République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA
FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE
DÉPARTEMENT DE MATHÉMATIQUE

Polycopié du Cours

BIOSTATISTIQUES

Statistiques Appliquées à l'Expérimentation En Sciences Biologique

Préparé par :
Dr. CHERFAOUI Mouloud

Université de Biskra, 2019/2020

Table des matières

Chapitre 1

Introduction à la théorie de test d'hypothèses ”

Introduction

Les tests statistiques sont des méthodes de la statistique inférentielle qui, comme l'estimation, permettent d'analyser des données obtenues par tirages au hasard. Ils consistent à généraliser les propriétés constatées sur des observations à la population d'où ces dernières sont extraites, et à répondre à des questions concernant par exemple la nature d'une loi de probabilité, la valeur d'un paramètre ou l'indépendance de deux variables aléatoires.

1.1 Test d'indépendance : Test de $Khi - Deux$

La mise en évidence de l'existence d'une liaison entre deux caractères aléatoires a beaucoup d'importance dans toutes les études biologique. Les techniques employées sont différentes suivant que les variables étudiées sont discrètes ou continues ; elles sont différentes aussi suivant que le type de loi des variables est connu ou non. Nous distinguerons trois cas fondamentaux qui donnent lieu chacun à diverses méthodes : les variables sont toutes les deux discrètes, une seule est continue et les deux le sont. Partant de là, nous allons introduire une méthode, plus générale, qui peut être appliqué dans les trois situations en question qui nous permet de mettre en évidence l'indépendance entre deux caractères (facteurs) aléatoires.

1.1.1 Position du problème

On veut savoir si le temps écoulé depuis la vaccination contre la petite vérole a ou non une influence sur le degré de gravité de la maladie lorsqu'elle apparaît. Les patients sont divisés en trois catégories selon la gravité de leur maladie : légère (L), moyenne (M), ou grave (G) et en trois autres quant à la durée écoulée depuis la vaccination : moins de 10 ans (A), entre 10 et 25 ans (B), plus de 25 ans (C).

Les résultats d'une observation portant sur $n = 1574$ malades sont les suivants :

Degré de gravité Y de la maladie	Durée X écoulée depuis la vaccination			Total
	A	B	C	
G	1	42	230	273
M	6	114	347	467
L	23	301	510	834
Total	30	457	1087	1574

Pour mettre en évidence, l'existence d'une liaison entre la durée écoulée depuis la vaccination et le degré de gravité de la maladie, on choisit de tester les hypothèses nulle et alternative :

H_0 **:** la durée écoulée depuis la vaccination et le degré de gravité de la maladie sont indépendantes",

H_1 **:** la durée écoulée depuis la vaccination et le degré de gravité de la maladie sont liées".

C'est dans ce genre de situations, que le test d'indépendance de $Khi - Deux$ peut intervenir.

1.1.2 Principe du test

De manière générale, le problème du test d'indépendance est posé de la manière suivante : soient X et Y deux variables discrètes (respectivement continues), X à r modalités (respectivement classes) et Y à k modalités (respectivement classes), notées respectivement $i = 1, ..., r$ et $j = 1, ..., k$ et n_{ij} l'effectif observé, dans le tableau croisé, des individus pour lesquels X vaut A_i et Y vaut B_j (voir le tableau 1.2).

X \ Y	B_1	B_2	$\cdots$	B_j	$\cdots$	B_k	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2k}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rj}	$\cdots$	n_{rk}	$n_{r\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet k}$	n

TABLE 1.2: *Présentation des effectifs observés sous forme tableau croisé*

On note $n_{\bullet j}$ le nombre total de ceux pour lesquels Y vaut B_j et qui figure au bas de la $j^{\text{ème}}$ colonne, et $n_{i\bullet}$ le nombre total de ceux pour lesquels X vaut A_i et qui figure à droite de la ligne i.

Pour mettre en évidence ou nie une liaison entre X et Y, on choisit de tester les hypothèses nulle et alternative :

H_0 **:** "X et Y sont indépendantes",

H_1 **:** "X et Y sont liées".

Sous l'hypothèse H_0 d'indépendance de X et Y :

$$P(X = i, Y = j) = P(X = i) \times P(Y = j) \tag{1.1}$$

$$p_{ij} = p_{i\bullet} \times p_{\bullet j}. \tag{1.2}$$

Comme les estimateurs de chacune de ces probabilités à partir du tableau des effectifs du tableau des observations, sont

$$\hat{p}_{ij} = \frac{n_{ij}}{n}, \quad \hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n} \quad \text{et} \quad \hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}. \tag{1.3}$$

Alors, d'après les formules (1.2) et (1.3) on aura :

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n} \times \frac{n_{\bullet j}}{n} \Longrightarrow n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}.$$

Si H_0 est vraie alors les écarts

$$e_{ij} = \hat{p}_{ij} - \hat{p}_{i\bullet} \times \hat{p}_{\bullet j} \left(\text{ ou encore } E_{ij} = n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n} \right),$$

ne doivent être dus qu'aux fluctuations d'échantillonnage.

Afin de répondre à notre objectif nous allons définir la statistique des erreurs suivante :

$$K_n^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} \times n_{\bullet j}}{n}}.$$

On peut démontrer que la variable K_n^2 suit une loi proche de celle d'un χ^2 à $(r-1)(k-1)$ degrés de liberté, pourvu que les dénominateurs $n_{i\bullet}p_{\bullet j}$ soient tous supérieurs à 5 (si ce n'est pas le cas, on regroupe plusieurs classes de tel sorte que cette condition soit vérifié), alors la règle de décision du test d'indépendance de $Khi - Deux$ sera sous la forme suivante :

- Si $k_n^2 < \chi^2_{((r-1)(k-1),\alpha)}$, alors les deux variables en question sont indépendantes.
- Si $k_n^2 > \chi^2_{((r-1)(k-1),\alpha)}$, alors les deux variables en question sont liées.

où k_n^2 est la réalisation de la statistique K_n^2, et $\chi^2_{((r-1)(k-1),\alpha)}$ le fractile d'ordre $1-\alpha$ d'une loi de $Khi - Deux$ à $(r-1)(k-1)$ ddl

1.1.3 Exemple d'application

Reprenant l'exemple exposé dans la section 1.1.1. Sous l'hypothèse H_0 (les deux variables sont indépendantes) les effectifs n_{ij} devrai être comme suit :

$n_{ij} =$

Y	X		
	A	B	C
G	5.20	79.26	188.53
M	8.90	135.59	322.51
L	15.90	242.15	575.96

Ainsi, les écarts E_{ij} entre les effectifs théoriques et observées sont :

$E_{ij} =$

Y	X		
	A	B	C
G	-4.20	-37.26	41.47
M	-2.90	-21.59	24.49
L	7.10	58.86	-65.96

$\Rightarrow \frac{E_{ij}^2}{\frac{n_{i\bullet} \times n_{\bullet j}}{n}} =$

Y	X		
	A	B	C
G	3.3981	17.5193	9.1206
M	0.9461	3.4379	1.8598
L	3.1742	14.3043	7.5534

On a d'une part, d'après les résultats ci-dessous, la réalisation de la variable K_n^2 vaut $k_n^2 = 61.3137$. D'autre part, sous l'hypothèse H_0, K_n^2 suit une loi du χ^2 à (r -1) (k - 1) = 4 degrés de liberté, ainsi la valeur critique du test pour un risque $\alpha = 5\%$ vaut 11.143 (voir table de la loi de $Khi - Deux$ en annexe).

On constate que la valeur critique du test est inférieur à la valeur observée k_n^2, alors on rejette l'hypothèse d'indépendance de la gravité de la maladie et du délai écoulé depuis la vaccination.

1.2 Analyse de la variance à un facteur (ANOVA 1)

Dans cette section, nous allons intéressé à un cas plus générale pour la comparaison de moyennes et cela lorsque le nombre d'échantillon est supérieur strictement à deux. Plus précisément nous allons intéressé à la technique d'analyse de la variance à un seul facteur qui est la plus adéquate avec la situation.

1.2.1 Position du problème

Supposons que nous ayons 3 forêts contenant un type d'arbre bien déterminé où nous désirons savoir si ces forêts ont une influence sur la hauteur des arbres ou non. À cet effet, nous avons réalisés un recueil de hauteur de six (06) arbres dans chaque forêt, dont les mesures sont rangées dans le tableau suivant.

N°	forêt 1	forêt 2	forêt 3
1	23.3	18.9	22.5
2	24.4	21.1	22.9
3	24.6	21.1	23.7
4	24.9	22.1	24.0
5	25.0	22.5	24.0
6	26.2	23.5	24.5

TABLE 1.3: *Tailles des arbres selon la forêt*

Soit les notions et les notations suivantes :

- Les forêts : Variable qualitative contenant trois modalités, appelée facteur.
- Hauteur des arbres : Réponse, notée X, et μ_i la hauteur moyenne des arbres de la $i^{\text{ème}}$ forêt $(i = \overline{1,3})$.

Répondre à notre objectif consiste à la réalisation du test suivant :

$$H_0 : \ ''\mu_1 = \mu_2 = \mu_3 = \mu'' \text{ contre } H_1 : \ ''\exists\ i, j \in \{1,2,3\} \text{ tel que } \mu_i \neq \mu_j;'' \, .$$

Pour réaliser ce test nous pourrons le décomposer en trois sous-tests où nous comparons la hauteur moyenne des arbres deux à deux selon les forêts. Mais afin de contourner le problème d'erreur α gonflé, le fait elle ne réalise qu'une seule comparaison à la fois, nous utilisons la technique statistique connue sous le nom d'analyse de variance (en anglais : Analyse Of Variance (ANOVA)) plutôt que des tests de Student t (voir section **??**) multiples. Remarquez que l'ANOVA peut aussi être utilisée quand $p = 2$ puisque, elle retourne la même conclusion qu'un test t.

1.2.2 Analyse de la variance à un seul facteur

L'identification de l'*ANOVA* 1 au sens littéraire peut être résumée dans la définition suivante :

Définition 1.1 *(ANOVA 1)*
L'analyse de la variance à un facteur teste l'effet d'un facteur contrôlé A ayant p modalités (groupes) sur les moyennes d'une variable quantitative X.

Les problèmes concernés par la technique *ANOVA* 1 s'écrivent en générale de la manière suivante :

N	*groupe* 1	*groupe* 2		*groupe p*
1	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,p}$
2	$X_{2,1}$	$X_{2,2}$	$\cdots$	$X_{2,p}$
3	$X_{3,1}$	$X_{3,2}$	$\cdots$	$X_{3,p}$
4	$X_{4,1}$	$X_{4,2}$	$\cdots$	$X_{4,p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_j	$X_{n_1,1}$	$X_{n_2,2}$	$\cdots$	$X_{n_p,p}$

et le modèle mathématique leurs associés est donné par :

$$X_{ij} = \mu_i + \epsilon_{ij}, \text{ avec } i = \overline{1,n}, j = \overline{1,p} \text{ et } \epsilon_{ij} \rightsquigarrow N(0,\sigma^2), \tag{1.4}$$

où X_{ij} est la $j^{\text{ième}}$ réalisation de la variable quantitative X dans le $i^{\text{ième}}$ échantillon et ϵ_{ij} sont les erreurs de mesure.

Si on retient ce modèle alors le test à réaliser est défini par :

$$H_0 : \text{''}\mu_1 = \mu_2 = ... = \mu_p = \mu\text{''} \text{ contre } H_1 : \text{''}\exists\, i,j \in \{1,2,...,p\} \text{ tel que } \mu_i \neq \mu_j\text{''}. \tag{1.5}$$

Dans ce qui suit, nous allons énumérer les étapes de la mise en oeuvre de l'ANOVA 1 qui nous permet de réaliser ce test.

1.2.3 Les étapes de l'$ANOVA$ 1

Afin de réaliser le test définie dans (1.5), trois conditions doit être vérifiées préalablement, à savoir :

- Les p échantillons comparés sont indépendants.
- La variable quantitative étudiée suit une loi normale dans les p populations comparées.
- Les p populations comparées ont même variance : *Homogénéité* des variances ou *homoscédasticité*.

Si ces dernières conditions sont vérifiées alors, on peut utiliser la technique $ANOVA$ 1 pour réaliser le test (1.5), et pour ce faire nous avons besoin des quantités (statistiques) suivantes :

- La moyenne de toutes les observations : $\overline{X} = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j} X_{ij}$ avec $n = \sum_{j=1}^{p} n_j$;
- Moyenne de chaque échantillon : $\overline{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$, pour $j = \overline{1,p}$;
- Variance de chaque échantillon : $\hat{\sigma}_i^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2$, pour $j = \overline{1,p}$;
- La variance de toutes les observations : $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2$ avec $n = \sum_{j=1}^{p} n_j$.

On peut démonter facilement que la variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances des p échantillons, c'est-à-dire :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2 = \frac{1}{p}\sum_{j=1}^{p}\sigma_i^2 + \frac{1}{p}\sum_{j=1}^{p}(\overline{X}_j - \overline{X})^2, \tag{1.6}$$

ou encore :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2 + \frac{1}{p}\sum_{j=1}^{p}(\overline{X}_j - \overline{X})^2. \tag{1.7}$$

On multipliant (1.7), par n on obtient :

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2}_{SC_{Tot}} = \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2}_{SC_{Res}} + \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(\overline{X}_j - \overline{X})^2}_{SC_{Fac}}, \tag{1.8}$$

où,

SC_{Tot} : est la variation totale qui représente la dispersion des données autour de la moyenne générale.

SC_{Fac} : est la variation due au facteur (variation inter-groupes) qui représente la dispersion des moyennes autour de la moyenne générale.

SC_{Res} : est la variation résiduelle (variation intra-groupes) qui représente la dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

L'idée la plus naturelle est que le facteur n'a pas d'impact sur le caractère étudié si la variation totale n'est engendrée que par la variation intra-groupes (résiduelle) associée au caractère, c'est-à-dire,

- Si H_0 est vraie, alors la variation SC_{Fac} due au facteur doit être petite par rapport à la variation résiduelle SC_{Res}.
- Par contre, si H_1 est vraie alors la variation SC_{Fac} due au facteur doit être grande par rapport à la quantité SC_{Res}.

Pour comparer ces quantités, Fisher a considéré le rapport des carrés moyens associés au facteur CM_{Fac} et les carrés moyens résiduels CM_{Res}, où

le carré moyen associé au facteur est : $CM_{Fac} = \frac{SC_{Fac}}{p-1}$.

le carré moyen résiduel est : $CM_{Res} = \frac{SC_{Res}}{n-p}$.

Si les 3 conditions d'application d'ANOVA (Indépendance, Normalité et Homogénéité) sont vérifiées et H_0 est vraie, alors

$$F_{obs} = \frac{CM_{Fac}}{SC_{Res}} \rightsquigarrow f_{(p-1,n-p)}.$$

Décision : Pour un seuil de risque donné α les tables de Fisher nous fournissent une valeur critique f_α telle que :

$$P\left(\frac{CM_{Fac}}{SC_{Res}} < f_\alpha\right) = 1 - \alpha,$$

- si $f_{obs} < f_\alpha \implies$ on ne peut pas rejeter H_0 (le facteur n'a aucune influence sur le caractère étudier),
- si $f_{obs} \geq f_\alpha \implies$ on rejette H_0 (le facteur influe sur le caractère étudié),

avec f_{obs} est la réalisation de la variable (statistique) F_{obs}.

Les résultats d'une ANOVA 1 sont souvent présentés dans un tableau sous la forme suivante :

	Somme des carrés	Degrés de libertés	Carré moyen	ratio	Ficher
source de variation	SC	ddl	CM	F_{obs}	c
Inter-groupe (Fac)	SC_{Fac}	$p-1$	CM_{Fac}	$\frac{CM_{Fac}}{CM_{Res}}$	c
Intra-groupe (Rés)	SC_{Res}	$n-p$	CM_{Res}		
Total	SC_{Tot}	$n-1$			

1.2.4 Exemple d'application

Reprenant l'exemple présenter dans la section 1.1.1. Les étapes qu'on doit suivre pour réaliser le test

$$H_0 : \text{ " } \mu_1 = \mu_2 = \mu_3 = \mu \text{ " contre } H_1 : \text{ " } \exists\, i, j \in \{1,2,3\} \text{ tel que } \mu_i \neq \mu_j \text{ ",}$$

à l'aide de la technique ANOVA 1, sont les suivantes :

- Calculer les moyennes des différents échantillons : $\overline{X}_1 = 24.73$, $\overline{X}_2 = 21.53$ et $\overline{X}_3 = 23.60$.

- Calculer la moyenne globale de toutes les observations : $\overline{X} = \frac{1}{n}(n_1\overline{X}_1 + n_2\overline{X}_2 + n_3\overline{X}_3) = 23.2889$.
- Compléter le tableau de l'ANOVA à un seul facteur :

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio F_{obs}	Ficher c
Inter-groupe	31.5911	2	15.7956	12.02	3.6823
Intra-groupe	19.7067	15	1.3138		
Total	51.2978	17			

- Décision : on constate que $f_{obs} = 12.02 > f_\alpha = 3.6823$ (pour un risque de $\alpha = 5\%$), donc les hauteurs moyennes des arbres sont significativement différentes d'une forêt à une autre. Cela signifié que le facteur forêt influe sur la hauteur des arbres.

1.3 Analyse de la variance à deux facteurs (ANOVA 2)

Cette section est consacrée à l'étude des situations expérimentales dans lesquelles l'effet de deux facteurs (variables qualitatives) est étudié simultanément, c'est-à-dire dans le même protocole expérimental. En cela, elle constitue une extension à la situation précédente dans laquelle on n'étudiait qu'un seul facteur à la fois (ANOVA d'ordre 1).

1.3.1 Position du problème

Nous avons réalisés un recueil de rendement de trois variétés du blé selon le type d'engrais utilisé, les mesures obtenus sont rangées dans la table 1.4.

	Variété 1	Variété 2	Variété 3
Engrais 1	46	41	35
	35	26	21
	19	11	31
Engrais 2	37	49	45
	18	37	66
	18	35	61
Engrais 3	32	65	34
	43	67	66
	32	58	58

TABLE 1.4: *Variation de rendement du blé selon la variété et le type d'engrais*

Si l'on s'intéresse, à l'effet des trois type d'engrais et les trois variétés du blé sur les rendements séparément, on pourrait réaliser deux expériences dans lesquelles on manipulerait chacun des deux facteurs, et analyser les résultats à l'aide d'ANOVA 1 s'il y a plus de 2 modalités ou niveaux pour chaque facteur : on saurait ainsi si le type d'engrais affecte sensiblement les rendements mesurées, et également si la variété affecte les rendements. Mais, on ne saurait pas si l'effet du type d'engrais est le même quelque soit la variété du blé ; en d'autres termes, on perd l'information concernant l'interaction entre ces deux facteurs.

Les modèles d'ANOVA d'ordre 2 sont comparables sur le fond au modèle précédent de l'ANOVA à un seul facteur mais elle incluent dans certaines situations, en plus de l'étude des effets principaux des deux facteurs, celle du l'effet d'interaction des deux facteurs.

L'identification de l'ANOVA à deux facteurs (ANOVA 2) au sens littéraire peut être résumée dans la définition suivante :

Définition 1.2 *(ANOVA 2)*
L'analyse de la variance à deux facteurs teste l'effet de deux facteurs contrôlés A et B (variables qualitatives) ayant respectivement I et J modalités sur les moyennes d'une variable quantitative X.

1.3.2 ANOVA 2 avec répétitions : cas de plan équilibré

Les problèmes concernés par la technique ANOVA 2 se présente en générale de la manière suivante :

N	B_1	B_2	$\cdots$	B_J
A_1	$X_{1,1,1}$	$X_{1,2,1}$	$\cdots$	$X_{1,J,1}$
	$X_{1,1,2}$	$X_{1,2,2}$	$\cdots$	$X_{1,J,2}$
	$\vdots$			$\vdots$
	$X_{1,1,K}$	$X_{1,2,K}$	$\cdots$	$X_{1,J,K}$
A_2	$X_{2,1,1}$	$X_{2,2,1}$	$\cdots$	$X_{2,J,1}$
	$X_{2,1,2}$	$X_{2,2,2}$	$\cdots$	$X_{2,J,2}$
	$\vdots$			$\vdots$
	$X_{2,1,K}$	$X_{2,2,K}$	$\cdots$	$X_{2,J,K}$
$\vdots$		$\vdots$		
A_I	$X_{I,1,1}$	$X_{I,2,1}$	$\cdots$	$X_{I,J,1}$
	$X_{I,1,2}$	$X_{I,2,2}$	$\cdots$	$X_{I,J,2}$
	$\vdots$			$\vdots$
	$X_{I,1,K}$	$X_{I,2,K}$	$\cdots$	$X_{I,J,K}$

et sont modèle mathématique est donné par :

$$X_{ijk} = \mu_{ij} + \epsilon_{ijk}, \text{ avec } i = \overline{1,I}, j = \overline{1,J} \text{ et } k = \overline{1,K}, \tag{1.9}$$

où X_{ijk} est la $k^{\text{ième}}$ réalisation de la variable quantitative X, lorsque on fixe le premier facteur à la $i^{\text{ième}}$ modalité et le deuxième facteur à la $j^{\text{ième}}$ modalité et ϵ_{ijk} sont les erreurs de mesure (inconnues) de plus $E(\epsilon_{ijk}) = 0$.

Le modèle (1.9) peut être réécrit sous sa forme détaillée comme suit :

$$X_{ijk} = \mu + a_i + b_j + c_{ij} + \epsilon_{ijk}, \text{ avec } i = \overline{1,I}, j = \overline{1,J} \text{ et } k = \overline{1,K}, \tag{1.10}$$

ce qui s'explique que la réalisation de la variable X est un cumule d'une constate μ (indépendante des deux facteurs), de l'effet du premier facteur a, de l'effet du deuxième facteur b, l'effet d'interaction des deux facteurs c et de l'erreur de mesure ϵ.

Si le modèle de référence retenu est le modèle (1.9), alors le test pour lequel nous nous intéressons à réaliser sera formulé comme suit :

$$\begin{aligned} &H_0 :'' \forall \left\{ \begin{array}{l} i \in \{1,2,\cdots,I\} \\ j \in \{1,2,\cdots,J\} \end{array} \right\}, \ \mu_{ij} = \mu'' \\ &\text{contre} \\ &H_1 :'' \exists \left\{ \begin{array}{l} i_1, i_2 \in \{1,2,\cdots,I\} \\ j_1, j_2 \in \{1,2,\cdots,J\} \end{array} \right\} \text{ tel que } \mu_{i_1 j_1} \neq \mu_{i_2 j_2}.'' \end{aligned} \tag{1.11}$$

Par contre, si le modèle de référence retenu est le modèle (1), alors l'analyse de la variance à deux facteurs avec repetitions consiste en réalisation de trois tests de Fisher à la fois, dont la formulation est :

1. **Effet du premier facteur** :
   H_0 : "*les paramètres a_i sont tous nuls*" contre H_1 : "*les paramètres a_i ne sont pas tous nuls*"
2. **Effet du second facteur** :
   H_0 : "*les paramètres b_j sont tous nuls* " contre H_1 : "*les paramètres b_j ne sont pas tous nuls*"
3. **Effet de l'interaction des deux facteurs** :
   H_0 : "*les paramètres c_{ij} sont tous nuls* " contre H_1 : "*les paramètres c_{ij} ne sont pas tous nuls*"

1.3.3 Les étapes de l'ANOVA 2

La lise en oeuvre d'une ANOVA 2, se fait principalement en 4 étapes. Les détails de ces étape sont comme suit :

Étape 0 : (Conditions)
Afin de réaliser une analyse de la variance à deux facteurs, les conditions suivantes doivent être vérifiées préalablement :

- Les $I * J$ échantillons comparés sont mutuellement indépendants.
- La variable quantitative étudiée suit une loi normale dans les $I * J$ populations comparées.
- Les $I * J$ populations comparées ont même variance : *Homogénéité* des variances (*homoscédasticité*).

Étape 1 : (Moyennes et variations quadratiques)
Quantifier les différentes statistiques intervenant dans l'ANOVA à 2 facteurs et qui sont :

- La moyenne globale de toutes les observations :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk} \text{ avec } n = I * J * K;$$

- Moyenne de chaque échantillon :

$$\overline{X}_{ij\bullet} = \frac{1}{K}\sum_{k=1}^{K} X_{ijk} \text{ pour } i = \overline{1, I} \text{ et } j = \overline{1, J};$$

- Moyenne de chaque modalité du premier facteur :

$$\overline{X}_{i\bullet\bullet} = \frac{1}{J * K}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk} \text{ pour } i = \overline{1, I};$$

- Moyenne de chaque modalité du deuxième facteur :

$$\overline{X}_{\bullet j\bullet} = \frac{1}{I * K}\sum_{i=1}^{I}\sum_{k=1}^{K} X_{ijk} \text{ pour } j = \overline{1, J};$$

- La somme des carrés des erreurs totale :

$$SC_{tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \overline{X})^2;$$

- La somme des carrés des erreurs résiduelles :

$$SC_{res} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \overline{X}_{ij\bullet})^2;$$

- La somme des carrés des erreurs du premier facteur :

$$SC_a = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{X}_{i\bullet\bullet} - \overline{X})^2;$$

- La somme des carrés des erreurs du deuxième facteur :

$$SC_b = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{X}_{\bullet j\bullet} - \overline{X})^2;$$

- La somme des carrés des erreurs des deux facteurs facteur :

$$SC_c = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{X}_{ij\bullet} - \overline{X}_{i\bullet\bullet} - \overline{X}_{\bullet j\bullet} + \overline{X})^2.$$

Avec le même raisonnement que dans l'ANOVA 1, on peut démontrer que la variation quadratique totale des observations autour de la moyenne $\overline{X}$ peut être décomposé comme suit :

$$SC_{Tot} = SC_{Res} + SC_a + SC_b + SC_c. \tag{1.12}$$

Étape 2 : (Les Carrés moyens)
A partir de la décomposition (1.12), l'idée la plus naturelle est que le facteur ou l'interaction des facteurs n'a pas d'impact sur le caractère étudié si la variation inter-groupes (engendrée par les deux facteur ou/et leurs interaction) associée au caractère est négligeable par rapport aux fluctuations individuelles. Pour comparer ces quantités, on considère les carrés moyens suivants :

Carré moyen due aux fluctuations individuelles : $CM_{Res} = \frac{SC_{Res}}{I*J*(K-1)}$.

Carré moyen de mesure de l'effet du premier facteur :$CM_a = \frac{SC_a}{(I-1)}$.

Carré moyen de mesure de l'effet du second facteur : $CM_b = \frac{SC_b}{(J-1)}$.

Carré moyen de mesure de l'effet de l'interaction entre les deux facteurs : $CM_c = \frac{SC_c}{(I-1)*(J-1)}$.

Notons que, si les trois conditions citées précédemment (Indépendance, Normalité et Homogénéité) sont vérifiées alors sous l'hypothèse H_0

$$\frac{CM_a}{CM_{Res}} \rightsquigarrow F_{((I-1),I*J(K-1))}, \tag{1.13}$$

$$\frac{CM_b}{CM_{Res}} \rightsquigarrow F_{((J-1),I*J(K-1))}, \tag{1.14}$$

$$\frac{CM_c}{CM_{Res}} \rightsquigarrow F_{((I-1)*(J-1),I*J(K-1)),} \tag{1.15}$$

où F la loi de Fisher.

Étape 3 : (Décision)
Pour un seuil de risque donné α, nous quantifions les valeurs critiques c_α, c_β et c_γ(par la lecture sur la table de Fisher), telle que :

$$P\left(\frac{CM_a}{SC_{Res}} < f_a\right) = 1-\alpha,$$
$$P\left(\frac{CM_b}{SC_{Res}} < f_b\right) = 1-\alpha,$$
$$P\left(\frac{CM_c}{SC_{Res}} < f_c\right) = 1-\alpha,$$

Ainsi les décisions des test se font comme suit :

Décision sur le premier facteur :

- Si $\frac{CM_a}{SC_{Res}} < f_a$, alors le premier facteur n'a pas une influence significative sur le caractère étudié.
- Si $\frac{CM_a}{SC_{Res}} \geq f_a$, alors le premier facteur a une influence significative sur le caractère étudié.

Décision sur le deuxième facteur :

- Si $\frac{CM_b}{SC_{Res}} < f_b$, alors le deuxième facteur n'a pas une influence significative sur le caractère étudié.
- Si $\frac{CM_b}{SC_{Res}} \geq f_b$, alors le deuxième facteur a une influence significative sur le caractère étudié.

Décision sur l'interaction des deux facteurs :

- Si $\frac{CM_c}{SC_{Res}} < f_c$, alors l'interaction des deux facteurs n'a pas une influence significative sur le caractère étudié.
- Si $\frac{CM_c}{SC_{Res}} \geq f_c$, alors l'interaction des deux facteurs a une influence significative sur le caractère étudié.

Remarque 1.1 *Les résultats d'une ANOVA 2 sont souvent présentés dans un tableau de la forme suivante :*

source de variation	*Somme des carrés* SC	*Degrés de libertés* ddl	*Carré moyen* CM	*ratio* F_{obs}	*Ficher* F_c
due à F_A	SC_a	$I-1$	CM_a	CM_a/CM_{Res}	f_a
due à F_B	SC_b	$J-1$	CM_b	CM_b/CM_{Res}	f_b
due à $F_A \times F_B$	SC_c	$(I-1)*(J-1)$	CM_c	CM_c/CM_{Res}	f_c
Résiduelle	SC_{Res}	$I*J*(K-1)$	CM_{Res}		
Totale	SC_{Tot}	$n-1$			

TABLE *1.5: Tableau de l'ANOVA d'ordre 2*

1.3.4 Exemple d'application

Afin de concrétiser les différentes étapes citées auparavant, reprenant l'exemple présenté dans la section 1.3.1. Supposons que les trois conditions d'application de la technique d'ANOVA 2 sont vérifiées et on désir prendre nos décisions pour un risque $\alpha = 5\%$ de se tromper, alors le reste des étapes se fait comme suit :

- Calculer les moyennes des différents échantillons :

	Variété 1	Variété 2	Variété 3	$\overline{X}_{i\bullet\bullet}$
Engrais 1	33.3333	26.0000	29.0000	29.4444
Engrais 2	24.3333	40.3333	57.3333	40.6667

Engrais 3	35.6667	63.3333	52.6667	50.5556
$\overline{X}_{\bullet j\bullet}$	31.1111	43.2222	46.3333	40.2222

TABLE 1.6: *Moyennes des différents échantillons*

- Compléter la table 1.5 :

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio f_{obs}	Ficher f_α
due à F_a	1164.222	2	582.111	4.766	$f_a = 3.55$
due à F_b	2008.222	2	1004.111	8.220	$f_b = 3.55$
due à $F_a \times F_b$	1719.556	4	429.889	3.519	$f_c = 2.93$
Résiduelle	2198.667	18	122.148		
Totale	50772.000	27			

- Décision : Dans les trois situations on constate que $F_{obs} > f_\alpha$ ($4.766 > 3.55$, $8.220 > 3.55$ et $3.519 > 2.93$), cela signifié qu'on doit rejeter l'hypothèse H_0 dans les trois tests. L'interprétation de ces résultats vis-à-vis le problème étudié est que le facteur variété et le facteur type d'engrais ainsi que l'interaction entre eux influent significativement sur le rendement du blé.
  Rappelons que cette décision est prise pour un risque de 5% mais si on souhaite diminuer ce risque à 1% alors la décision sera différente (pour cet exemple seulement).
  En effet, pour un seuil de risque 2%, on constate que le type d'engrais qui influe significativement sur le rendement le fait que $8.220 > f_b = 6.01$. Par contre la variété (respectivement l'interaction des deux facteurs) n'a pas une influence significative sur le rendement car $4.766 < f_a = 6.01$ (respectivement $3.519 < f_c = 4.58$).

1.3.5 ANOVA 2 sans répétitions

Considérons l'exercice suivant :

Exercice 1 Supposons que lors d'une étude statistique d'un certain phénomène, nous nous somme intéressés à l'influence de deux facteurs F_1 (ayant *4 modalités*) et F_2 (ayant *5 modalités*) sur un caractère quantitatif X. Pour cela nous avons utilisé l'ANOVA 2 dont certains résultats, fournis par cette méthode, sont donnés comme suit :
$SC_{tot} = 1350, CM_{F_1} = 140$ et le ratio du premier facteur $f_c = 3.5$.

Si l'expérience réaliser pour répondre à notre objectif est *sans répétitions* alors :

1. Donner et compléter la table d'ANOVA correspondante au problème.
2. Que peut-on conclure sur l'effet des facteurs sur la variable X, à un seuil de risque $\alpha = 5\%$?

Avant de répondre à l'exercice introduisons quelques informations utiles pour la réalisation d'une ANOVA à deux facteurs lorsque l'expérience n'est réalisée qu'une seule fois c'est-à-dire y a une seule observation par combinaison des deux facteurs.

Rappelons que la Table de décomposition de la variance en fonction des sources de variabilité, pour un plan factoriel à 2 facteurs avec réplications ce présente comme suit :

Variation	SC	ddl	CM	F_{obs}
Facteur A	$KJ \sum_{i=1}^{I} \left(\overline{X}_{i\bullet\bullet} - \overline{X}\right)^2$	$I-1$	$CM_\alpha = SC_\alpha/dll$	CM_α/CM_{Res}
Facteur B	$KI \sum_{j=1}^{J} \left(\overline{X}_{\bullet j\bullet} - \overline{X}\right)^2$	$J-1$	$CM_\beta = SC_\beta/dll$	CM_β/CM_{Res}

Inter. $A \times B$	$K \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\overline{X}_{ij\bullet} - \overline{X}_{i\bullet\bullet} - \overline{X}_{\bullet j\bullet} + \overline{X}\right)^2$	$(I-1)*(J-1)$	$CM_\gamma = SC_\gamma/dll$	CM_γ/CM_{Res}
Résiduelle	$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(X_{ijk} - \overline{X}_{ij\bullet}\right)^2$	$I*J*(K-1)$	$CM_{Res} = SC_{Res}/dll$	
Totale	$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(X_{ijk} - \overline{X}\right)^2$	$n-1$		

TABLE 1.8: *Table de l'ANOVA d'ordre 2 cas : avec réplications équilibrées.*

avec I et J représentent le nombre de modalités ou niveaux des facteurs A et B, K l'effectif pour chaque combinaison des facteurs, où le nombre de réplications pour un plan est équilibrée, SC représente la somme des carrés, ddl le nombre de degrés de libertés et CM le carré moyen.

Cependant, lorsque le protocole expérimental ne comporte pas de réplications, c'est-à-dire qu'on ne dispose que d'une observation par combinaison de facteurs, il est toujours possible d'effectuer une ANOVA, mais l'analyse est limitée à la seule étude des effets principaux, avec comme hypothèse implicite supplémentaire l'absence d'interaction entre les facteurs. En effet, puisqu'il y n'y a qu'une seule observation pour chaque combinaison de chaque niveau des différents facteurs, il n'est plus possible d'estimer la variabilité intra pour cette combinaison particulière et l'on ne peut plus estimer la variabilité résiduelle à partir de ces intra-variabilités. Celle-ci doit donc être estimée à partir du CM de l'interaction (en fait, les composantes d'interaction et résiduelles sont confondues).

La décomposition de la variance suit le même principe que dans le cas des plans avec réplications (exception faite de l'indice de réplications) et est illustrée dans la table 1.9.

Variation	SC	ddl	CM	F_{obs}
Facteur A	$J \sum_{i=1}^{I}\left(\overline{X}_{i\bullet} - \overline{X}\right)^2$	$I-1$	$CM_A = SC_A/dll$	CM_A/CM_{Res}
Facteur B	$I \sum_{j=1}^{J}\left(\overline{X}_{\bullet j} - \overline{X}\right)^2$	$J-1$	$CM_B = SC_B/dll$	CM_B/CM_{Res}
Résiduelle	$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(X_{ij} - \overline{X}_{i\bullet} - \overline{X}_{\bullet j} + \overline{X}\right)^2$	$(I-1)*(J-1)$	$CM_{Res} = SC_{Res}/dll$	
Totale	$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(X_{ij} - \overline{X}\right)^2$	$n-1$		

TABLE 1.9: *Table de l'ANOVA d'ordre 2 cas sans réplications*

La table 1.9 représente (résume) la décomposition de la variance en fonction des sources de variabilité, pour un plan factoriel à 2 facteurs sans réplications (SC représente la somme des carrés, ddl le nombre de degrés de libertés et CM le carré moyen.

1. *La table de décomposition de la variance en fonction des sources de variabilité, pour un plan factoriel à 2 facteurs sans réplications correspondante au problème est donnée par 1.8. À partir de l'énoncé on a $I = 4$ et $J = 5$. De plus, le fait qu'il n'y a pas de réplications (sans répétions) alors $n = I * J = 20$, ces données nous permet de déterminer facilement les différents degrés de liberté (d.d.l de la troisième colonne de la table 1.9).*

   (a) $I - 1 =$ ***3***.

   (b) $J - 1 =$ ***4***.

   (c) $(I-1)*(J-1) =$ ***12***.

   (d) $n - 1 =$ ***19***.

   Pour le reste des quantités on a :

- $CM_{F_1} = SC_{F_1}/(I-1) = 140$, *alors* $SC_{F_1} = CM_{F_1} * (I-1) = 140 * 3 =$ ***420***.
- $F_c = CM_{F_1}/CM_{Res} = 3.5$, *alors* $CM_{Res} = CM_{F_1}/F_c = 140/3.5 =$ ***40***.
- $CM_{Res} = SC_{Res}/((I-1)(J-1))$, *alors* $SC_{Res} = CM_{Res} * ((I-1)(J-1)) =$ ***480***.
- $SC_{Tot} = CM_{Res} + SC_{F_1} + SC_{F_2}$, *alors* $SC_{F_2} = 1350 - 480 - 420 =$ ***450***.
- $CM_{F_2} = SC_{F_2}/(J-1)$, *alors* $CM_{F_2} =$ ***112.5***
- $F_{c_{F_2}} = CM_{F_2}/CM_{Res}$, *alors* $F_{c_{F_2}} = 112.5/40 =$ ***2.8125***

Ainsi, on aura la table suivante :

Variation	*SC*	*ddl*	*CM*	F_c
Facteur F_1	*420*	*3*	*140*	*3.5*
Facteur F_2	*450*	*4*	*112.5*	*2.8125*
Résiduelle	*480*	*12*	*40*	
Totale	*1350*	*19*		

TABLE *1.10: Table de l'ANOVA associée au problème*

2. Sur la table de la loi de Ficher pour un seuil de confiance 0.95 *(*$\alpha = 5\%$*) on obtient :*

$f_\alpha = f_{(3,12,0.95)} = 3.49 < 3.5 \Rightarrow$ *le premier facteur a une influence significative sur la variable* X.
$f_\alpha = f_{(4,12,0.95)} = 3.26 > 2.8125 \Rightarrow$ *le deuxième facteur n'a pas une influence significative sur* X.

Remarque 1.2

- *Dans la littérature on distingue trois types d'ANOVA (I, II et III) et cela selon la nature du/des facteur(s) étudié :*

  type I : *modèle à effet fixe lorsque les modalités des facteurs sont choisies délibérément par l'expérimentateur. C'est le cas dans la plupart des protocoles expérimentaux, et c'est le type que nous avons développé dans ce document.*

  type II : *modèle à effet aléatoire lorsque les modalités des facteurs sont issues d'un processus d'échantillonnage.*

  type III : *modèle à effets mixtes, lorsque on dispose des facteurs à effet fixe et des facteurs à effet aléatoire simultanément.*

- *Une autre classification de l'ANOVA existe également et cela selon la présentation des observations :*

  Plan avec répétitions *: chaque cellule contient plusieurs observations où elles peuvent contenir le même nombre d'observation (Plan équilibré) ou elles peuvent contenir un nombre différent d'observations (Plan non équilibré).*

  Plan sans répétitions *: chaque cellule contient une seule observation.*

  Plan à mesures répétées *: les mêmes sujets ont été utilisés pour les observations au sein de différentes cellules (appariement).*

Chapitre 2

Régression linéaire simple et multiple

Introduction et problématique

La régression est l'une des méthodes les plus connues et les plus appliquées en statistiques pour l'analyse de données quantitatives sous forme d'un modèle. Si on s'intéresse à la relation entre deux variables, on parlera de *régression simple* en exprimant l'une des deux variables en fonction de l'autre. Tandis que, si la relation porte entre une variable et plusieurs autres variables (≥ 2), on parlera de *régression multiple*.

La mise en œuvre d'une régression impose l'existence d'une relation de cause à effet entre les variables prises en compte dans le modèle. Cette méthode peut être mise en place sur des données quantitatives observées sur n individus et présentées sous la forme :

$$y = f(x) + \epsilon, \tag{2.1}$$

où

- y est une variable quantitative prenant la valeur y_i pour l'individu i $(i = 1, ..., n)$, appelée variable à expliquer ou variable réponse.
- $x_1, x_2, ..., x_p$ sont p variables quantitatives prenant respectivement les valeurs $x_{1i}, x_{2i}, ...x_{pi}$ pour le $i^{\text{ième}}$ individu, appelées variables explicatives ou prédicteurs.
- ϵ est une variable aléatoire (résidus).

Considérons un couple de variables quantitatives (X, Y). S'il existe une liaison entre ces deux variables, la connaissance de la valeur prise par X change notre incertitude concernant la réalisation de Y. Si l'on admet qu'il existe une relation de cause à effet entre X et Y, le phénomène aléatoire représenté par X peut donc servir à prédire celui représenté par Y et la liaison s'écrit sous la forme (2.1) et on dit que l'on fait de la régression de y sur x (dans le cas d'une régression multiple de y sur $x_1, x_2, ..., x_p$ la liaison peuvent être écrite sous la forme $y = f(x_1, x_2, ..., x_p)$).

Dans les cas les plus fréquents, on choisit l'ensemble des fonctions affinées du type :

Cas de régression linéaire simple :

$$f(x) = ax + b. \tag{2.2}$$

Cas de régression linéaire multiple :

$$f(x) = f(x_1, x_2, ..., x_p) = a_0 + a_1x_1 + a_2x_2 + ... + a_px_p. \tag{2.3}$$

2.1 Le modèle de régression linéaire simple

Soit un échantillon de n individus. Pour un individu i $(i = 1, ..., n)$, on a observé y_i la valeur de la réalisation de la variable quantitative Y et x_i la valeur de la variable quantitative x.

On veut étudier la relation entre ces deux variables, et en particulier, l'effet de x (variable explicative) sur y (variable réponse).

Dans un premier temps, on peut représenter graphiquement cette relation en traçant le nuage des n points de coordonnées (x_i, y_i) et on constate que la relation entre y_i et x_i s'écrit sous la forme d'un modèle de régression linéaire

La relation entre y et x est supposée n'être qu'approximative : elle est perturbée par un "terme d'erreur" additif, noté ϵ_i avec $E(\epsilon_i),\ i = \overline{1:n}$.

L'équation de la régression linéaire simple (ou le "modèle de régression") s'écrit donc de la façon suivante :

$$Y = a + b + \epsilon, \tag{2.4}$$

$$E(Y) = a + bE(x), \tag{2.5}$$

ou encore,

$$E(Y/x) = a + bx, \tag{2.6}$$

où a et b sont les paramètres du modèle, et ϵ est le terme d'erreur qui est une variable aléatoire.

Remarque 2.1

1. *a représente le point d'intersection de la droite de régression avec l'ordonnée ("intercept", "constante").*
2. *b représente la pente de la droite de régression.*
3. *La valeur de b donne le nombre d'unités supplémentaires de Y associées à une augmentation par une unité de x.*
4. *$E(Y/x)$ est la moyenne de Y pour une valeur de x donnée.*

Exemple 1

(a) : *$Y_i = a + bx_i + cx_i^2 + \epsilon_i$, est un modèle linéaire tandis que la relation entre x et y n'est pas linéaire mais de type polynomial.*

(b) : *$Y_i = a + b\ cos(x_i) + \epsilon_i$, est un modèle linéaire.*

(c) : *$Y_i = ae^{bx_i} + \epsilon_i$, n'est pas un modèle linéaire.*

(d) : *$Y_i = ab + cx_i + \epsilon_i$, n'est pas un modèle linéaire.*

Enfin, la linéarité est relier aux paramètres du modèle et non pas aux variables explicatives.

2.2 Analyse du modèle de régression linéaire simple

Soit le couple (X, Y) de variable aléatoire où X est une variable indépendante et Y la variable dépendante. On cherche une relation du type

$$Y = a + bx + \epsilon.$$

Notons que la mise en oeuvre et l'exploitation de ce modèle nécessite une quantification préalable des paramètres inconnus a et b.

2.2.1 Estimation des paramètres du modèle

On suppose que la variable X est contrôlée par l'expérimentateur où il réalise n expériences y_1 , y_2 , $y_3, \ldots$, y_n aux points x_1 , x_2 , x_3 , $\ldots$, x_n fixés. De plus, on suppose que les Y_i sont mutuellement indépendants.

Le modèle s'écrit

$$y_i = a + bx_i$$

pour $i = \overline{1:n}$, tel que :

- $E(\epsilon_i) = 0$
- $Cov(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$
- $Var(\epsilon_i) = \sigma^2 \quad \forall i = \overline{1:n}$

Supposons qu'on opte pour la méthode des moindre carrées pour quantifier a et b, alors les estimateurs des paramètres a et b sont $\hat{a}$ et $\hat{b}$ qui minimise la fonction $Q(a, b)$, définie par :

$$Q(a,b) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - a - bx)^2. \tag{2.7}$$

Cela revient à la détermination d'un optimum minimal de la fonction des erreurs quadratique $Q(a, b)$, qui consiste à résoudre le système des équations suivant :

$$\left\{ \begin{array}{rcl} \frac{\partial Q(a,b)}{\partial a} & = & 0, \\ \frac{\partial Q(a,b)}{\partial b} & = & 0, \end{array} \right. \tag{2.8}$$

c'est-à-dire,

$$\left\{ \begin{array}{rcl} -2\sum\limits_{i=1}^{n}(Y_i - a - bx) & = & 0 \\ -2\sum\limits_{i=1}^{n} x_i(Y_i - a - bx) & = & 0 \end{array} \right. \Rightarrow \left\{ \begin{array}{rcccccl} \sum\limits_{i=1}^{n} Y_i & - & \sum\limits_{i=1}^{n} a & - & b\sum\limits_{i=1}^{n} x_i & = & 0 \\ \sum\limits_{i=1}^{n} Y_i x_i & - & a\sum\limits_{i=1}^{n} x_i & - & b\sum\limits_{i=1}^{n} x_i^2 & = & 0 \end{array} \right. . \tag{2.9}$$

Finalement, le système à résoudre, pour estimer les coefficients de régression a et b, ni rien d'autre qu'un système linéaire à deux équations et à deux inconnus, qui est donné par :

$$\left\{ \begin{array}{rcccl} a & \left(\sum\limits_{i=1}^{n} 1\right) & + & b\left(\sum\limits_{i=1}^{n} x_i\right) & = & \sum\limits_{i=1}^{n} Y_i \\ a & \left(\sum\limits_{i=1}^{n} x_i\right) & + & b\left(\sum\limits_{i=1}^{n} x_i^2\right) & = & \sum\limits_{i=1}^{n} Y_i x_i \end{array} \right. \tag{2.10}$$

La résolution du système (2.10), nous fournis la solution suivante :

$$\boxed{\hat{b} = \frac{\frac{1}{n}\sum\limits_{i=n}^{n} x_i y_i - \overline{X}\ \overline{Y}}{\frac{1}{n}\sum\limits_{i=1}^{n} x_i^2 - \overline{X}^2}} \text{ et } \boxed{\hat{a} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{b}\frac{1}{n}\sum_{i=1}^{n} x_i}, \tag{2.11}$$

ou encore :

$$\boxed{\hat{b} = \frac{Cov(x,y)}{Var(x)}} \text{ et } \boxed{\hat{a} = \overline{Y} - \hat{b}\overline{X},} \tag{2.12}$$

où :

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right] - \overline{X}\ \overline{Y}. \tag{2.13}$$

2.2.2 Estimation de σ^2

En plus de l'estimation des paramètres du modèle (a et b), l'une des caractéristique statistique importante liée au modèle est bien que la variance inconnue σ^2. Pour cela, nous allons estimer σ^2, où nous proposons d'utiliser la méthode MLE (voir chapitre 2). A cet effet, on suppose que $\epsilon_i \rightsquigarrow \mathcal{N}(0, \sigma^2)$, alors

$$Y_i \rightsquigarrow \mathcal{N}(a + bx_i, \sigma^2).$$

Dans ce cas, la fonction de vraisemblance correspondante au modèle est donnée par :

$$\mathcal{L}(Y_1, Y_2, ..., Y_n, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2\right],$$

L'expression de la variance qui maximise cette fonction est donnée par :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2. \tag{2.14}$$

Mais, cet estimateur est un estimateur avec Biais, alors il doit être corriger. Ainsi, après sa correction on aura l'estimateur sans Biais de σ^2 suivant :

$$\hat{\sigma}_c^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 \rightsquigarrow \chi^2_{n-2}. \tag{2.15}$$

2.2.3 Qualité et validation du modèle :

Dans cette section, nous allons présenter deux manière du jugé la qualité et l'adéquation du modèle linéaire :

$$Y_i = a + bx_i + \epsilon_i \quad , i = 1, ..., n.$$

pour l'explication de la variable Y à l'aide de la variable x.

2.2.3.1 Coefficients de correlation et de détermination

En probabilité et en statistique, étudier la corrélation entre deux ou plusieurs variables aléatoire, c'est étudier l'intensité de la liaison qui peut être existée entre ces variables.

Une mesure de cette corrélation dans le cadre linéaire est obtenue par le calcul du coefficient appelé coefficient de corrélation. Ce coefficient est égal au rapport de leurs covariances et du produit non nul de leurs écarts types :

$$\rho = Cor(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = r(x, y). \tag{2.16}$$

avec,

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2, \tag{2.17}$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{Y})^2, \tag{2.18}$$

et $Cov(x, y)$ est donnée dans (2.13).

Le coefficient de corrélation est toujours compris entre -1 et +1. De plus, son signe donne le sens de la corrélation où le signe positif indique que les deux variables sont proportionnelles

dans le même sens, tandis que le signe négatif indique que les deux variables sont inversement proportionnelles.

Plus $|\rho|$ est près de 1, plus la corrélation est grande donc le modèle linéaire décrit bien le phénomène étudié. Par contre, si $|\rho|$ est près de zéro le modèle linéaire n'est pas adéquat pour la modélisation du problème étudié.

Le coefficient de corrélation nous donne des informations sur l'existence d'une relation linéaire entre les deux variables considérées. A cet effet, un coefficient de corrélation nul ne signifie pas l'absence de toute relation entre les deux variables mais seulement l'absence d'une relation linéaire. Pour cela, il ne faut pas confondre la corrélation et la relation causale. Une bonne corrélation entre deux variables peut révéler une relation de cause à effet entre elle, mais pas nécessairement.

Pour mieux juger la qualité d'une régression linéaire, on définit un autre indicateur compris entre 0 et 1, nommé : *coefficient de détermination*, noté R^2 :

$$R^2 = \rho^2.$$

Ce nombre mesure l'adéquation entre le modèle et les données observées où plus, R^2 est près de 1, plus le modèle est adéquat et le contraire est vrai.

2.2.3.2 Le test de Fisher

Une autre technique, plus puissante que le calcule de coefficient de corrélation, pour mesuré la pertinence et l'adéquation d'un modèle est l'utilisation du test de Ficher qui se base sur l'analyse de la variance.

On peut démontrer que la variation totale de Y se décompose comme suit :

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{Y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SCR} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2}_{SCE}$$

où

SCT : Variation de Y ou variation totale.

SCR : Variation des résidus.

SCE : Variation de la régression ou variation expliqué par la régression.

Dans la logique des choses, le modèle est validé, si la variation totale du modèle n'est engendré que par la variation des résidus et non pas par la variation de la régression, autrement dit la variation moyenne des résidus doit être supérieur à la variation moyenne de la régression pour valider le modèle,

$$\frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)} > k > 1,$$

donc, il nous reste à savoir comment déterminer la valeur critique k.

Sachant que :

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \rightsquigarrow \chi^2_{n-2}$$

et

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2 \rightsquigarrow \chi^2_1,$$

alors,

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)} \rightsquigarrow f_{(1,n-2)}$$

où la notation $f_{(1,n-2)}$ désigne est une loi de Fisher de degrés de liberté $n_1 = 1$ et $n_2 = n-2$, cela signifie que, pour un risque α, la valeur critique k n'est rien d'autre que le fractale d'ordre $1-\alpha$ d'une loi de Fisher de degrés de liberté 1 et $n-1$ ($k = f_{(1,n-2,1-\alpha)}$) ainsi on décide que :

- Si $f_c > f_{(1,n-2,1-\alpha)}$ alors le modèle est valide.
- Si $f_c \leq f_{(1,n-2,1-\alpha)}$ le modèle n'est pas valide.

où f_c est la réalisation de la statistique F.

2.3 Régression linéaire multiple

Dans la pratique les principales étapes d'une analyse de régression Multiple sont :

1. Définir la variable dépendante et les variables explicatives.
2. Spécifier la nature de la relation entre la variable dépendante et les variables explicatives.
3. Estimer les paramètres du modèle, en suite, quantifier sa qualité et vérifier sa validité.
4. Dans le cas où le modèle est retenu, interpréter sa signification par rapport au problème posé.

Dans cette section, nous nous intéresserons à la régression multiple dans le cadre du modèle linéaire. La régression linéaire multiple est la généralisation de la régression linéaire simple qui ne considère qu'une seule variable explicative. Considérons le modèle linéaire multiple dont la forme est la suivante :

$$Y = b_1 x_1 + b_2 x_2 + ... + b_k x_k + \epsilon,$$

pour la $i^{\text{ième}}$ observation le modèle peut être représenté de la manière suivante :

$$Y_i = b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + \epsilon_i \ , \qquad i = 1, ..., n,$$

ou encore, sous sa forme Matricielle :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & ... & x_{k1} \\ x_{12} & x_{22} & ... & x_{k2} \\ \vdots & & & \vdots \\ x_{1n} & x_{2n} & ... & x_{kn} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

$$Y \quad = \quad X \quad b \quad + \quad \epsilon$$

A partir des étapes de l'analyse de régression multiple, cité précédemment, on est au niveau de l'étape (3), c'est-à-dire on doit estimer les paramètres (coefficients) du modèle.

2.3.1 Estimation des paramètres du modèle

Supposons qu'on a :

$$Y = Xb + \epsilon,$$

avec,

$$X = \begin{bmatrix} x_{11} & x_{21} & ... & x_{k1} \\ x_{12} & x_{22} & ... & x_{k2} \\ \vdots & & & \vdots \\ x_{1n} & x_{2n} & ... & x_{kn} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad et \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

de plus,

- $E(\epsilon) = 0$,
- $Var(\epsilon) = \sigma^2 I_n$, où I_n est une matrice d'identités d'ordre n.

On utilisant la méthode des moindres carrés, pour estimer les coefficients du modèle, on aura un système linéaire à k équations et k variables. Ce système s'écrit sous sa forme matricielle comme suit :

$$\underbrace{\begin{bmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & ... & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{1i}^2 & ... & \sum_{i=1}^{n} x_{2i}x_{ki} \\ \vdots & & & \vdots \\ \sum_{i=1}^{n} x_{1i}x_{ki} & ... & ... & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}}_{M} \times \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}}_{b} = \underbrace{\begin{bmatrix} \sum_{i=1}^{n} y_i x_{1i} \\ \sum_{i=1}^{n} y_i x_{2i} \\ \vdots \\ \sum_{i=1}^{n} y_i x_{ki} \end{bmatrix}}_{m}$$

où $M = X^t\ X$ et $m = X^t\ Y$.

Finalement, l'estimation des coefficients du modèle sont données par le calcul matriciel suivant :

$$\hat{b} = (X^t.X)^{-1}.X^t.Y.$$

2.3.2 Test sur la validité du modèle

Avec le même raisonnement abordé dans le cas de la regression linéaire simple on peut construire le test de validation du modèle. En effet, la variation totale de Y se décompose comme suit :

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{Y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SCR} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2}_{SCE}$$

où

SCT : Variation de Y ou variation totale.

SCR : Variation des résidus.

SCE : Variation de la régression ou variation expliqué par la régression.

Pour valider le modèle, on test

$$H_0 \text{ "} b_1 = b_2 = ... = b_k = 0\text{" contre } H_1 \text{ "} \exists j \in \{1,\ 2, ...,\ k\} /\ b_j \neq 0\text{"},$$

avec le même raisonnement que dans le cas de régression linéaire on obtient la statistique du test suivante :

$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{Y})^2/(k-1)}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-k)} \rightsquigarrow f_{(k-1,n-k)},$$

où la notation $f_{(k-1,n-k)}$ désigne la loi de Fisher à $k-1$ et $n-k$ degrés de liberté.

Ainsi, on décide :

- Si $f_c > f_{(k-1,n-k,1-\alpha)} \Rightarrow$ de valider le modèle.
- Si $f_c \leq f_{(k-1,n-k,1-\alpha)} \Rightarrow$ de ne pas valider le modèle.

avec f_c est la réalisation de la statistique F.

Chapitre 3

Exercices corrigés

Introduction

Dans ce chapitre nous proposons quelques exercices avec solutions détaillées qui sont ordonnés selon les notions introduites dans les chapitres précédents afin de permettre aux utilisateurs de ce polycopié de contrôler l'acquisition des notions essentielles qui ont été introduites.

3.1 Énoncés des exercices

Exercice 2 *(Analyse de la variance à un seul facteur)*

Lors d'une expérimentation pédagogique, on désire comparer l'efficacité de quatre méthodes d'enseignement. On dispose des notes obtenues à un examen par quatre groupes d'étudiants (chaque groupe contient 25 étudiants) ayant chacun reçu un des 4 types d'enseignement a, b, c ou d. Pour répondre à l'objectif la technique statistique la plus adéquate est bien que l'ANOVA à seul facteur. L'application de cette dernière nous a fournis ce qui suit :

	SC	ddl	MC	F
Inter-groupes (Fac)			31.82	6.64
Intra-groupes (Rés)				
Total				

1. Completer la table d'analyse de variance ci-dessous.
2. A un seuil de risque $\alpha = 5\%$, que-peut-on conclure sur l'efficacité moyens des 4 méthodes.

Exercice 3 *(ANOVA 1 à un plan équilibré)*

Nous souhaitons comparer quatre traitements, notés A, B , C et D. Nous répartissons par tirage au sort les patients, et nous leur affectons l'un des quatre traitements. Nous mesurons sur chaque patient la durée, en jours, séparant de la prochaine crise d'asthme. Les mesures sont reportées dans le tableau ci-dessous :

Traitement A	*Traitement B*	*Traitement C*	*Traitement D*
36 ; 37 ; 35 ; 38 ; 41	*42 ; 38 ; 39 ; 42 ; 44*	*26 ; 26 ; 30 38 ; 34*	*42 ; 45 ; 50 ; 56 ; 58*

Pouvons-nous conclure, à un seuil de risque 1%, que les facteur traitement a une influence sur le critère retenue ? (On donne $SCT = 1324.55$.)

Exercice 4 *(ANOVA 1 à un plan non équilibré)*

On s'intéresse au rendement d'orge pour quatre variétés différentes. On dispose de quatre parcelles avec une variété d'orge pour chacune. On répète cette expérience à des endroits différents. On a obtenu :

	variété 1	variété 2	variété 3	variété 4
	46	57	50	39
	43	53	41	51
	48	43	47	45
		54	42	43
		48		
Somme	137	255	180	178

1. Calculer les estimations des moyennes μ_1, μ_2, μ_3, μ_4 et m.
2. Considérons l'hypothèse (H_0) : les rendements moyens de chaque variété sont égaux.
   a) Donner la table d'analyse de variance du problème posé, sachant que la somme des carrés résiduels $SC_{Res} = 263.667$.
   b) Que peut-on conclure sur le rendement de chaque variété (l'hypothèse H_0) à un seuil de risque $\alpha = 1\%$.

Exercice 5 *(ANOVA 2 à un plan équilibré)*

Les données sont issues d'une expérience dans laquelle la concentration de calcium dans le plasma a été mesurée chez 20 sujets des deux sexes ayant subi ou non l'administration d'un traitement hormonal.
Les données individuelles en fonction des deux traitements (h_1 et h_2) et les deux sexes (S_1 et S_2) sont résumées dans le tableau suivant

h_1		h_2	
S_1	S_2	S_1	S_2
16.5	14.5	39.1	32.0
18.4	11.0	26.2	23.8
12.7	10.8	21.3	28.8
14.0	14.3	35.8	25.0
12.8	10.0	40.2	29.3

Supposons qu'on désir tester l'influence du traitement, du sexe et les deux simultanément sur concentration de calcium.

1. Donner les hypothèses à tester.
2. Quelle techniques statistiques qui nous permet de réaliser le test citer en (1).
3. Appliquer la techniques citer en (2), pour répondre à notre objectif.

Exercice 6 *(ANOVA 2 à un plan sans répétitions)*

Trois équipes (matin, midi, soir) se relaient sur une chaîne de montage. Elles occupent quatre post de travail A, B, C et D. Sur la production d'un mois, on note le nombre total de pièces défectueuses ventilées par équipe et par post de travail :

	post				
équipe	A	B	C	D	moyenne
équipe du matin	26	13	35	6	20
équipe du soir	18	17	31	2	17
équipe de nuit	31	24	33	4	23
moyenne	25	18	33	4	20

On souhaite interpréter ce tableau à l'aide d'une ANOVA.

1. Dans un premier temps, on analyse les deux facteurs équipe et post séparément.
   (a) Peut-on affirmer l'existence d'une différence entre les performances globales des équipes ?
   (b) Peut-on conclure que les post présentent des difficultés de montage inégales ?
2. On désire maintenant tenir compte simultanément des deux facteurs. On modélise le tableau par un modèle additif d'analyse de la variance à deux facteurs :

$$Y_{ij} = m + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij},$$

   où les variables aléatoires, ϵ_{ij} sont indépendantes, normales centrées de même variance σ^2
   (a) Serait-il pertinent de modéliser le tableau par un modèle complet à deux facteurs ?
   (b) Y-a-t-il un effet "post de travail" ?
   (c) Y-a-t-il un effet "équipe" ?
   (d) Y-a-t-il un effet "équipe" et "post de travail" ? que peut-on conclure ?
   (e) Comparer les résultats obtenus dans 1 et 2, que peut-on conclure ?

Exercice 7 *(ANOVA 2 à un plan sans répétitions)*

Un botaniste veut déterminer si la présence d'insectes a un effet sur la fécondité des plantes dans un champ. Afin d'empêcher les insectes d'attaquer ses plantes, le botaniste a l'idée d'installer des cages au-dessus des quadrants expérimentaux. Il propose d'utiliser trois traitements : contrôle (plantes non couvertes), plantes recouvertes de cages laissant les insectes entrer et plantes recouvertes de cages empêchant les insectes d'atteindre les plantes. Pour s'assurer que les différences qu'il observe à la fin de l'étude sont dues aux manipulations et non à un effet quelconque des propriétés des cages utilisées, il utilise 3 quadrants par traitement, et il échantillonne 6 plantes par quadrant. La variable mesurée est la fécondité moyenne (le nombre moyen de graines produit par les 6 plantes). Les données se trouvent ci-dessous.

		Contrôle	Cages fermées	Cages ouvertes
Quadrants	1	78.4	71.2	76.2
	2	75.2	61.8	77.3
	3	83.5	68.5	77.7

Question : A l'aide d'une technique statistique adéquate, indiquer si la présence d'insectes ainsi que les cages ont un effet significatif sur la fécondité moyenne des plantes ou non.

Exercice 8 *(Test d'indépendance de $Khi - Deux$)*

En vue de comparer deux traitements T_1 et T_2 d'une affection bénigne, on répartit entre ces deux traitements 250 malades par tirage au sort. Les résultats, sur l'état du malade après 5 jours de traitement, sont indiqués dans le tableau ci-dessous.

	État du malade après 5 jours		
Traitement	Stationnaire	Amélioré	Guéri
T_1	15	70	35
T_2	25	85	20

Les deux variables Traitement et l'état du malade après 5 jours de traitement sont ils indépendantes ?

Exercice 9 *(Test d'indépendance de $Khi - Deux$)*

On s'intéresse à l'association entre le mode de vie, "seul" ou "en famille", et la présence ou

l'absence d'une névrose. Dans un échantillon aléatoire d'individus d'une certaine population on a trouve les fréquences ci-dessous :

	Névrose		
Mode de vie	Présente	Absente	Total
En famille	40	60	100
Seul	100	60	160
Total	140	120	260

Question : Peut-on rejeter, au seuil de 1%, l'hypothèse de non association entre le mode de vie et la présence d'une névrose ?

Exercice 10 *(Test d'indépendance de Khi − Deux)*

Lors de l'interrogation du module "*Analyse de Données en Biosciences*", des troisièmes années Licence option Biologie et physiologie végétale qui contiens trois groupes ($G01$, $G02$ et $G03$), dans la grande salle numéro 1 qui contiens quatre (04) rangées, le responsable de module s'est posé la question suivante :
Est-ce-que les étudiants choisissent le rang pour s'asseoir selon leurs groupes ou non ?
Après l'analyse de la situation l'enseignant a conclu que le choix du rang est indépendant du groupe. Indiquer, pour un seuil de risque $\alpha = 5\%$, si l'enseignant à raison ou non. Sachant que la répartition des étudiants dans la salle selon leurs groupes est comme suite :

Groupe	Rang 1	Rang 2	Rang 3	Rang 4	$\sum$
Groupe 1	4	6	8	6	24
Groupe 2	4	3	5	5	17
Groupe 3	6	5	3	2	16
$\sum$	14	14	16	13	57

Exercice 11 *(Test d'indépendance de Khi − Deux)*

Afin de comparer l'action de deux types de levures (A et B) sur une pâte à gâteaux, on prélève, pour chacune des levures, un échantillon aléatoire de gâteaux. L'aptitude des pâtes à lever est définie par les critères suivants : Moyenne, Bonne et Très bonne. Les résultats constatés sont rassemblés dans le tableau suivant :

Aptitude à lever	Moyenne	Bonne	Très bonne
A	41	16	63
B	22	27	51

Quel est le test statistique adéquat pour déterminer s'il y a une différence d'activité des deux levures ? À l'aide de ce test, au risque de 5%, peut-on conclure à une différence d'activité des deux levures ?

Exercice 12 *(ANOVA 1 et test d'indépendance de Khi − Deux)*

Trente sept étudiants d'une promotion ont été répartis, en début d'année académique, de manière strictement aléatoire dans trois séries de travaux pratiques de statistique dirigés par trois assistants différents $A1$, $A2$ et $A3$. Les résultats obtenus par les étudiants de chaque série sont notés sur 10 et regroupés dans le tableau suivant.

Assistant	*Note des étudiants (sur 10)*												
$A1$	*9*	*5.5*	*6*	*3*	*3*	*2*	*7*	*5*	*0*	*8*	*4.5*	*7*	×
$A2$	*4*	*3*	*6*	*8*	*2*	*3*	*5*	*7*	*7*	*4.5*	*3.5*	*0*	×
$A3$	*8*	*6*	*4*	*8*	*10*	*4*	*4.5*	*5*	*7*	*8*	*10*	*9*	*6*

Y-a-t-il un effet d'appartenance sur le niveau des étudiants ?

Afin d'étudier l'indépendance des résultats par rapport à la série d'appartenance de l'étudiant, un chercheur fait un décompte en termes de nombre de réussites et d'échecs par série de travaux pratiques.
Quel test que le chercheur doit utiliser dans ce cas ? Réaliser ce test sur les présentes données.

Exercice 13 *(Régression linéaire simple)*

Dans le cadre de travaux de recherche sur la *Biomasse* (mg), d'un certain type de plante, en fonction de la concentration de l'Azote NH_4^+ (μmol), nous avons réalisé des expériences dont la biomasse moyenne (Y) ainsi que la concertation du l'Azote (X) en question sont données dans le tableau ci-dessus :

Concentration μmol	0	100	200	400	600
Biomasse mg	305	378	458	540	565

On donne : $\sum x_i = 1300$; $\sum y_i = 2246$; $\sum x_i^2 = 570000$; $\sum y_i^2 = 1056498$; $\sum x_i y_i = 684400$;

Afin de modéliser ces données, nous avons proposé le modèle linéaire suivant :

$$Y \ = \ a \ + \ bx.$$

1. Présenter graphiquement le nuage des points (X_i, Y_i). Que peut-on conclure sur le modèle proposer ?
2. Calculer les estimations des paramètres a et b et donner la droite de régression.
3. Calculer le coefficient de corrélation linéaire. Que peut-on conclure ?
4. Pour un seuil de risque $\alpha = 5\%$, le modèle proposé est-il pertinent ?
5. Quelle Biomasse prévoyez-vous à une concentration 500 μmol ?

Exercice 14 *(Régression linéaire simple)*

Dans le cadre de travaux de recherche sur la durée de la saison de végétation en montagne, des stations météorologiques sont installées à différentes altitudes. La température moyenne (variable Y en degrés Celsius) ainsi que l'altitude (variable X en mètres) de chaque station données dans le tableau ci-dessous :

altitude	*1040*	*1230*	*1500*	*1600*	*1740*	*1950*	*2200*	*2530*	*2800*	*3100*
température	*7.4*	*6*	*4.5*	*3.8*	*2.9*	*1.9*	*1*	*-1.2*	*-1.5*	*-4.5*

On donne : $\sum x_i = 19690$; $\sum y_i = 20.3$; $\sum x_i^2 = 42925500$; $\sum y_i^2 = 162.41$; $\sum x_i y_i = 17671$;

1. Calculer le coefficient de corrélation linéaire.
2. Calculer les estimations des paramètres a, b et σ^2 pour la régression linéaire de Y sur X.
3. Quelle température moyenne prévoyez-vous à 1100 m ? à 2300 m ?

Exercice 15 *(Régression linéaire simple)*

Dans le cadre d'une enquête visant à comparer, selon certains critères, différents *Sandwich* vendus dans les fast-foods, nous avons retenu les informations se trouvant dans le tableau ci-dessous.

Sandwich	S_1	S_2	S_3	S_4	S_5	S_6
Poids (g)	150	92	193	90	135	169
Prix (DA)	190	140	270	90	180	130

On donne :
$\sum Poids = 829$; $\sum Poids^2 = 123099$; $\sum Prix = 1000$; $\sum Prix^2 = 186000$; et
$\sum Poids * Prix = 147860$;

1. Y-a-t-il une relation linéaire entre les variables Poids et Prix ? Pour répondre à cette question, faire un graphique, calculer le coefficient de corrélation des deux variables et l'équation de la droite de régression.
2. Supposons qu'on désir d'augmenter le poids du Sandwich S_6 à 180 g, alors quelle sera son nouveau prix ?
3. Le modèle linéaire proposé est-il adéquat pour la description de la relation entre les variables Poids et Prix ?

Exercice 16 *(Régression linéaire simple et transformation des variables)*

On veut prédire la hauteur H d'un arbre en fonction de son diamètre D. Pour faire une régression linéaire, on effectue un changement de variable en posant $X = \ln(D)$ et $Y = \ln(H)$. Voici les mesures faites sur 5 arbres :

D	*0.1999*	*0.3012*	*0.3791*	*0.6005*	*0.6570*
H	*9.2073*	*9.6794*	*10.8049*	*13.4637*	*14.1540*

1. Donner le coefficient de corrélation linéaire entre X et$Y$.
2. Donner l'équation de la droite de régression de Y par rapport à X.
3. Tester la pertinence de la régression au seuil de 5%.
4. Donner la hauteur prévue d'un arbre de diamètre 0.7.

Exercice 17 *(Régression linéaire simple et changement des variables)*

Dans le cadre de travaux de recherche sur l'absorbance, d'un produit en fonction de sa concentration, par une certaine plante, nous avons réalisé des expériences dont l'absorbance moyenne (Y) ainsi que la concertation du produit (x) en question sont données dans le tableau ci-dessus :

							Somme
X $\mu g/\mu l$	0	20	40	60	80	100	300
Y	0	0.205	0.331	0.515	0.584	0.671	2.3060

a) Afin de modéliser ces données, nous avons proposé le modèle linéaire suivant :

$$Y \ = \ a_1 \ x \ + \ b_1.$$

1. Calculer les estimations des paramètres a_1 et b_1 et donner la droite de régression.
2. Quelle absorbance prévoyez-vous à une concentration 40 $\mu g/\mu l$? Que peut-on conclure ?
3. Calculer le coefficient de corrélation linéaire, ce résultat confirme t-il les résultats obtenue en 3) ?
4. Pour un seuil de risque $\alpha = 5\%$, le modèle proposé est-il pertinent ?

b) Vue les doutes qu'on a sur le modèle précèdent, nous avons proposé le modèle suivant :

$$Z \ = \ e^Y \ = \ a_2 \ x \ + \ b_2.$$

1. Complétez le tableau suivant :

							Somme
X $\mu g/\mu l$	0	20	40	60	80	100	300
Z	1.0000						

2. Calculer les estimations des paramètres a_2 et b_2 pour la régression linéaire de Z sur X.
3. Quelle absorbance prévoyez-vous à une concentration 40 $\mu g/\mu l$. Que peut-on conclure par rapport au premier modèle ?
4. Calculer le coefficient de corrélation linéaire de ce nouveau modèle.
5. Indiquer quel est le meilleur modèle parmi les deux proposés (avec justification).

3.2 Solution des exercices

Solution 1 *(Analyse de la variance à un seul facteur)*

À partir de l'énoncé on a $p = 4$ et $n = p * 25 = 100$, ces données nous permet de déterminer facilement les différents degrés de liberté ($d.d.l$ de la troisième colonne de la table d'ANOVA 1.

1. $p - 1 = \mathbf{3}$.
2. $n - p = \mathbf{96}$.
3. $n - 1 = \mathbf{99}$.

Pour le reste des quantités on a :

- $CM_{F_1} = SC_{F_1}/(I-1) = 31.82$, alors $SC_{F_1} = CM_{F_1} * (I-1) = 31.82 * 3 =$**95.46**.
- $F_c = CM_{F_1}/CM_{Res} = 6.64$, alors $CM_{Res} = CM_{F_1}/F_c = 31.82/6.64 =$ **4.7922**.
- $CM_{Res} = SC_{Res}/(n-p)$, alors $SC_{Res} = CM_{Res} * (n-p) = 4.7922 * 96 =$ **460.0512**.
- $SC_{Tot} = CM_{Res} + SC_{F_1}$, alors $SC_{Tot} = 95.46 + 464.8434 =$ **555.5112**.

Ainsi, on aura la table suivante :

	Somme des carrés	*ddl*	Moyenne des carrés	*F*
Inter-groupes	95.46	3	31.82	6.64
Intra-groupes	460.0512	96	4.7922	
Total	555.5112	99		

TABLE 3.5: *Table de l'ANOVA associée au problème*

Sur la table de la loi de Ficher pour un seuil de confiance 0.95 ($\alpha = 5\%$) on obtient :
$f_\alpha = f_{(3,96,0.95)} \approx f_{(3,100,0.95)} = 2.70 < 6.64 \Rightarrow$ le facteur "Méthode d'enseignement" a une influence significative sur le niveau des étudiants.

Solution 2 *(ANOVA 1 à un plan équilibré)*

Les différentes moyennes (de chaque échantillon et globale) sont données par :

$$\overline{X}_1 = 37.40,\ \overline{X}_2 = 41.00,\ \overline{X}_3 = 30.80,\ \overline{X}_4 = 50.20 \text{ et } \overline{X} = 39.85.$$

En exploitant ces dernières quantités pour le calcul des différentes variations on obtient :

	SC	*ddl*	*CM*	*f*	f_α
Inter-groupes	961.2667	3	320.4222	14.1123	5.29
Intra-groupes	363.2833	16	22.7052		
Total	1324.55	19			

On constate que $f > f_\alpha$ cela signifie qu'on doit rejeter H_0. C'est-à-dire le facteur traitement a une influence significative sur les durées séparant deux crise d'asthme.

Solution 3 *(ANOVA 1 à un plan non équilibré)*

1. Les estimations des différentes moyennes μ_1, μ_2, μ_3, μ_4 et m sont données respectivement par :

   ▷ $\overline{X}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j} = \frac{1}{3}(137) = 45.67,$ ▷ $\overline{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j} = \frac{1}{5}(255) = 51.00,$

   ▷ $\overline{X}_3 = \frac{1}{n_3}\sum_{j=1}^{n_3} x_{3j} = \frac{1}{4}(180) = 45.00,$ ▷ $\overline{X}_4 = \frac{1}{n_4}\sum_{j=1}^{n_4} x_{4j} = \frac{1}{4}(178) = 44.50,$

   et $\overline{X} = \frac{1}{n}\sum_{i=1}^{4} n_i\overline{X}_i = \frac{1}{16}(3 * 45.67 + 5 * 51.00 + 4 * 45.00 + 4 * 44.50) = 46.8750.$

2. Considérons l'hypothèse (H_0) : les rendements moyens de chaque variété sont égaux.

   a) Dans ce cas, le problème est le même que celui de l'exercice 3. Contrairement à ce dernier, on remarque que les tailles des échantillons ne sont pas les mêmes, mais le raisonnement de l'ANOVA ne changera pas . En effet, la décomposition de la variation totale dans cette situation sera :

   $$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2}_{SC_{Tot}} = \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2}_{SC_{Res}} + \underbrace{\sum_{j=1}^{p} n_j(\overline{X}_j - \overline{X})^2}_{SC_{Fac}}, \tag{3.1}$$

   où, n_j : est la taille du $j^{\text{ième}}$ échantillon (groupe).

   SC_{Tot} : est la variation totale qui représente dispersion des données autour de la moyenne générale.

   SC_{Fac} : est la variation due au facteur qui représente dispersion des moyennes autour de la moyenne générale.

   SC_{Res} : est la variation résiduelle qui représente dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio F_{obs}
Inter-groupe	126.0833	3	42.0278	1.9128
Intra-groupe	263.6667	12	21.9722	
Total	389.7500	15		

   b) On a d'une part $F_{obs} = 1.9128$ et d'autre part $f_\alpha = f(3, 12, 1 - 0.01) = 5.9525$, alors on rejette pas H_0 car $F_{obs} < f_\alpha$, cela signifier qu'il y a pas une différence significatives entre les rendement des différentes variétés d'orge.

Solution 4 *(ANOVA 2 à un plan équilibré)*

Dans cet exercice le problème posé est l'analyse de l'influence d'un traitement hormonal, du sexe et leurs interactions sur la concentration de calcium dans le plasma d'un être humain.

1. Sachant que le modèle correspondant au problème est :

   $$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \text{ avec } i = \overline{1, I}, j = \overline{1, J} \text{ et } k = \overline{1, K},$$

   alors les hypothèses à tester sont :

Effet du premier facteur (le traitement hormonal) :

H_0 : "*les paramètres* α_i *sont tous nuls*" contre H_1 : "*les paramètres* α_i *ne sont pas tous nuls*"

Effet du second facteur (le sexe) :

H_0 : "*les paramètres* β_j *sont tous nuls* " contre H_1 : "*les paramètres* β_j *ne sont pas tous nuls*"

Effet de l'interaction des deux facteurs (le traitement et le sexe simultanément) :

H_0 : "*les paramètres* γ_{ij} *sont tous nuls* " contre H_1 : "*les paramètres* γ_{ij} *ne sont pas tous nuls*"

2. La technique adéquate pour l'analyse du problème est bien que l'ANOVA à deux facteur.
3. L'application de l'ANOVA 2 sur les données nous fournies ce qui suit :

   (a) Les caractéristiques descriptives des données qui sont résumées dans la table suivante :

Hormone	Sexe	Moyenne	Variance
Hormone 1	Homme	14.880	4.9736
	Femme	12.120	3.5816
	Total	13.500	6.1820
Hormone 2	Homme	32.520	55.7736
	Femme	27.780	8.9456
	Total	30.150	37.9765
Total	Homme	23.700	108.1660
	Femme	19.950	67.5725
	Total	21.825	91.3849

TABLE 3.9: *Analyse descriptive de la variable dépendante : Concentration de Calcium dans le plasma.*

   (b) La table de l'ANOVA 2 suivante :

Source	Somme des carrés	ddl	Moyenne des carrés	f_c
Hormone	1386.113	1	1386.113	60.534
Sexe	70.312	1	70.312	3.071
Hormone * Sexe	4.900	1	4.900	.214
Résiduelles	366.372	16	22.898	
Total	1827.698	19		

TABLE 3.10: *Décomposition de la variance en fonction des sources de variabilité, pour un plan factoriel à 2 facteurs avec réplications.*

Supposons qu'on desire prendre la décision avec risque $\alpha = 5\%$, de se tromper. Alors d'après les résultats obtenus (voir table 3.10), on constate que :
La présence hormonal lors du traitement, à un effet significatif sur la concentration de calcium dans le plasma (le fait que $f_c > f_{(1,16,1-0.05)} = 4.49$), par contre y a aucune différence de concentration du calcium chez la femme ou l'homme, ce qui reste vraie pour l'interaction Sexe et Hormone qui n'a aucun effet significatif sur la concentration du calcium dans le plasma (le fait que $f_c < 4.49$

pour la variable Sexe et l'interaction des deux variables).

Solution 5 *(ANOVA 2 à un plan sans répétitions)* La réponse à cet exercice consiste à réaliser une ANOVA à deux facteurs lorsque le plan d'expérience est sans répétitions de mesures. L'objectif du présent exercice est de mettre en évidence le lien de l'ANOVA 1 (facteur par facteur) avec l'ANOVA 2 lorsque n'y ont pas de répétitions. Le tableau suivant résume les caractéristiques des différents échantillons du problème posé :

	post					
équipe	A	B	C	D	Moyenne	Variance
équipe du matin	26	13	35	6	20	126.5
équipe du soir	18	17	31	2	17	105.5
équipe de nuit	31	24	33	4	23	131.5
Moyenne	25	18	33	4	20	113.5
Variance	28.6667	20.6667	2.6667	2.6667	6	127.1667

ANOVA 1 : Le tableau de l'analyse de la variance des pièces défectueuses selon le *facteur post*, pour $\alpha = 5\%$, est :

Source	SC	d.d.l	MC	f_c	$f_{(n_1,n_2,1-\alpha)}$
Inter-groupes	1362	3	454	22.146	4.0662
Intra-groupes	164	8	20.5		
Total	1526	11			

Le tableau de l'analyse de la variance des pièces défectueuses selon le *facteur équipe*, pour $\alpha = 5\%$, est :

Source	SC	d.d.l	MC	f_c	$f_{(n_1,n_2,1-\alpha)}$
Inter-groupes	72	2	36	0.223	4.2565
Intra-groupes	1454	9	161.556		
Total	1526	11			

D'après le tableau d'ANOVA du facteur *post* on constate que le facteur post a un effet significatif sur le nombre de pièces défectueuses et cela le fait que $f_c > f_\alpha$. D'après le tableau d'ANOVA du facteur *équipe* on constate que le facteur équipe n'a pas un effet significatif sur le nombre de pièces défectueuses et cela le fait que $f_c < f_\alpha$.

ANOVA 2 : Certes, il est pertinent de modéliser le tableau par un modèle à deux facteurs, mais au lieu du modèle complet :

$$Y_{ij} = m + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij},$$

on utilise le modèle suivant :

$$Y_{ij} = m + \alpha_i + \beta_j + \epsilon_{ij}.$$

A cet effet, pour $\alpha = 5\%$, on aura le tableau de l'analyse de la variance des pièces défectueuses selon le *facteur équipe* et le *facteur post* simultanément suivant :

Variation	*SC*	*ddl*	*CM*	F_{obs}	$f_{(n_1,n_2,1-\alpha)}$
post	1362	3	454	29.6093	4.7571
Équipe	72	2	36	2.3479	5.1433
Résiduelle	92	6	15.333		
Total	1526	11			

L'analyse de ce dernier tableau nous permet de faire les mêmes conclusions que dans la première partie de l'exercice (ANOVA 1).

Remarque : La réponse à cet exercice peut ce faire également par l'étude d'indépendance entre la variable équipe et la variable post on utilisons le test de $Khi - Deux$.

Solution 6 *(ANOVA 2 à un plan sans répétitions)* Deux solution sont possible dans cette exercice, et cela selon l'hypothèse posée. En effet, si nous considérons *les Quadrants* comme facteur donc on réalise ANOVA 2 sans répétions pour répondre au problème, si nous considérons *les Quadrants* comme réplication donc on réalise une ANOVA 1 pour répondre au problème. Les résultats correspondant aux deux situations sont résumés, respectivement, dans les deux tables suivantes.

Source	*SC*	*ddl*	*MC*	f_c	$f_{(n_1,n_2,1-\alpha)}$
Quadrants	*42.736*	*2*	*21.368*	*2.117*	*6.9443*
Cages	*242.696*	*2*	*121.348*	*12.025*	*6.9443*
Erreur	*40.364*	*4*	*10.091*		
Total	*325.796*	*8*			

TABLE *3.15: Table d'ANOVA 2 sans répétions*

Source	*SC*	*ddl*	*MC*	f_c	$f_{(n_1,n_2,1-\alpha)}$
Inter-groupes	*242.696*	*2*	*121.348*	*8.762*	*5.1433*
Intra-groupes	*83.100*	*6*	*13.850*		
Total	*325.796*	*8*			

TABLE *3.16: Table d'ANOVA 1*

- A partir les résultats rangés dans la table 3.15, on constate que pour un risque de 5%, les *Quadrants* n'ont pas d'influence sur la fécondité moyenne tandis que le facteur *cage* a une influence significative sur la fécondité. On conclue que la présence des insectes ont une influence sur fécondité moyenne des plantes en question.
- A partir les résultats rangés dans la table 3.16, on constate que pour un risque de 5%, le facteur *cage* a une influence significative sur la fécondité. On conclut que la présence des insectes ont une influence sur fécondité moyenne des plantes en question.

Solution 7 *(Test d'indépendance de $Khi - Deux$)*
Afin de vérifier si les deux variables *Traitement* et l'*état du malade après 5 jours de traitement* sont indépendantes, nous utilisons le test d'indépendance de $Khi - Deux$.
Les effectifs observés, attendus et ainsi que l'écart entre eux sont résumés dans le tableau suivant :

			État			Total
			Stationnaire	Amélioré	Guéri	
Traitement	T_1	Effectif	15	70	35	120
		Effectif théorique	19.2	74.4	26.4	120
		Résidu	-4.2	-4.4	8.6	
	T_2	Effectif	25	85	20	130
		Effectif théorique	20.8	80.6	28.6	130
		Résidu	4.2	4.4	-8.6	
Total	Effectif		40	155	55	250
	Effectif théorique		40.0	155.0	55.0	250

Pour répondre à notre objectif, il suffit de comparé la réalisation de la statistique de *Khi-Deux* avec la valeur tabulée de *Khi-Deux* pour $r-1$ degré de liberté et un risque α qu'on doit fixer préalablement. Supposons que ce dernier est fixé à $\alpha = 5\%$.
On a d'une part, la réalisation de la statistique du test est donnée par :

$$k_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} = 7.6548.$$

et d'autre par la valeur tabulée de $\chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((2-1)(3-1),0.05)} = 5.991$. A cet effet, on conclut que le traitement et l'état du malade après 5 jours sont dépendants (liés) le fait que $7.6548 > 5.991$.

Solution 8 *(Test d'indépendance de Khi − Deux)*

Le test statistique adéquat est bien que le test d'indépendance de χ^2. Les résultats fournis par ce test, appliqué sur nos données, sont résumés dans le tableau suivant :

			Névrose		Total
			Présente	Absente	
Mode de vie	En famille	Effectif	40	60	100
		Effectif théorique	53.8	46.2	100
		Résidu	-13.8	13.8	
	Seul	Effectif	100	60	160
		Effectif théorique	86.2	73.8	160
		Résidu	13.8	-13.8	
Total		Effectif	140	120	260

on a :

$$K_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(N_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} = 12.536. \tag{3.2}$$

$$et \quad \chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((2-1)(2-1),0.01)} = \chi^2_{(1,0.01)} = 6.635. \tag{3.3}$$

De (3.2) et (3.3), on conclut qu'il y a un lien entre le mode de vie et la névrose le fait que $K_n^2 > \chi^2_{(1,0.01)}$.

Solution 9 *(Test d'indépendance de Khi − Deux)*

Le calcul des $n_{ij} = \frac{n_{i\bullet}\times n_{\bullet j}}{n}, i = 1, 2, 3$ *et* $j = 1, 2, 3, 4$, nous fournis ce qui suit :

			Rang				Total
			Rang 1	Rang 2	Rang 3	Rang 4	
Groupe	Groupe 1	Effectif	4	6	8	6	24
		Effectif théorique	5.9	5.9	6.7	5.5	24.0
		Résidu	-1.9	0.1	1.3	0.5	
	Groupe 2	Effectif	4	3	5	5	17
		Effectif théorique	4.2	4.2	4.8	3.9	17.0
		Résidu	-0.2	-1.2	0.2	1.1	
	Groupe 3	Effectif	6	5	3	2	16
		Effectif théorique	3.9	3.9	4.5	3.6	16.0
		Résidu	2.1	1.1	-1.5	-1.6	
Total		Effectif	14	14	16	13	57
		Effectif théorique	14.0	14.0	16.0	13.0	57.0

Pour répondre à notre objectif, il suffit de comparé la réalisation de la statistique de *Khi-Deux* avec la valeur tabulée de *Khi-Deux* pour $r - 1$ degré de liberté et un risque α qu'on doit fixer préalablement.

On a d'une part, la réalisation de la statistique du test est donnée par :

$$k_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} = 4.195.$$

et d'autre par la valeur tabulée de $\chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((4-1)(3-1),0.05)}$ est de 12.592. A cet effet, on conclu que le choix de rang est indépendant du groupe le fait que $k_n^2 < \chi^2_{(6,0.05)}$, cela signifié que les étudiants choisissent leurs places arbitrairement sans prendre en considération leurs appartenance à un groupe bien déterminé.

Solution 10 *(Test d'indépendance de Khi − Deux)*

Le test statistique adéquat pour déterminer s'il y a une différence d'activité des deux levures ou non est bien que le test d'indépendance ce χ^2. Les résultats fournis par ce test, appliqué sur nos données, sont résumés dans le tableau suivant :

			Aptitude à lever			Total
			Moyenne	Bonne	Très Bonne	
Levure	*A*	Effectif	41	16	63	120
		Effectif théorique	34.4	23.5	62.2	120
		Résidu	6.6	-7.5	0.8	
	B	Effectif	22	27	51	100
		Effectif théorique	28.6	19.5	51.8	100
		Résidu	-6.6	7.5	-0.8	
Total	Effectif		43	63	114	220

Pour répondre à notre objectif, il suffit de comparé la réalisation de la statistique du test avec la valeur critique qui est le fractile d'ordre $\alpha = 5\%$ d'une loi de *Khi-Deux* à $(r-1)*(k-1)$ degré de liberté, avec $r = 2$ (nombre de levures) et $k = 3$ (nombre de critères).

$$K_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(N_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} \approx 8.1, \tag{3.4}$$

et la valeur critique du test est donnée par :

$$\chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((2-1)(3-1),0.05)} = 5.991 \tag{3.5}$$

De (3.4) et (3.5), on conclu qu'il y a une différence entre l'activité des deux levures.

Solution 11 *(Test d'indépendance de $Khi - Deux$)*

La réponse à la première question de cet exercices consiste à réalisé une analyse de la variance à un seul facteur sur les notes des étudiants selon l'assistant, tandis que la deuxième concerne l'indépendance des deux variables note de l'étudiant et l'assistant dont la réponse se fait par le test d'indépendance du $Khi - Deux$.

1. Le tableau suivant résume les caractéristiques descriptives (moyenne et variance) des différents échantillons du problème posé :

	n	Moyenne	Variance
Assistant 1	12	5.0000	6.3750
Assistant 2	12	4.4167	4.9514
Assistant 3	13	6.8846	4.2367
Total	37	5.4730	6.2966

La table 3.22 résume les résultats de l'analyse de la variance des notes des étudiants selon le facteur Assistant.

Source	SC	d.d.l	MC	f_c	$f_{(n_1,n_2,1-\alpha)}$
Inter-groupes	41.979	2	20.990	3.737	3.2759
Intra-groupes	190.994	34	5.617		
Total	232.973	36			

TABLE 3.22: *Table de l'ANOVA associée au problème des notes selon l'Assistant*

A partir de cette dernière table, on constate que $f_c > f_\alpha$ cela signifie que le choix de l'Assistant influe sur les notes moyenne des groupes.

2. On admet que l'étudiant a échoué s'il a une note inférieur strictement à 5 ($note < 5$) et il a

réussit si sa note est supérieure ou égale à 5 ($note \geq 5$), ansi le tableau croisé "Assistant" $\times$ "réussite et échec" est donné comme suit :

	État	
Assistant	Échec	réussite
A_1	5	7
A_2	7	5
A_3	3	10

Les effectifs observés, attendus et ainsi que l'écart entre eux sont résumés dans le tableau suivant :

			État		Total
			Échec	Réussite	
Assistant	A_1	Effectif	7	5	12
		Effectif théorique	4.9	7.1	12.0
		Résidu	2.1	-2.1	
	A_2	Effectif	5	7	12
		Effectif théorique	4.9	7.1	12.0
		Résidu	.1	-.1	
	A_3	Effectif	3	10	13
		Effectif théorique	5.3	7.7	13.0
		Résidu	-2.3	2.3	
Total		Effectif	15	22	37
		Effectif théorique	15.0	22.0	37.0

On a d'une part, la réalisation de la statistique du test est donnée par :

$$k_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} = 3.8027.$$

et d'autre par la valeur tabulée de $\chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((3-1)(2-1),0.05)}$ est 5.991. A cet effet, on conclu que les résultats des étudiants(échec ou réussite) sont indépendant du l'assistant qui assure le TP.

Solution 12 *(Régression linéaire simple)*

1. À partir de la présentation graphique (voir figure 3.1), on constate que le nuage des points est distribué sous une forme linéaire, à priori le modèle proposé est adéquat pour l'explication de Y en fonction de x.

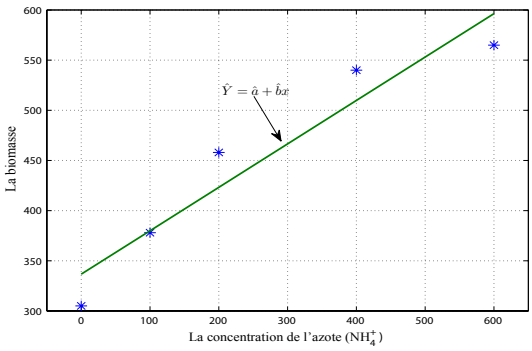

FIGURE 3.1: *Présentation graphique du nuage des points* (X_i, Y_i)

2. On a d'une part :

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\ \overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2} \qquad \text{et} \qquad \hat{a} = \overline{Y} - \hat{b}\overline{X}. \tag{3.6}$$

et d'autre part :

$$\begin{aligned} \overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}1300 = 260, \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}2246 = 449.2, \\ Cov(x,y) &= \frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\ \overline{Y} = \frac{1}{5}(684400) - (260)\ (449.2) = 20088, \\ Var(x) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 = \frac{1}{5}(570000) - (260)^2 = 46400, \end{aligned}$$

ainsi,

$$\hat{b} = 0.4329, \qquad \text{et} \qquad \hat{a} = 336.6460,$$

de ce fait, la droite de régression de la biomasse (Y) en fonction de la concentration (x) est :

$$\hat{Y} = 0.4329\ x\ +\ 336.6460.$$

3. On a d'une part,

$$r = r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}, \tag{3.7}$$

et d'autre part : $Cov(x,y) = 20088$, Écart-type$(x) = \sqrt{46400} = 215.4066$ et Écart-type$(Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2} = \sqrt{\frac{1}{5}(1056498) - (449.2)^2} = \sqrt{9518.36} = \mathbf{97.5621}$, alors

$$\rho = \rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} = 0.9559 = 95.59\%, \tag{3.8}$$

Le fait que la valeur de $\rho \approx 1$, on déduit qu'il y a une forte liaison linéaire entre x et Y.

4. Afin de valider le modèle nous aurons besoin des $\hat{Y}_i = 0.4329\ x_i\ +\ 336.6460$ dont leurs valeurs sont rangées dans le tableau suivant :

Concentration (μmol)	0	100	200	400	600
Biomasse (mg)	305	378	458	540	565
$\hat{y}_i\ (mg)$	336.646	379.936	423.226	509.806	596.386
$e_i = y_i - \hat{y}_i$	-31.646	-1.936	34.774	30.194	-31.386

On a d'une part

$$f_c = \frac{\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y})^2/(n-2)} = \frac{\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum_{i=1}^{n} e_i^2/(n-2)} = \frac{43477.3591/1}{4111.2071/(5-2)} = 31.7260,$$

et d'autre par

$$f_\alpha = f(1, n-2, 1-\alpha) = f(1,3,0.95) = 10.1.$$

On constate que $f_c > f_\alpha$, alors le modèle est valide (pertinent), c'est-à-dire on admet qu'on peut expliquer la Biomasse de la plante en fonction de la concentration de l'azote par la droite

$$\hat{Y} = 0.4329\ x\ +\ 336.6460.$$

5. On a : $\hat{Y} = 0.4329\ x\ +\ 336.6460$ alors la Biomasse qu'on peut prévoir à une concentration 500 μmol est $\hat{Y} = 0.4329 * 500\ +\ 336.6460 = 553.0960 mg$.

Solution 13 *(Régression linéaire simple)*

1. Par définition le coefficient de corrélation linéaire est donné par :

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y}.$$

on a : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{19690}{10} = 1969, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{20.3}{10} = 2.03,$

$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2} = \sqrt{\frac{1}{10}42925500 - 1969^2} = 679.5333,$

$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2} = \sqrt{\frac{1}{10}162.4100 - 2.03^2} = 3.6697,$

$Cov(x,y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{X}\ \overline{Y} = \left(\frac{1}{10}17671\right) - 1969 \times 2.03 = -2.22997,$

alors le coefficient de corrélation est :

$$r = -0.9936$$

2. Calculer les estimations des paramètres a, b et σ^2 pour la régression linéaire de Y sur X.
3. Le modèle linéaire de Y sur X est donné par :

$$Y = aX + b + \epsilon,$$

en utilisant la méthode des moindres carrées les estimateurs de a et b sont définis comme suite :

$$\boxed{\hat{a} = \frac{Cov(x,y)}{Var(x)} = -0.0054.} \text{ et } \boxed{\hat{b} = \overline{Y} - \hat{a}\overline{X} = 12.5953} \tag{3.9}$$

c'est-à-dire, la droite de régression est :

$$\hat{Y} = -0.0054X + 12.5953.$$

on a,

$$\hat{\sigma}_c^2 = var(\epsilon) = var(y - \hat{a} - \hat{b}x) = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2, \tag{3.10}$$

donc

$$\hat{\sigma}_c^2 = \frac{1}{10-2}\sum_{i=1}^{10}(y_i - \hat{a} - \hat{b}x_i)^2 = 0.1931.$$

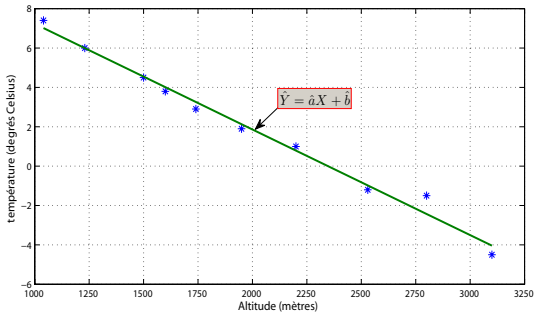

FIGURE 3.2: *Nuage des points observés et la droite de régression*

4. Les températures moyennes correspondantes à aux altitudes 1100 m et 2300 m.
   (a) $1100m$ est : $y = -0.0054 * 1100 + 12.5953 = 6.6929$.
   (b) $2300m$ est : $y = -0.0054 * 2300 + 12.5953 = 0.2539$.

Solution 14 On note X est le poids, Y est le Prix et le modèle de régression est $Y = aX + b$.

1. A partir des données on a :

variable	Moyenne	Variance	Carrée Moyenne
X	138.1667	1426.4722	20516.50
Y	166.6667	3222.2222	31000
$X * Y$	24643.33		

   d'où : $Cov(X, Y) = 1615.54$, $\rho = 0.754$, $\hat{a} = 1.133$, $\hat{b} = 10.186$ et $\hat{Y} = 1.133X + 10.186$.

2. Si on augmente le poids du Sandwich S_6 à 180 g, alors son nouveau prix sera :
   $\hat{Y} = 1.133(180) + 10.186 = 214.126DA$.

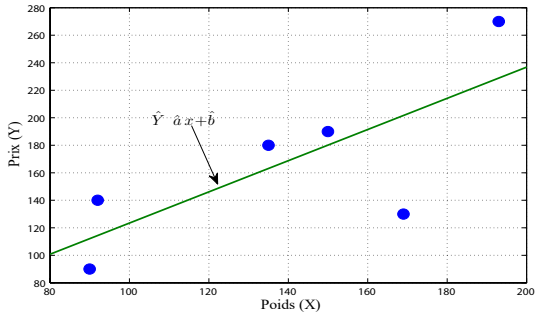

FIGURE 3.3: *Nuage de variation du Prix des Sandwichs en fonction de leurs Poids.*

3. La table d'analyse de la variance, du modèle, est donnée comme suite :

Source	SC	ddl	MC	f_c	Sig.
Régression	10978.215	1	10978.215	5.256	0.084
Résidu	8355.118	4	2088.780		
Total	19333.333	5			

TABLE 3.25: *Table d'ANOVA du modèle*

   A partir de ces résultats, on constate que pour un risque de 5%, le modèle linéaire n'est pas adéquat pour la description de la relation entre les variables Poids et Prix. Mais on peut conclure que ce modèle est adéquat pour un risque de 10%.

Solution 15 *(Régression linéaire simple et transformation des variables)*
Pour faire une régression linéaire, on effectue un changement de variable en posant $X = \ln(D)$ et $Y = \ln(H)$. Après le calcul des valeurs des variable X et Y on aura les résultats suivants :

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65
$\hat{Y}$	2.1690	2.3255	2.4133	2.5889	2.6232
ϵ	0.0510	-0.0555	-0.0333	0.0111	0.0268

1. Le calcul du coefficient de corrélation linéaire entre X et Y nécessite les quantités suivantes :
   $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = -0.9420$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2.4240$,

$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2} = \sqrt{0.1945} = 0.4410,$
$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2} = \sqrt{0.0299} = 0.1728,$
$Cov(x,y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{X}\ \overline{Y} = 0.0742,$
ainsi on aura le coefficient de corrélation :

$$r = 0.9737.$$

2. On a,
$$\hat{a} = \frac{Cov(x,y)}{Var(x)} = 0.3817 et \hat{b} = \overline{Y} - \hat{a}\overline{X} = 2.7836.$$
alors,
$$Y = 0.38172X + 2.78358, \tag{3.11}$$

3. Le test de validation du modèle se base sur la statistique :
$$F = \frac{\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y})^2/(n-2)} \rightsquigarrow f_{(1,n-2)},$$
or on a,
$$\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1 = 0.1417$$
et
$$\sum_{i=1}^{n}(y_i - \hat{y})^2/(n-2) = 0.0025$$
alors la réalisation, f, de la statistique F est égale à : 55.7266.
A partir de la table de Fisher pour un seuil de risque $\alpha = 5\%$, on obtient $f_{(1,n-2,1-\alpha)} = f_{(1,3,0.95)} = 10.1$, on constate que la valeurs de la réalisation de la statistique F est supérieur à la valeurs tabulée de fisher, cela signifier que le modèle est valide c'est-à-dire le modèle linéaire définie dans (3.11) est adéquat pour l'explication de la variable Y en fonction de la variable X.

4. Donner la hauteur prévue d'un arbre de diamètre 0.7. on a,
$$\hat{Y} = 0.38172X + 2.78358 \Rightarrow \ln(\hat{H}) = 0.38172\ln(D) + 2.78358 \Rightarrow \hat{H} = e^{0.38172\ln(D)+2.78358}.$$
Alors, pour un diamètre D=0.7, on prévoit une hauteur $H = e^{0.38172\ln(0.7)+2.78358} = 14.1177$.

Solution 16 *(Régression linéaire simple et changement des variables)*

							Somme
X $\mu g/\mu l$	0	20	40	60	80	100	300
Y	0	0.205	0.331	0.515	0.584	0.671	2.3060
X^2	0	400	1600	3600	6400	10000	22000
Y^2	0	0.0420	0.1096	0.2652	0.3411	0.4502	1.2081
$X * Y$	0	4.10	13.24	30.90	46.72	67.10	162.06

a) Afin de modéliser ces données, nous avons proposé le modèle linéaire suivant :

$$Y \;=\; a_1\, x \;+\; b_1.$$

1. Calcul des estimateurs des paramètres a_1 et b_1. On a :

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}300 = 50.\\
\overline{Y} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}3002.3060 = 0.3843.\\
Cov(x,y) &= \frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\;\overline{Y} = \frac{1}{6}(162.06) - (50)\;(0.3843) = 7.7950\\
Var(x) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 = \frac{1}{6}(22000) - (50)^2 = 1166.6667
\end{aligned}$$

alors,

$$\hat{a}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\frac{1}{n}\sum\limits_{i=n}^{n} x_i y_i - \overline{X}\;\overline{Y}}{\frac{1}{n}\sum\limits_{i=1}^{n} x_i^2 - \overline{X}^2} = 0.0067.$$

$$\hat{b}_1 = \overline{Y} - \hat{a}_1\overline{X} = 0.0503,$$

de ce fait la droite de régression de l'absorbance (Y) en fonction de la concentration (x) est donnée par :

$$\hat{Y} = 0.0067\; x \;+\; 0.0503.$$

2. Quelle absorbance prévoyez-vous à une concentration 50 $\mu g/\mu l$?

$$\hat{Y} = 0.0067\;(50) \;+\; 0.0503 \;=\; 0.3853.$$

3. Quelle absorbance prévoyez-vous à une concentration 40 $\mu g/\mu l$? Que peut-on conclure ?

$$\hat{Y} = 0.0067\;(40) \;+\; 0.0503 = 0.3183.$$

On constate que la valeur de régression est très proche de la vraie valeur (0.331), donc à priori le modèle retenu est adéquate pour la représentation des données du tableau.

4. Calcul du coefficient de corrélation linéaire.

$$r = r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} = 0.9851,$$

avec $\sigma_y = \sqrt{var(Y)} = \sqrt{\frac{1}{n}\sum\limits_{i=1}^{n} y_i^2 - \overline{Y}} = \sqrt{0.0536} = 0.2316$; La valeur du coefficient de corrélation est très proche de 1, i.e. X et Y sont fortement linéairement liés donc le modèle est efficace ce qui confirme les résultats de la question 3).

5. Pour un seuil de risque $\alpha = 5\%$, le modèle proposé est-il pertinent ?
Pour répondre à cette question on utilise le test de validation du modèle (Fisher). On d'une part

$$f_c = \frac{\sum\limits_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum\limits_{i=1}^{n}(y_i - \hat{y})^2/(n-2)} = \frac{0.3124/1}{0.0095/(6-2)} = 131.5368,$$

et d'autre par

$$f_\alpha = f(1, n-2, 1-\alpha) = f(1, 4, 0.95) = 7.71.$$

On constate que $f_c > f_\alpha$, alors on accepte le modèle proposé, c'est-à-dire le modèle est valide (pertinent)

b) Vue les doutes qu'on a sur le modèle précèdent, nous avons proposé le modèle suivant :

$$Z \;=\; e^Y \;=\; a_2\; x \;+\; b_2.$$

1. Complétez le tableau suivant :

							Somme
X $\mu g/\mu l$	0	20	40	60	80	100	300
Z	1.0000	1.2275	1.3924	1.6736	1.7932	1.9562	9.0429
Z^2	1.0000	1.5068	1.9387	2.8011	3.2156	3.8267	14.2888
$X * Z$	0	24.5505	55.6944	100.4183	143.4558	195.6193	519.7382

2. Calculer les estimations des paramètres a_2 et b_2 pour la régression linéaire de Z sur X.

$$\begin{aligned}
\overline{X} &= 50. \\
\overline{Z} &= \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{6}(9.0429) = 1.5072. \\
Cov(x,z) &= \frac{1}{n}\sum_{i=n}^{n} x_i z_i - \overline{X}\;\overline{Z} = \frac{1}{6}(162.06) - (50)\;(0.3843) = 7.7950 \\
Var(x) &= 1166.6667
\end{aligned}$$

alors,

$$\hat{a}_2 = \frac{Cov(x,z)}{Var(x)} = 0.0097.$$

$$\hat{b}_2 = \overline{Z} - \hat{a}_2\overline{X} = 1.0243,$$

de ce fait la droite de régression de (Z) en fonction de (x) est donnée par :

$$\hat{Z} = 0.0097\; x \;+\; 1.0243.$$

3. Quelle absorbance prévoyez-vous à une concentration 40 $\mu g/\mu l$. Que peut-on conclure par rapport au premier modèle ?
On $Z = 0.0097\;(40) \;+\; 1.0243 = 1.4123$ donc l'absorbance $y = log(z) = log(1.4123) = 0.3452$.
On constate que se modèle nous fournit une valeur proche à la vraie valeur mais c'est le premier modèle qui nous fournis une valeur plus proche d se fait il se peut que c'est le premier modèle qui est meilleur.
4. Calculer le coefficient de corrélation linéaire de ce nouveau modèle.

$$r = r(x,y) = \frac{Cov(x,z)}{\sigma_x \sigma_z} = 0.9946,$$

5. On constate que le coefficient de corrélation est plus grand pour le deuxième modèle donc le meilleur modèle est le deuxième.