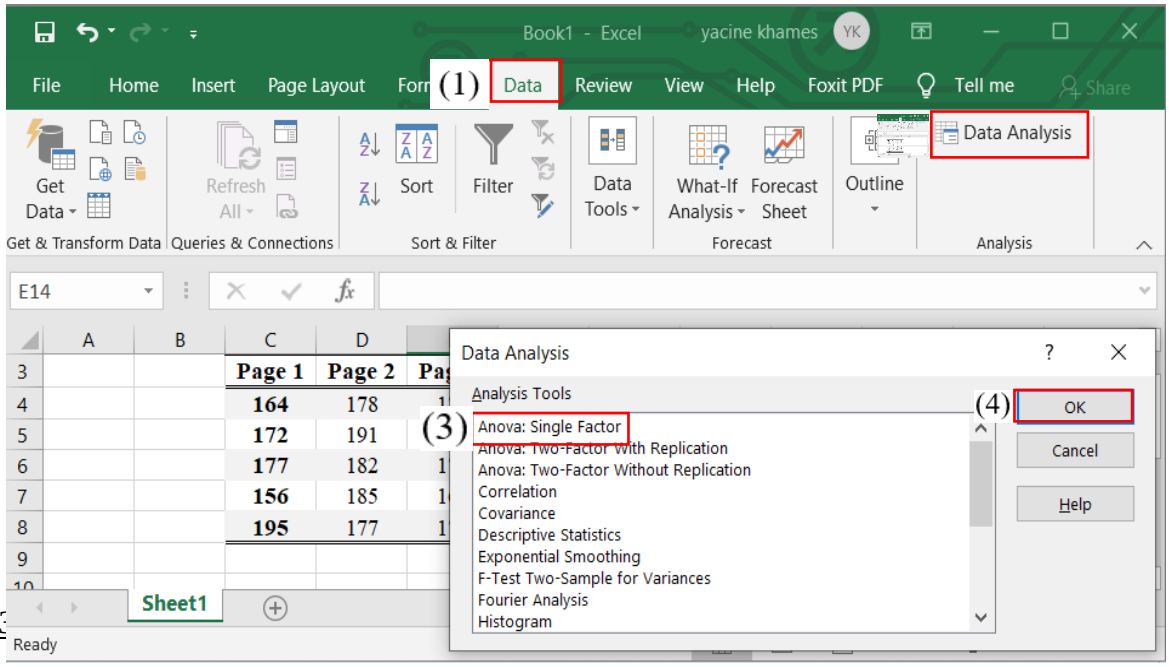
Etape 1

Ouvrer votre tableau sur la feuille Excel



Etape 2

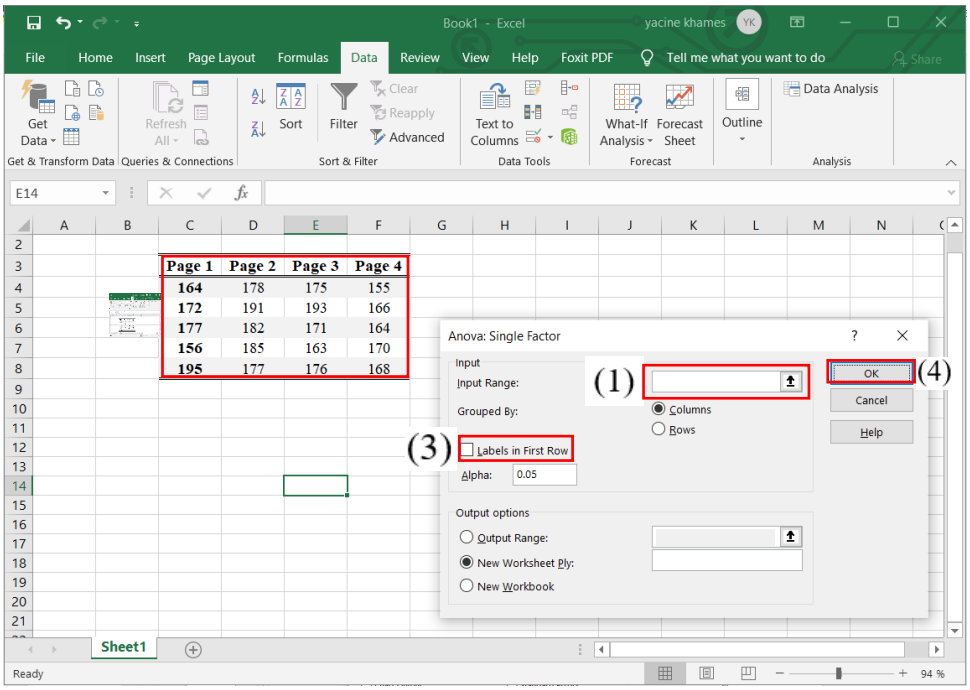
Choisir le menu : (1) **Données** (***Data***)+ (2) **Utilitaire d'analyse** (***Data Analysis***) + (3) **Anova à un facteur** (***Anova : Single Factor***) + (4) **OK**



Etape 3

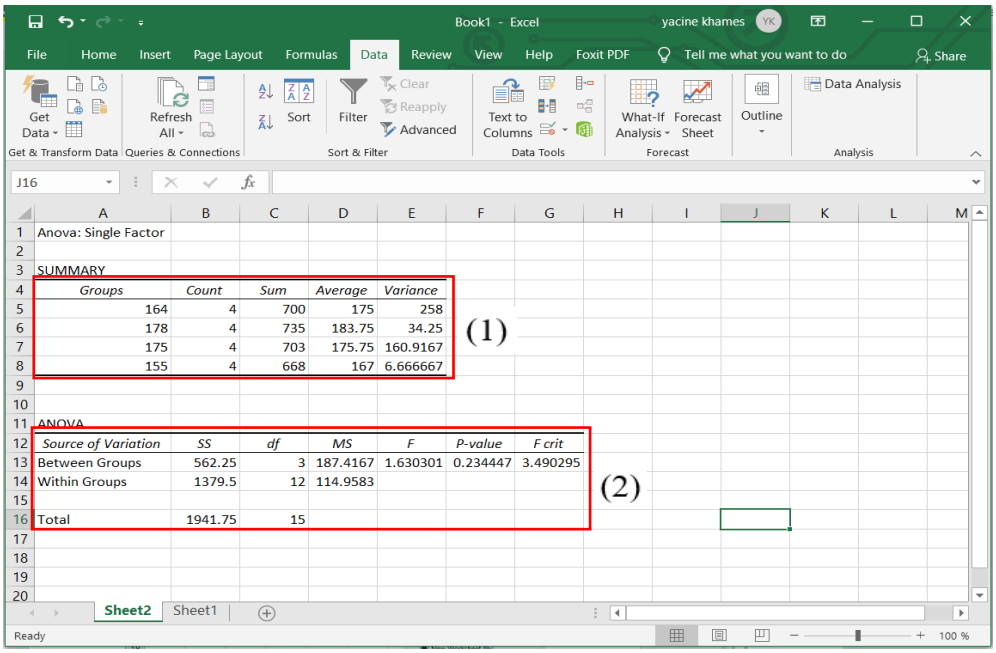
Cliquer ensuite sur la (1) **Page d'entrée** (***Input Range***), sélectionnez votre tableau (les quatre colonnes) en mettant le curseur de la souris sur (2) le tableau.

Cochez ensuite (3) **Intitulée en première ligne** (***Labels in First Row***) pour prendre en considération le nom des variables de la première ligne et puis cliquez sur (4) **OK.**



Résultat

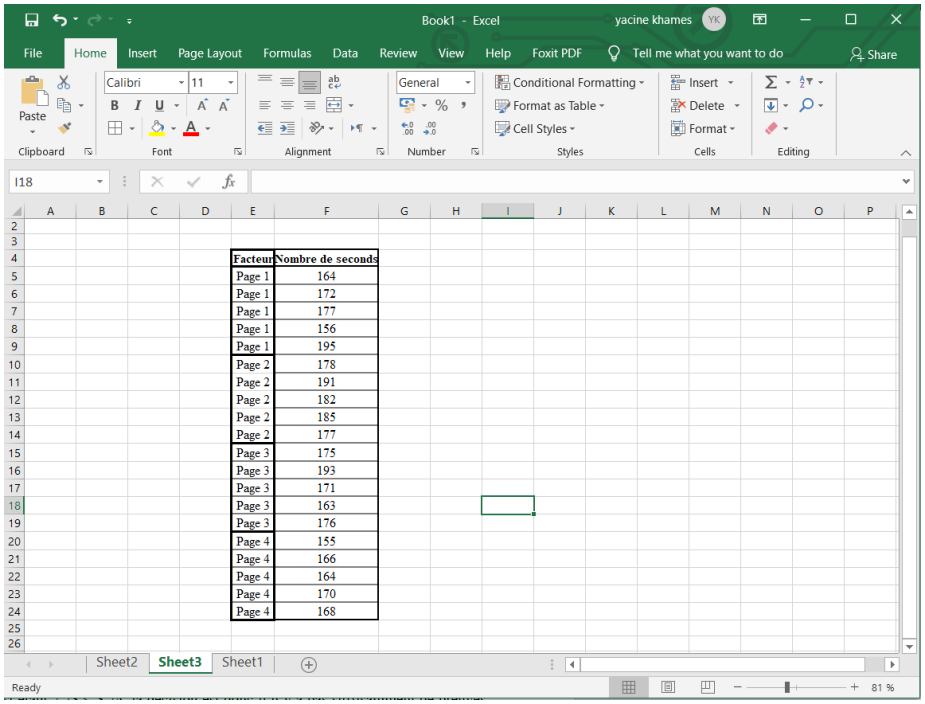
Le premier tableau (1) est un résumé statistique l'adhésivité de quatre pages
Le deuxième tableau (2) est le résultat d'analyse de l'Anova



Nous pouvons représenter graphiquement les données sous forme d'une boit à moustache (box plot).

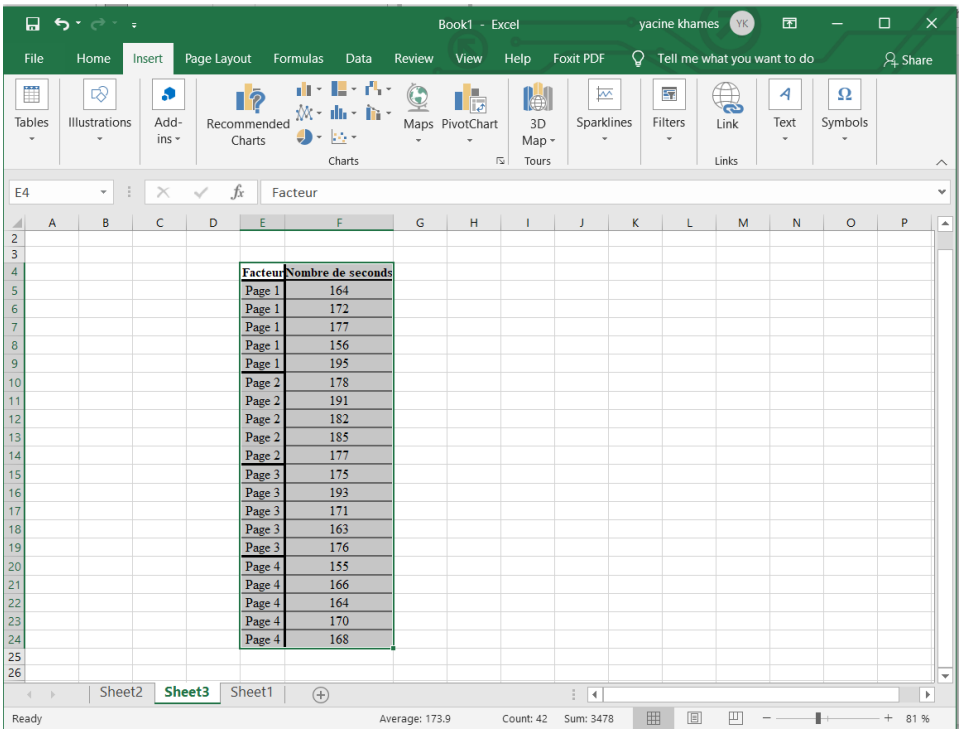
Etape 1

Transformer le tableau (a) en tableau (b) :



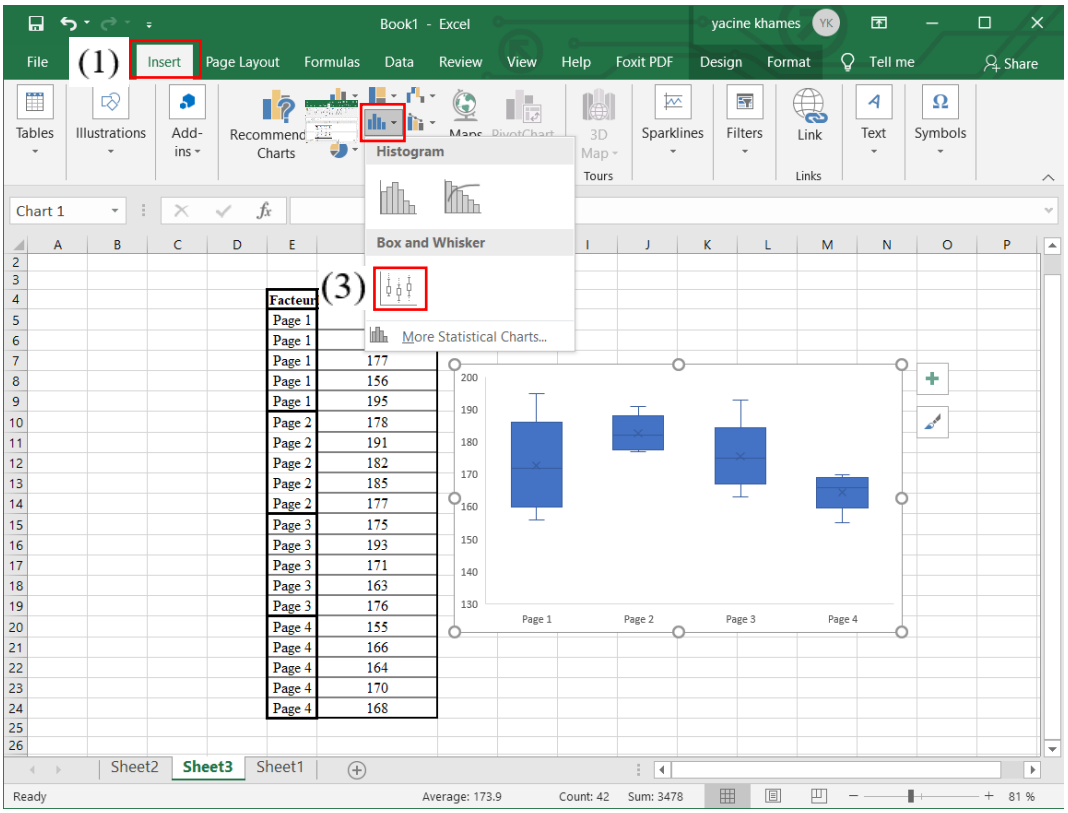
Etape 2

Sélectionner votre le tableau



Etape 3

Sur la barre d'outils, cliquez sur Insérer (1) (***Insert***) + (2) **Insérerun graphique statistique** (***Insert Statistic Chart***) + (3) **Boite à moustache** (**box plot**). Votre boite à moustaches apparaîtra directement dans votre feuille de calcul



Identification des moyennes qui sont différentes en utilisant le test de Scheffé

Après une analyse de variance (*Anova*), nous pouvons éventuellement rejeter l'hypothèse nulle du test *F*, mais le test Anova ne permet pas de savoir où se situe cette différence (quelle moyenne particulière est différente des autres ?). Plusieurs procédures ont été développées pour déterminer où se situent les différences significatives entre les moyennes après qu'une Anova ait été effectuée. Une méthode informelle (**facultatif**) est de construire soit des boites à moustaches (box plot) à la même échelle ou bien des intervalles de confiances, pour voir si l'un ou plusieurs se chevauchent.

Plusieurs méthodes formelles pour identifier les moyennes qui différent des autres. Parmi ces méthodes les plus couramment utilisés figurent le **test de Scheffé**.

Test de Scheffé

Pour réaliser le test Scheffé, vous devez comparer les moyennes deux à deux, en utilisant toutes les combinaisons possibles. Par exemple, s'il y a quatre moyennes, 6 comparaisons doivent être faites : $\bar{X}_1 \; et \; \bar{X}_2$; $\bar{X}_1 \; et \; \bar{X}_3$; $\bar{X}_1 \; et \; \bar{X}_4$; $\bar{X}_2 \; et \; \bar{X}_3$; $\bar{X}_2 \; et \; \bar{X}_4$; $\bar{X}_3 \; et \; \bar{X}_4$.

Etape 1 :
Calculez la statistique de Scheffé(F_s)

La formule de Scheffé est : $$F_s = \frac{(\bar{X}_i - \bar{X}_j)^2}{S_W^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \qquad (11)$$

$\bar{X}_i \; et \; \bar{X}_j$sont les moyennes des échantillons comparés : *i* et *j*

$n_i$etn_j sont respectivement les tailles des échantillons comparés

S_W^2est la variance intragroupe (résiduelle ou erreur)

Etape 2 :
Trouvez la valeur critique ($F_{Critique(Scheffé)}$)

Pour trouver la valeur critique pour le test de Scheffé ($F_{Critique(Scheffé)}$), multipliez la valeur critique pour le test d'anova($F_{Critique(anova)}$) par $(k-1)$
Après avoir calculé cette valeur pour une paire de donnée d'échantillon, vous la comparez à la valeur calculée F_s.

Si $F_s \geq F_{Critique(Scheffé)}$, rejetez l'hypothèse nulle d'égalité des moyennes
Si $F_s < F_{Critique(Scheffé)}$, ne pas rejetez l'hypothèse nulle d'égalité des moyennes.

Faites les mêmes étapes pour toutes les paires d'échantillons restantes.

Exemple

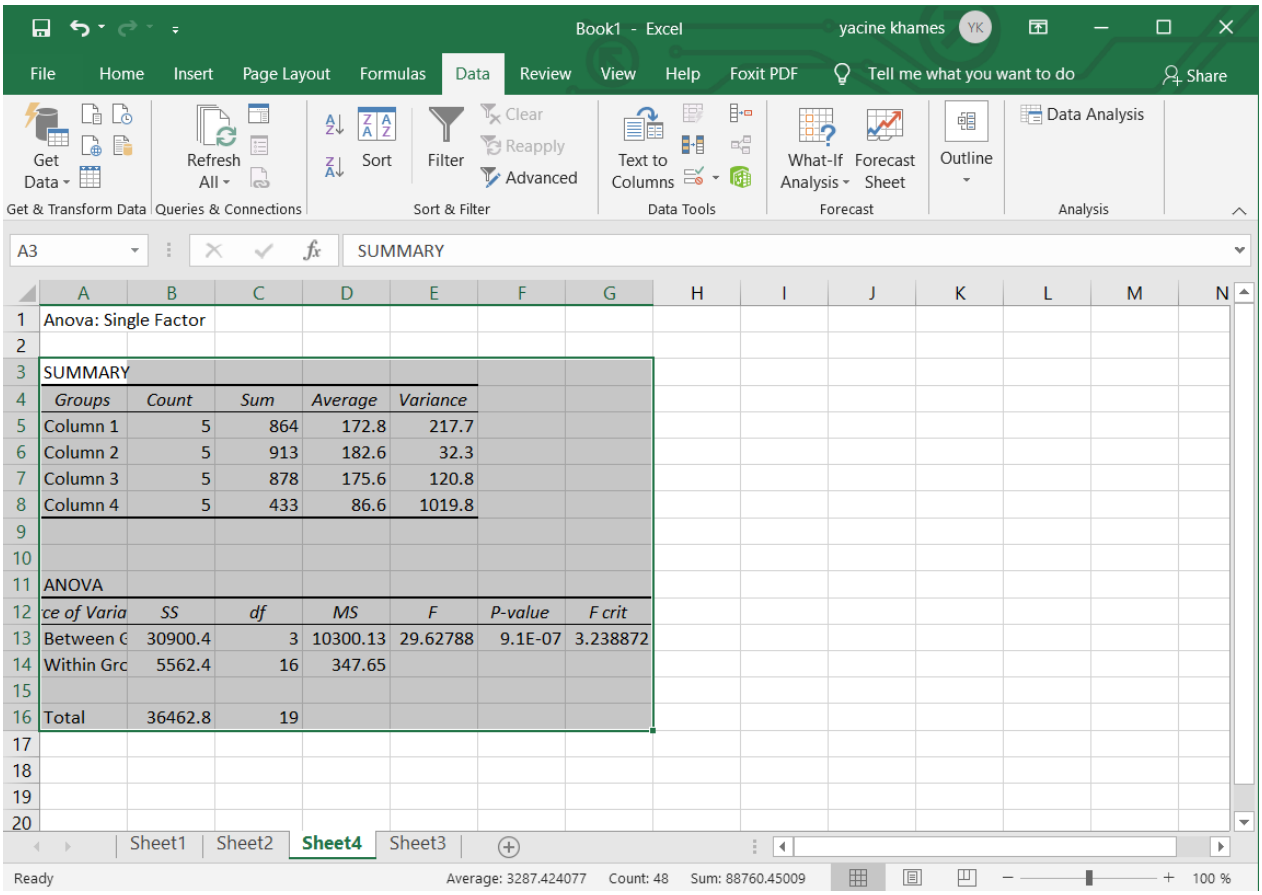
Utilisez le test de Scheffé 'l'adhésivité de quatre pages web' pour savoir s'il y a une différence significative entre les quatre pages web. Mais cette fois-ci nous allons prendre la page 5 au lieu de la page 4.

	Page 1	Page 2	Page 3	Page 5
	164	178	175	120
	172	191	193	90
	177	182	171	34
	156	185	163	100
	195	177	176	89
Moyenne ($\bar{X}$)	172.8	182.6	175.6	86.6

Solution

Faites une analyse de variance pour savoir s'il y a une différence significative entre les quatre moyennes.

Utiliser le logiciel pour effectuer une Anova (l'objectif ce n'est pas l'analyse de variance)



Le tableau d'Anova montre que la $F_{Calculé} = 29.62$ est supérieure à la valeur $F_{Critique} = 3.23$, c'est-à-dire rejet de l'hypothèse nulle. On conclut que la moyenne d'au moins un groupe est différent des autres

Pour identifier quelle moyenne particulière est différente des autres nous utiliserons le test de Scheffé.

Le test de Scheffé requiert un test séparé pour chaque paire possible c'est-à-dire 6 paires d'échantillons :

$$H_0: \bar{X}_1 = \bar{X}_2 \qquad H_0: \bar{X}_1 = \bar{X}_3 \qquad H_0: \bar{X}_1 = \bar{X}_5$$
$$H_0: \bar{X}_2 = \bar{X}_3 \qquad H_0: \bar{X}_2 = \bar{X}_5 \qquad H_0: \bar{X}_3 = \bar{X}_5$$

Nous commençons par $H_0 : \bar{X}_1 = \bar{X}_2$ l'exemple 'l'adhésivité de quatre pages web' nous indique que $\bar{X}_1$ =172.8 et $\bar{X}_2$ = 182.6 avec $n_i = n_j = 5$.

Etape 1

Nous pouvons calculer la statistique F_s

$S_W^2 = 347{,}65$, lue sur le tableau Anova.

$$F_s = \frac{(\bar{X}_i - \bar{X}_j)^2}{S_W^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = \frac{(172.8-182.6)^2}{347{,}65\left(\frac{1}{5}+\frac{1}{5}\right)} = 0.69$$

Etape 2

Le nombre de degrés de liberté intragroupe est $ddl = k - 1 = 4 - 1 = 3$

$(F_{Critique(Scheffé)} = F_{Critique(Anova)} \times (k-1)$

$(F_{Critique(Scheffé)} = 3.24 \times (4-1) = 9.72$

Le résultat ci-dessus montre que $0.69 < 9.72$; ne pas rejetez l'hypothèse nulle d'égalité des moyennes. On conclut qu'il n'y a pas suffisamment de preuves pour confirmer l'affirmation selon laquelle la moyenne $\bar{X}_1$ =172.8 est différente de la moyenne $\bar{X}_2$ = 182.6. Par conséquent, il n'y a pas de différence significative entre ces deux moyennes.

Nous pouvons faire un tableau pour résumer les résultats de comparaison de toutes les paires des échantillons.

Combinaison	F_s	$F_{Critique(Scheffé)}$	Rejet H_0
Page1-Page2	0.69	9.72	Non
Page1-Page3	0.05	9.72	Non
Page1-Page5	**53.43**	**9.72**	**Oui**
Page2-Page3	0.35	9.72	Non
Page2-Page5	**66.27**	**9.72**	**Oui**
Page3-Page5	**56.96**	**9.72**	**Oui**

Selon le test de Sheffé, la différence existe entre la page 1 et 5, la page 2 et 5, la page 3 et 5.

Anova à deux facteurs

Dans la section précédente, nous avons utilisé l'analyse de variance pour décider si plus deux de population avaient la même moyenne. Cette analyse était une analyse de variance à un seul facteur. La technique d'analyse de la variance présentée précédemment est appelée Anova à sens unique car il n'y a qu'une seule variable indépendante (page). L'Anova à deux facteurs implique deux variables indépendantes.

Par exemple, supposons qu'un chercheur souhaite tester les effets de deux types différents de **traitement** pour plantes (peupliers) et deux types de **sol** différents (humide et sec) (Tableaux ci-dessous).

Tableau 5. (a) Les donnéesprésentées en trois colonnes sont utilisées pour une analyse de variance à deux facteurs à l'aide du logiciel Excel. (b) données présentées en lignes et en colonnes

(a)

Facteur 1	Facteur 2	Variables quantitative
Type	Traitement	Poids du peupliers
Humide	Aucun	0.15
Humide	Aucun	0.02
Humide	Aucun	0.22
Sec	Aucun	0.15
Sec	Aucun	0.15
Sec	Aucun	0.15
Humide	Engrais	1.5
Humide	Engrais	1.38
Humide	Engrais	2.1
Sec	Engrais	1.34
Sec	Engrais	1.34
Sec	Engrais	1.34
Humide	Irrigation	0.2
Humide	Irrigation	0.02
Humide	Irrigation	0.23
Sec	Irrigation	0.23
Sec	Irrigation	0.23
Sec	Irrigation	0.23
Humide	Engrais et Irrigation	2.15
Humide	Engrais et Irrigation	2.03
Humide	Engrais et Irrigation	2.01
Sec	Engrais et Irrigation	2.03
Sec	Engrais et Irrigation	2.03
Sec	Engrais et Irrigation	2.03

(b)

		Traitement			
		Aucun	**Engrais**	**Engrais et Irrigation**	**Irrigation**
Type	**Humide**	0.15	1.5	2.15	0.2
		0.02	1.38	2.03	0.02
		0.22	2.1	2.01	0.23
	Sec	0.15	1.34	2.03	0.23
		0.15	1.34	2.03	0.23
		0.15	1.34	2.03	0.23

Procédure pour l'Anova à deux facteurs (le calcul manuel est facultatif)

H_0 : il n'y a pas d'effet d'interaction entre le type de sol et le traitement utilisé.

H_1 : Il existe un effet d'interaction entre le type de sol et le traitement utilisé.

Dans l'analyse de variance à deux facteurs, commencez à tester l'hypothèse nulle qu'il n'y a pas d'interaction entre les deux facteurs. En suivant ces étapes :

Etape 1

Calculez la Somme des carrés des écarts totales (SC_{Totale}) : $\sum n_i(x_i - \bar{X}_G)^2$ = 16.945

		Traitement			
		Aucun	Engrais	Engrais et Irrigation	Irrigation
Type	Humide	0.15	1.50	2.15	0.20
		0.02	1.38	2.03	0.02
		0.22	2.10	2.01	0.23
	Sec	0.15	1.34	2.03	0.23
		0.15	1.34	2.03	0.23
		0.15	1.34	2.03	0.23
	$\bar{X}$	0.14	1.5	2.046667	0.19
	$\bar{X}_G$	0.969			

SCT= $(0.15 - 0.96)^2 + (0.02 - 0.96)^2 \ldots \ldots . (0.23 - 0.96)^2$= 16.945

Etape 2

Somme des carrés des écarts intragroupe à deux facteurs (SC_{Intra})

=SC_{Humide}+SC_{Sec} = $\sum n_i(x_i - \bar{X})_{Humide} + \sum n_i(x_i - \bar{X})_{Sec}$= 0.355+0 =0.35

	Traitement			
	Aucun	Engrais	Engrais et Irrigation	Irrigation
Humide	0.15	1.50	2.15	0.20
	0.02	1.38	2.03	0.02
	0.22	2.10	2.01	0.23
$\bar{X}$	0.13	1.66	0.15	2.06

$SC_{Humide} = (0.15 - 0.13)^2 \ldots (0.22 - 0.13)^2 2 + (1.50 - 1.66)^2 2 \ldots (2.10 - 1.66)^2 + (2.15 - 0.15)^2 + . . (2.01 - 0.15)^2 + (0.20 - 2.06)^2 + \cdots + (0.23 - 2.06)^2$= 0.355

	Traitement			
	Aucun	Engrais	Engrais et Irrigation	Irrigation
Sec	0.15	1.34	2.03	0.23
	0.15	1.34	2.03	0.23
	0.15	1.34	2.03	0.23
$\bar{X}$	0.15	1.34	2.03	0.23

SC_{Humide} = $(0.15 - 0.15)^2$ … $(0.15 - 0.15)^2 2$ + $(1.34 - 1.34)^2 2$ … $(1.34 - 1.34)^2$ + $(2.03 - 2.03)^2$+.. $(2.03 - 0.2.03)^2$ + $(0.23 - 2.23)^2 + \cdots + (0.23 - 2.23)^2 = 0$

Etape 3

Calculez la Somme des carrés des écarts intergroupe (SC_{Inter}) : $\sum n_i(x_i - \bar{X}_G)^2$ = 16.34

		Traitement			
		Aucun	Engrais	Engrais et Irrigation	Irrigation
Type	Humide	0.15	1.5	2.15	0.2
		0.02	1.38	2.03	0.02
		0.22	2.1	2.01	0.23
	Sec	0.15	1.34	2.03	0.23
		0.15	1.34	2.03	0.23
		0.15	1.34	2.03	0.23
		0.14	1.50	2.046	0.19
		0.96			

$SC_{Inter} = 6 \times (0.14 - 0.96)^2 + 6 \times (1.50 - 0.96)^2 + 6 \times (2.046 - 0.96)^2 + 6 \times (0.19 - 0.96)^2$ =16.34

Etape 4

Calculez la Somme des carrés des écarts ligne/colonne = $\frac{(Somme^2{}_{ligne})}{n} - \frac{(Somme^2{}_{colonne})}{n}$

=22.56 − 22.54

= 0.024

		Traitement					
		Aucun	Engrais	Engrais et Irrigation	Irrigation		
Type	Humide	0.15	1.50	2.15	0.20	$Somme$ = 12.01 n = 12 $\frac{Somme^2}{n}$=12.02	$\frac{(Somme^2{}_{ligne})}{n}$ = 12.02+10.54 =22.56
		0.02	1.38	2.03	0.02		
		0.22	2.10	2.01	0.23		
	Sec	0.15	1.34	2.03	0.23	$Somme$ = 11.25 n = 12 $\frac{Somme^2}{n}$=10.54	
		0.15	1.34	2.03	0.23		
		0.15	1.34	2.03	0.23		
	∑	0.84	9	12.28	1.14	∑0 =23.26	
	n	6	6	6	6	n =24	
		$\frac{(Somme^2{}_{colonne})}{n} = \frac{23.26^2}{24}$= 22.54					

Etape 5

Somme des carrés des écarts interaction

$SC_{Totale} = SC_{Intra} + SC_{Inter} + SC_{ligne/colonne} + SC_{Interaction}$

$SC_{Interaction} = SC_{Totale} - (SC_{Intra} + SC_{Inter} + SC_{ligne/colonne})$

$SC_{Interaction} = 16.94 - 0.35 - 16.34 - 0.024 = 0.141$

Etape 6

$ddl_{Totale} = n - 1 = 24 - 1 = 23$

$ddl_{Inter} = k - 1 = 3$ $\quad$ $ddl_{LigneColone} = l - 1 = 2 - 1 = 1$

Avec k le nombre de facteur colonne ; l est le nombre facteur ligne

$ddl_{Interaction} = (k - 1) \times (l - 1) = 3 \times 1 = 3$

$ddl_{Intra} = ddl_{Totale} - (ddl_{Inter} + ddl_{Ligne/colonne} + ddl_{Interaction})$

$ddl_{Intra} = 23 - (3 + 1 + 3) = 16$

Etape 7

Trois statistiques sont déterminées

$$F_{Interaction} = \frac{\frac{SC_{Interaction}}{ddl_{Interaction}}}{\frac{SC_{Intra}}{ddl_{Intra}}} = \frac{\frac{0.141}{3}}{\frac{0.35}{16}} = 2.13$$

On peut avoir également les autres statistiques

$$F_{Traitement} = \frac{\frac{SC_{Inter}}{ddl_{Inter}}}{\frac{SC_{Intra}}{ddl_{Intra}}} = \frac{\frac{16.34}{3}}{\frac{0.35}{16}} = 250.20$$

$$F_{Type} = \frac{\frac{SC_{Ligne/colonne}}{ddl_{Ligne/colonne}}}{\frac{SC_{Intra}}{ddl_{Intra}}} = \frac{\frac{0.024}{1}}{\frac{0.35}{16}} = 1.09$$

L'Anova à deux facteurs appliqués aux données du tableau est utilisée pour étudier trois effets possibles sur les poids des peupliers. (1) effet interaction entre traitement et type de sol ; (2) effet du traitement ; (3) effet type de sol. Les calculs sont complexes (voir l'exemple ci-dessus) et nécessitent l'emploi d'un logiciel. Les résultats de l'analyse à l'aide du logiciel Excel sont présentés ci-après :

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	0.024067	1	0.024067	1.083271	0.31344	4.493998
Columns	16.42445	3	5.474817	246.4284	1.34E-13	3.238872
Interaction	0.1414	3	0.047133	2.12153	0.137637	3.238872
Within	0.355467	16	0.022217			
Total	16.94538	23				

Chapitre 5: Test de Khi 2 (Chi carré)

Objectifs :

- **Tester une distribution pour déterminer la qualité de l'ajustement (goodness of fit), en utilisant le test de Khi II.**
- **Tester deux variables pour l'indépendance, en utilisant le test de Khi 2.**

Tester l'homogénéité des proportions à l'aide du khi-deux (Chi carré)

La statistique du chi carré peut être utilisée pour voir si une distribution de fréquences correspond à un modèle spécifique.

Exemple :

Supposons que vous vouliez voir s'il y a une différence dans le nombre d'arrestations dans une certaine ville pour quatre types de crimes. Un échantillon aléatoire de 160 arrestations a montré la répartition suivante.

Vol	*Crime*	*Consommation de drogues*	*Conduire sous l'influence de l'alcool*
36	48	26	42

Les fréquences pour chaque cas de crime ont été obtenues à partir d'un échantillon, ces fréquences réelles sont appelées les **fréquences observées**. Les fréquences obtenues par calcul sont appelées les **fréquences théoriques**. Pour calculer les fréquences théoriques, il y a deux règles à suivre :

1. Si toutes les **fréquences théoriques** sont égales, la fréquence théorique ($\boldsymbol{t}$) peut être calculée en utilisant

$$t = \frac{n}{k} \qquad (12)$$

$n$est le nombre total d'observations

$k$est le nombre de catégories

2. Si toutes les **fréquencesthéoriques** ne sont pas égales, alors la fréquence théorique ($\boldsymbol{t}$) peut être calculée par

$$\boldsymbol{t = n \times p} \qquad \textbf{(13)}$$

$n$est le nombre total d'observations

$p$est la probabilité pour cette catégorie.

Test d'adéquation de Khi deux (χ^2)

En regardant l'exemple du nombre d'arrestations, s'il n'y avait pas de différence, on s'attendrait à $^{152}/_{4} = 38$ arrestations pour chaque catégorie. Un tableau complet s'affiche.

	Vol	***Crime***	***Consommation de drogues***	***Conduire sous l'influence de l'alcool***
Observé	36	48	26	42
Théorique	38	38	38	38

Il y a toujours une différence entre les fréquences observées et théoriques. Les différences entre les différentes catégories sont-elles significatives ? (Y a-t-il une différence dans le nombre d'arrestations pour ces types de crimes) ou sont-elles dues au hasard ?

Le test de qualité d'ajustement du khi deux permettra de déterminer la réponse.

Les étapes

Etape 1

H_0 : Il n'y a aucune différence significative

H_1 : Il y a une différence significative

Etape 2

Formule pour le test du khi deux

$$\chi^2 = \sum \frac{(O-T)^2}{T} \qquad \textbf{(14)}$$

ddlégale (l-1)*(k-1)

O est la fréquence observée

T est la fréquence Théorique

Etape 3

Lire la valeur critique sur la table de khi deux (χ^2)

Etape 4

Conclusion

$\chi^2_{Calculé} < \chi^2_{Critique}$;ne pas rejeter H_0

$\chi^2_{Calculé} > \chi^2_{Critique}$; H_0 rejetée

Solution

Etape 1

H_0 : il n'y a pas de différence dans le nombre d'arrestations pour chaque type de crime.

H_1 : Il y a une différence dans le nombre d'arrestations pour chaque type de crime.

Etape 2

Trouvez la somme de χ^2 avec la formule de $\chi^2 = \sum \frac{(O-T)^2}{T}$

	Vol	***Crime***	***Consommation de drogues***	***Conduire sous l'influence de l'alcool***
Observé	36	48	26	42
Théorique	38	38	38	38
χ^2	$\frac{(36-38)^2}{38}$= 0.10	$\frac{(48-38)^2}{38}$=2.63	$\frac{(26-38)^2}{38}$=3.78	$\frac{(42-38)^2}{38}$ = 0.42
$\sum \chi^2$	$= 0.10 + 2.63 + 3.78 + 0.42$ = 6.93			

Etape 3

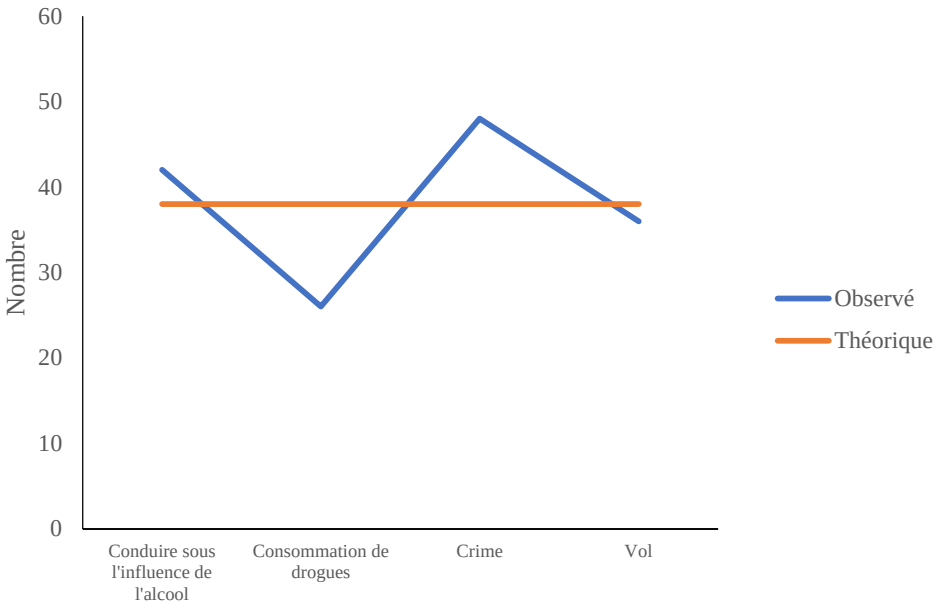
La valeur critique est de 7.378 pour $ddl = 4 - 1\ et\ \alpha = 0.05$

	α									
ddl	**0.995**	**0.99**	**0.975**	**0.95**	**0.9**	**0.1**	**0.05**	**0.025**	**0.01**	**0.**
1	–	–	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.
2	0.01	0.02	0.051	0.103	0.211	4.605	5.991	7.378	9.21	10
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12

Etape 4

Conclusion

6.93 < 7.815 $\gg\gg {\chi^2}_{Calculé} < {\chi^2}_{Critique}$, ne pas rejeter H_0; Il n'y a pas suffisamment de preuves pour rejeter l'affirmation selon laquelle il n'y a pas de différence dans le nombre d'arrestations pour chaque type de crime.



Test de Khi deux (χ^2) à l'aide d'un tableau de contingence

Lorsque les données peuvent être présentées sous forme de tableau en termes de fréquences, plusieurs types d'hypothèses peuvent être vérifiées à l'aide du test du khi deux pour le test d'indépendance des variables et pour le test d'homogénéité des proportions.

Test d'indépendance

Le test d'indépendance du khi deux est utilisé pour vérifier si deux variables sont indépendantes l'un de l'autre.

Les étapes

Etape 1

H_0 : Les variables sont indépendantes les unes des autres.

H_1 : Les variables sont dépendantes les unes des autres.

Etape 2

Trouvez la fréquence théorique

T est la fréquence Théorique :

$$T = \frac{(Somme\ des\ lignes) \times (Somme\ des\ colonnes)}{Somme\ totale} \qquad (15)$$

Etape 3

Formule pour le test du khi deux

$$\chi^2 = \sum \frac{(O-T)^2}{T}$$

Etape 4

ddlégale (l-1)*(k-1)

Lire la valeur critique sur la table de khi deux (χ^2)

Etape 5

Conclusion

$\chi^2{}_{Calculé} < \chi^2{}_{Critique}$;ne pas rejeter H_0

$\chi^2{}_{Calculé} > \chi^2{}_{Critique}$; H_0 rejetée

Exemple

Supposons, par exemple, qu'une nouvelle procédure de traitement postopératoire soit administrée à un certain nombre de patients dans un grand hôpital.

Le chercheur peut se poser la question suivante : les médecins ont-ils une opinion différente de celle des infirmières sur cette procédure, ou ont-ils essentiellement la même opinion ?

Groupe	Préfère la nouvelle procédure	Préfère l'ancienne procédure	Pas de préférence
Infirmières	200	180	120
Docteurs	150	220	130

Solution

Etape 1

H_0 : Les variables sont indépendantes les unes des autres.

H_1 : Les variables sont dépendantes les unes des autres.

Etape 2

A- Les fréquences observées, sommes des lignes, somme des colonnes et la somme totale

Groupe	Préfère la nouvelle procédure	Préfère l'ancienne procédure	Pas de préférence	Somme
Infirmières	200	180	120	500
Docteurs	150	220	130	500
Somme	350	400	250	1000

B- Les fréquences théoriques : $T = \frac{(Somme\ des\ lignes) \times (Somme\ des\ colonnes)}{Somme\ totale}$

Groupe	Préfère la nouvelle procédure	Préfère l'ancienne procédure	Pas de préférence
Infirmières	$\frac{500\times350}{1000} = 175$	$\frac{500\times400}{1000} = 200$	$\frac{500 \times 250}{1000} = 125$
Docteurs	$\frac{500 \times 350}{1000} = 175$	$\frac{500 \times 400}{1000} = 200$	$\frac{500\times250}{1000} = 125$

Etape 3

Trouvez la somme de χ^2 avec la formule de $\chi^2 = \sum \frac{(O-T)^2}{T} = 3.75 + 2 + 0.2 + 3.75 + 2 + 0.2$

$$\chi^2 = 11.9$$

Groupe	Préfère la nouvelle procédure	Préfère l'ancienne procédure	Pas de préférence
χ^2	$\frac{(200-175)^2}{175} = 3.57$	$\frac{(180-200)^2}{200} = 2$	$\frac{(120-125)^2}{125} = 0.2$
χ^2	$\frac{(150-175)^2}{175} = 3.57$	$\frac{(220-200)^2}{200} = 2$	$\frac{(130-125)^2}{125} = 0.2$

Etape 4

La valeur critique est de 5.991 pour $ddl = (2-1) \times (3-1) = 2\ et\ \alpha = 0.05$

	α									
ddl	**0.995**	**0.99**	**0.975**	**0.95**	**0.9**	**0.1**	**0.05**	**0.025**	**0.01**	**0.**
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.
2	0.01	0.02	0.051	0.103	0.211	4.605	5.991	7.378	9.21	10
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12

Etape 5

Conclusion

11.9 > 5.991 $\gg\gg \chi^2_{Calculé} > \chi^2_{Critique}$, H_0 rejetée ;La conclusion est qu'il existe suffisamment de preuves pour soutenir l'affirmation selon laquelle l'opinion est liée à la profession (dépendante), c'est-à-dire que les médecins et les infirmières ont des opinions différentes sur la procédure.

Test de Khi deux (χ^2) pour des proportions

Le deuxième test du khi deux qui utilise un tableau de contingence est appelé test d'homogénéité des proportions.

Le test d'homogénéité des proportions est utilisé pour vérifier l'affirmation selon laquelle différentes populations ont la même proportion.

Exemple
Un psychologue a sélectionné au hasard 100 personnes dans chacune des quatre catégories de revenus et leur a demandé si elles étaient "très heureuses". Pour les personnes qui gagnaient moins de 30 000 dollars, 24 % ont répondu par l'affirmative. Pour les personnes qui gagnent entre 30 000 et 74 999 dollars, 33 % ont répondu oui. Pour les personnes qui gagnent entre 75 000 et 90 999 dollars, 38 % ont répondu oui, et pour les personnes qui gagnent 100 000 dollars ou plus, 49 % ont répondu oui.

Sur α = 0,05, testez l'affirmation selon laquelle il n'y a pas de différence dans la proportion de personnes de chaque groupe économique qui étaient très heureuses.

	Moins de 3000	Entre 30 000 et 74 999	Entre 75 000 et 90 999	100 000 ou plus
Oui	$\frac{24}{100} = 24\%$	$\frac{33}{100}$	$\frac{38}{100}$	$\frac{49}{100}$
Non	$\frac{76}{100} = 76\%$	$\frac{67}{100}$	$\frac{62}{100}$	$\frac{51}{100}$

Étape1
H_0: p1 = p2 = p3 = p4
H1 : Au moins une proportion diffère des autres**.**
Étape 2
Trouvez la valeur critique. La formule pour les degrés de liberté est la même comme auparavant : $k - 1 = 3$. La valeur critique est donc 7.815.
Étape 3
Calculer la valeur du test χ^2. Puisque nous voulons vérifier l'affirmation selon laquelle les proportions sont égales, nous utilisons les calculs attendus comme suit :

A- Les fréquences observées, sommes des lignes, somme des colonnes et la somme totale

	Moins de 3000	Entre 30 000 et 74 999	Entre 75 000 et 90 999	100 000 ou plus	Total
Oui	24	33	38	49	144
Non	76	67	62	51	256
Total	100	100	100	100	400

B- Les fréquences théoriques : $T = \frac{(Somme\ des\ lignes) \times (Somme\ des\ colonnes)}{Somme\ totale}$

	Moins de 3000	Entre 30 000 et 74 999	Entre 75 000 et 90 999	100 000 ou plus
Oui	$\frac{100 \times 144}{400} = 36$	$\frac{100 \times 144}{400} = 36$	$\frac{100 \times 144}{400} = 36$	$\frac{100 \times 144}{400} = 36$
Non	$\frac{100 \times 256}{400} = 64$	$\frac{100 \times 256}{400} = 64$	$\frac{100 \times 256}{400} = 64$	$\frac{100 \times 256}{400} = 64$

Etape 4

Trouvez la somme de χ^2 avec la formule de $\chi^2 = \sum \frac{(O-T)^2}{T} = 4 + 0.25 + 0.1 + 4.6 + 2.25 + 0.14 + 0.062 + 2.64$

$\chi^2 = 14.04$

	Moins de 3000	Entre 30 000 et 74 999	Entre 75 000 et 90 999	100 000 ou plus
Oui	$\frac{(24-36)^2}{36} = 4$	$\frac{(33-36)^2}{36} = 0.25$	$\frac{(38-36)^2}{36} = 0.1$	$\frac{(49-36)^2}{36} = 4.6$
Non	$\frac{(76-64)^2}{64} = 2.25$	$\frac{(67-64)^2}{64} = 0.14$	$\frac{(62-64)^2}{64} = 0.062$	$\frac{(51-64)^2}{64} = 2.64$

Etape 5

Conclusion

14.04 > 7.815 »»» $\chi^2_{Calculé} > \chi^2_{Critique}$, H_0 rejetée ;on conclut qu'il existe suffisamment de preuves pour soutenir l'affirmation selon laquelle il n'y a pas de différence dans les proportions. Par conséquent, les revenues semblent faire une différence dans les proportions.

Chapitre 6: Test non paramétriques

Objectifs :

- **Tester les hypothèses, en utilisant le test des rangs signés de Wilcoxon pour données appariées**
- **Tester les hypothèses, en utilisant le test des rangs de Wilcoxon pour deux échantillons indépendants.**
- **Tester les hypothèses en utilisant le test de Kruskal-Wallis.**
- **Calculer le coefficient de corrélation de rang de Spearman**

Les tests statistiques, tels que les tests z, t et Anova, sont appelés **tests paramétriques**. Ce sont des tests statistiques pour les paramètres de la population tels que les moyennes, les variances et les proportions qui impliquent des hypothèses sur les populations à partir desquelles les échantillons sont **normalement distribués** (suivent la loi normale).

Mais que se passe-t-il si la population dans une situation particulière **n'est pas normalement distribuée** ?

Les statisticiens ont développé une branche de statistiques appelée **statistiques non paramétriques** ou statistiques sans répartition à utiliser lorsque la population dans laquelle les échantillons sont sélectionnés **n'est pas normalement distribué** ou est distribué de toute autre manière particulière.

Remarque

Les tests paramétriques comme nous l'avons vu auparavant, toutes les formules contiennent les paramètres de la population (μ, σ ou p).

Test des rangs signés de Wilcoxon pour données appariées (facultatif)

Test non paramétrique utilise les rangs des données d'échantillons appariées.

Procédure

Etape 1

H_0 : pas de différence significative

H_1 : la différence est significative

Etape 2

a) Trouver les différences d.

b) Classer chaque valeur absolue de la plus basse à la plus haute (rangs).

c) Donner à chaque rang un signe positif ou négatif, selon le signe figurant dans **(b**).

d) Trouver la somme des rangs positifs et la somme des rangs négatifs séparément.

e) Sélectionner la plus petite des valeurs absolues des sommes des rangs **(e)**, et utilisez.cette valeur absolue comme valeur de test T.

Etape 3

Si $n \leq 30$, la statistique de test est T.

Si $n \geq 30$, la statistique de test est $$Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \qquad \textbf{(16)}$$

$n$est le nombre de paires

Etape 4

La valeur calculée va être comparer avec la valeur critique lue sur la table de Wilcoxon (voir la table ci-dessous)

Si $T \geq T_{Critique}$, ne pas rejeter H_0

Si $T < T_{Critique}$, H_0 rejetée.

Remarque Si $n \geq 30$.

$Z < 0.05$**,** ne pas rejeter H_0

$Z \geq \mathbf{0.05}$**,** H_0 rejetée

Table de Wilcoxon (Si $n \leq 30$)

	α			
n	**0.1**	**0.05**	**0.02**	**0.01**
5	1			
6	2	1		
7	4	2	0	
8	6	4	2	0
9	8	6	3	2
10	11	8	5	3
11	14	11	7	5
12	17	14	10	7
13	21	17	13	10
14	26	21	16	13
15	30	25	20	16
16	36	30	24	19
17	41	35	28	23
18	47	40	33	28
19	54	46	38	32
20	60	52	43	37
21	68	59	49	43
22	75	66	56	49
23	83	73	62	55
24	92	81	69	61
25	101	90	77	68
26	110	98	85	76
27	120	107	93	84
28	130	117	102	92
29	141	127	111	100
30	152	137	120	109

Exemple

Les données du tableau ci-dessous fournissent 2 types de grains de maïs (ordinaires et séchées au four). Utilisez le test des rangs signés de Wilcoxon avec un niveau de significativité de 0.05 pour tester qu'il n'y a pas de différence entre les rendements des de types de maïs.

Ordinaires	2203	2235	2210	2796	2408	2261
Séchées au four	2309	2215	2311	2763	2480	2225

Solution

Etape 1

H_0 : pas de différence significative

H_1 : la différence est significative

Etape 2

Ordinaires	2203	2235	2210	2796	2408	2261
Séchées au four	2309	2215	2311	2763	2480	2225
a) Différence (d)	$2203 - 2309 = -106$	20	-101	33	-72	36
b) Classement par ordre croissant (rangs)	1	4	2	5	3	6
c) Rangs signés	-1	4	-2	5	-3	6

d) La somme des rangs absolus négatifs $= 1 + 2 + 3 = 6$

La somme des rangs absolus positifs $= 4 + 5 + 6 = 15$

e) La valeur de T est la plus petits, donc $T = 6$.

Etape 3

Puisque n (nombre de paires) égale à 6, $n \leq 30$, La statistique de test $T = 6$.

Etape 4

La statistique $(T = 6) > (T_{Critique} = 1)$, Ne pas rejetée H_0. Il semble qu'il n'y a pas de différence significative entre les deux types de maïs.

Test des rangs de Wilcoxon pour deux échantillons indépendants (Mann-Whitney)

Test non paramétrique qui utilise les rangs de deux populations indépendantes.

Procédure

Etape 1

H_0 : pas de différence significative
H_1 : la différence est significative

Etape 2

a) Combiner temporairement les deux échantillons en un seul
b) Remplacer chaque valeur de l'échantillon par son rang(la valeur la plus basse vers la plushaute)

Etape 3

Calculer la valeur Z de la statistique de test

$$\mu_R = \frac{n_1(n_1+n_2+1)}{2} \qquad (17)$$

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}} \qquad (18)$$

$$Z = \frac{R_1-\mu_R}{\sigma_R} \qquad (19)$$

n_1 : effectif de l'échantillon 1

n_2 : effectif de l'échantillon 2

R_1 : somme des rangs de l'échantillon 1

Etape 4

La valeur calculée absolue est comparée avec la valeur critique $U_\alpha = 1.96$ pour $\alpha = 0.05$.

Si $|Z| < 1.96$, ne pas rejeter H_0

Si $|Z| < 1.96$, H_0 rejetée

Exemple

Le nombre d'accidents de voiture est indiqué dans le tableau ci-dessous. À α = 0,05, y a-t-il une différence dans le nombre d'accidents dans ces deux régions ?

Région 1	12	20	10	10	13	8
Région 2	16	7	12	8	5	6

Solution

Etape 1

H_0 : pas de différence significative
H_1 : la différence est significative

Etape 2

Région 1	12	20	9	10	12	8
Rang 1 (	**8.5**	12	6	7	**8.5**	**4.5**
Région 2	16	7	13	8	5	6
Rang 2	11	3	10	**4.5**	1	2

Remarque

Si deux valeurs occupent le même rang, vous le divisez alors en deux puis vous sautez le rang suivant.

Etape 3

n_1 =6 R_1 = 8.5 + 12 + 6 + 7 + 8.5 + 4.5 = 46.5

n_2 = 6

Calculer la valeur Z de la statistique de test

$\mu_R = \frac{6\times(6+6+1)}{2} = 39$

$\sigma_R = \sqrt{\frac{6\times6(6+6+1)}{12}} = 6.24$

$$Z = \frac{46.5 - 39}{6.24} = 1.20$$

Etape 4

$|1.20| < 1.96$, ne pas rejeter H_0.Il existe suffisamment de preuves pour soutenir l'affirmation selon laquelle il n'y a pas de différence dans le nombre d'accident entre les deux régions.

Kruskal-Wallis (ANOVA non-paramétrique)

Le test de Kruskal-Wallis est un test non paramétrique permet de savoir si trois échantillons ou plus viennent de la même population.

Procédure

Etape 1

H_0 : pas de différence significative
H_1 : la différence est significative

Etape 2

a) Classer les données de la plus basse à la plus haute et classez chaque valeur par son rang.
b) Trouver la somme des rangs de chaque groupe

Etape 3

Calculer la valeur du test$H$

$$\boldsymbol{H = \frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \cdots + \frac{R_k^2}{n_k}\right) - 3(N+1)} \qquad \textbf{(20)}$$

$N = n_1 + n_2 + \cdots + n_k$

R_k : somme des rangs de chaque groupe k

k : le nombre de groupe

Etape 4

Utiliser le tableau du khi-deux pour trouver la valeur critique $H_{critique}$,avec $ddl = k - 1$ (k = nombre de groupes).

Si $H < H_{critique}$, ne pas rejeter H_0

Si $H \geq H_{critique}$, H_0 rejetée

Exemple

Un chercheur souhaite vérifier si le nombre total d'infections survenues dans trois groupes d'hôpitaux sélectionnés au hasard est le même.À α = 0,05, existe-t-il suffisamment de preuves pour rejeter l'affirmation selon laquelle le nombre d'infections dans les trois groupes d'hôpitaux est le même ?

Groupe 1	Groupe 2	Groupe 3
656	470	103
414	226	107
1019	74	164
277	110	152

Solution

Etape 1

H_0 : pas de différence significative
H_1 : la différence est significative

Etape 2

Les valeurs entre parenthèses sont les rangs

	Groupe 1	Groupe 2	Groupe 3
	656 (11)	470 (10)	103 (2)
	414 (9)	226 (7)	107 (3)
	1019 (12)	74 (1)	152 (5.5)
	277 (8)	110 (4)	152 (5.5)
n	$n_1 = 4$	$n_2 = 4$	$n_3 = 4$
R	$R_1 = 40$	$R_2 = 22$	$R_3 = 16$

Etape 3

$N = 4 + 4 + 4 = 12$

$k = 3$

$$H = \frac{12}{12(12+1)}\left(\frac{(40)^2}{4} + \frac{(22)^2}{4} + \frac{(16)^2}{4}\right) - 3(12+1)$$

$$H = 0.0769\,(400 + 121 + 64) - 39$$

$$H = 5.89$$

Etape 4

La valeur critique $H_{critique} = 5.991$,avec $ddl = 3 - 1 = 2$

Si $H = 5.89 < H_{critique} = 5.991$, ne pas rejeter H_0.

Corrélation de rang de Spearman (Corrélation non paramétrique)

Le coefficient de corrélation de rang est plus simple à calculer que le coefficient de Pearson (voir la partie corrélation linéaire de Pearson). Pour calculer la statistique de test ($r_{s)}$ en utilisant la différence de rang après un classement effectuer sur l'ensemble des données.

Procédure

Etape 1

H_0 : Pas de corrélation

H_1 : la corrélation est significative

Etape 2

a) Classez les valeurs dans chaque série de données
b) Calculez la différence de rang de chaque paire (d)
c) Calculez le carré des différences (d^2)
d) Trouvez la somme des carrés ($\sum d^2$)

Etape 3

Remplacez dans la formule

La formule suggère qu'il n'y a pas d'égalité dans les rangs :

$$\boldsymbol{r_s = 1 - \left[\frac{6(\sum d^2)}{n(n^2-1)}\right]} \qquad \textbf{(21)}$$

d : différence de rang

n : nombre de paires de données

Etape 4

Utilisez le tableau du Corrélation de Spearman pour trouver la valeur critique $r_{critique}$

Si$r_s < r_{critique}$, ne pas rejeter H_0. On conclut qu'il n'y a pas de corrélation entre les deux séries de données

Si $r_s \geq r_{critique}$, H_0 rejetée. On conclut qu'il y a une corrélation entre les deux séries de données

Exemple

Utilisez le coefficient de corrélation de Spearman pour voir s'il existe une relation linéaire entre ces deux séries de données.

Calories	580	580	270	470	420	415	330	430
Cholestérols (mg)	205	225	285	270	185	215	185	220

Solution

Etape 1

H_0 : Pas de corrélation
H_1 : la corrélation est significative

Etape 2

Calories	Cholestérols (mg)	Calories (Rang)	Cholestérols (mg) (Rang)	d	d^2
580	205	7	3	4	16
580	225	6	6	0	0
270	285	1	8	-7	49
470	270	6	7	-1	1
420	185	3	1.5	1.5	2.25
415	215	4	4	0	0
330	185	2	2	0	0
430	220	5	5	0	0
					$\sum = 68.25$

Etape 3

Remplacez dans la formule

$n = 8$

Il 'n y pas d'égalité dans les rangs

$$r_s = 1 - \left[\frac{6(\sum d^2)}{n(n^2-1}\right]$$

$$r_s = 1 - \left[\frac{6(68.25)}{8(8^2-1}\right]$$

$$r_s = 1 - 0.77$$

$$r_s = 0.18$$

Etape 4

La valeur critique $r_{critique} = 0.738$

Si $0.18 < 0.738$, ne pas rejeter H_0. On conclut qu'il n'y a pas de corrélation entre le nombre de calories et le cholestérol.

Bibliographie

Ab Rahman, J. (2015). Brief guidelines for methods and statistics in medical research. Springer Singapore.

Armitage, P., Berry, G., & Matthews, J. N. S. (2008). Statistical methods in medical research. John Wiley & Sons.

Bluman, A. (2014). Elementary Statistics: A step by step approach 9e. McGraw Hill.

BURT, B. (2013). Basic Biostatistics: Statistics For Public Health Practice. Jones & Bartlett Learning.

Chernick, M. R., &Friis, R. H. (2003). Introductory biostatistics for the health sciences: modern applications including bootstrap. John Wiley & Sons.

D'Agostino, R. B. (Ed.). (2005). Tutorials in Biostatistics, Tutorials in Biostatistics: Statistical Modelling of Complex Medical Data (Vol. 2). John Wiley & Sons.

Daniel, W. W., & Cross, C. L. (2018). Biostatistics: a foundation for analysis in the health sciences. Wiley.

Daniel, W. W., & Cross, C. L. (2018). Biostatistics: a foundation for analysis in the health sciences. Wiley.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). An introduction to statistical learning: with applications in R. Spinger.

Gourgou, S., Paoletti, X., &Mathoulin-Pelissier, S. (2021). Méthodes biostatistiques appliquées à la recherche clinique en cancérologie (Vol. 2). John LibbeyEurotext.

Hebel, J. R., & McCarter, R. (2011). Study guide to epidemiology and biostatistics. Jones & Bartlett Publishers.

Hoffman, J. I. (2015). Biostatistics for medical and biomedical practitioners. Academic press.

Hothorn, T., & Everitt, B. S. (2006). A handbook of statistical analyses using R. Chapman and Hall/CRC.

Kirkwood, B. R., & Sterne, J. A. (2010). Essential medical statistics. John Wiley & Sons.

Lalanne, C., &Mesbah, M. (2017). Biostatistique et analyse informatique des données de santé avec Stata (Vol. 2). ISTE Group.

Mahajan, B. K., & Lal, S. (1999). Methods in biostatistics for medical students and research workers. Indian Journal of CommunityMedicine, 24(3), 140.

Motulsky, H. J. (2019). Biostatistique: une approche intuitive. De Boeck Supérieur.

Pagano, M., Gauvreau, K., &Mattie, H. (2022). Principles of biostatistics. CRC Press.

Rumsey, D. J. (2010). Statistics essentials for dummies. John Wiley & Sons.

Townend, J. (2013). Practical statistics for environmental and biological scientists. John Wiley & Sons.

Triola, M. F., &Iossi, L. (2008). Essentials of statistics. Boston, MA, USA:: Pearson Addison Wesley.

Triola, M. M., &Triola, M. F. (2013). Biostatistics for the Biological and Health Sciences with Statdisk: Pearson New International Edition. Pearson Higher Ed.

Van Belle, G., Fisher, L. D., Heagerty, P. J., & Lumley, T. (2004). Biostatistics: a methodology for the health sciences. John Wiley & Sons.

Zar, J. H. (1999). Biostatistical analysis. Pearson Education India.

Tableau 6. Table de la loi de Fisher-Snedecor

		Risque d'erreur $\alpha = 0.05$																
		$ddl = k - 1$																
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**
$ddl = n - k$	**1**	161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	240.5	241.9	243.9	245.9	248	249.1	250.1	251.1	252.2
	2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	19.41	19.43	19.45	19.45	19.46	19.47	19.48
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.7	8.66	8.64	8.62	8.59	8.57
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	5.91	5.86	5.8	5.77	5.75	5.72	5.69
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.5	4.46	4.43
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4	3.94	3.87	3.84	3.81	3.77	3.74
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.3
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.9	2.86	2.83	2.79
	10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.7	2.66	2.62
	11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49
	12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.6	2.53	2.46	2.42	2.38	2.34	2.3
	14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6	2.53	2.46	2.39	2.35	2.31	2.27	2.22
	15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.48	2.4	2.33	2.29	2.25	2.2	2.16
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11
	17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.1	2.06
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02
	19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98
	20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.28	2.2	2.12	2.08	2.04	1.99	1.95
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.1	2.05	2.01	1.96	1.92
	22	4.3	3.44	3.05	2.82	2.66	2.55	2.46	2.4	2.34	2.3	2.23	2.15	2.07	2.03	1.98	1.94	1.89
	23	4.28	3.42	3.03	2.8	2.64	2.53	2.44	2.37	2.32	2.27	2.2	2.13	2.05	2.01	1.96	1.91	1.86
	24	4.26	3.4	3.01	2.78	2.62	2.51	2.42	2.36	2.3	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84
	25	4.24	3.39	2.99	2.76	2.6	2.49	2.4	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.9	1.85	1.8
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.2	2.13	2.06	1.97	1.93	1.88	1.84	1.79
	28	4.2	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77
	29	4.18	3.33	2.93	2.7	2.55	2.43	2.35	2.28	2.22	2.18	2.1	2.03	1.94	1.9	1.85	1.81	1.75
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2	1.92	1.84	1.79	1.74	1.69	1.64
	60	4	3.15	2.76	2.53	2.37	2.25	2.17	2.1	2.04	1.99	1.92	1.84	1.75	1.7	1.65	1.59	1.53

Table de Khi deux (χ^2)

	α									
ddl	**0.995**	**0.99**	**0.975**	**0.95**	**0.9**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.01	0.02	0.051	0.103	0.211	4.605	5.991	7.378	9.21	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.86
5	0.412	0.554	0.831	1.145	1.61	9.236	11.071	12.833	15.086	16.75
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.69	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.18	2.733	3.49	13.362	15.507	17.535	20.09	21.955
9	1.735	2.088	2.7	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.94	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.92	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.3
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.66	5.629	6.571	7.79	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.39	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.26	9.591	10.851	12.443	28.412	31.41	34.17	37.566	39.997
21	8.034	8.897	10.283	11.591	13.24	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.98	45.559
25	10.52	11.524	13.12	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.16	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.29
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.42	76.154	79.49

Table de Spearman

	α			
n	0.1	0.05	0.02	0.01
5	0.9			
6	0.829	0.886	0.943	
7	0.714	0.786	0.893	0.929
8	0.643	0.738	0.833	0.881
9	0.6	0.7	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.536	0.618	0.709	0.818
12	0.497	0.591	0.703	0.78
13	0.475	0.566	0.673	0.745
14	0.457	0.545	0.646	0.716
15	0.441	0.525	0.623	0.689
16	0.425	0.507	0.601	0.666
17	0.412	0.49	0.582	0.645
18	0.399	0.476	0.564	0.625
19	0.388	0.462	0.549	0.608
20	0.377	0.45	0.534	0.591
21	0.368	0.438	0.521	0.576
22	0.359	0.428	0.508	0.562
23	0.351	0.418	0.496	0.549
24	0.343	0.409	0.485	0.537
25	0.336	0.4	0.475	0.526
26	0.329	0.392	0.465	0.515
27	0.323	0.385	0.456	0.505
28	0.317	0.377	0.488	0.496
29	0.311	0.37	0.44	0.487
30	0.305	0.364	0.432	0.478

Table de la loi de Student

d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291