UNIVERSITE DE BATNA_2 **FACULTE DE MEDECINE**

SERIE DE TD N° 11 *en BIOSTATISTIQUE 2019/2020*

Exercices sur l'analyse de la variance

Exercice 1 :

On veut savoir si l'addition de substances adjuvantes à un vaccin modifie la production d'anticorps. Pour cela, on mesure les quantités d'anticorps produites par des sujets après administration de quantités égales du vaccin, additionnées ou non d'une substance adjuvante. On obtient les taux dans le tableau suivant :

Sans substance adjuvante	Avec de l'alumine	Avec des phosphates
1, 3, 3, 0, 1	2, 4, 5, 4, 3, 6	1, 4, 2, 3, 3

a) Quelles hypothèses faut-il faire pour appliquer la technique de l'analyse de la variance à la résolution du problème posé ?
b) Sous les hypothèses adéquates, tester l'hypothèse selon laquelle les populations dont sont extraits les 3 échantillons ont la même variance.
c) En précisant toujours les hypothèses adéquates, l'efficacité du vaccin dépend-elle
   i) De la présence de substances adjuvantes ?
   ii) De leur nature ?

Solution :

a) Les échantillons doivent être indépendants et gaussiens.
b) Le test d'hypothèse selon lequel les populations dont sont extraits les 3 échantillons ont la même variance.

Test d'égalité des variances (On considère la plus grande et la plus petite variance) :

	Alumine	*Phosphate*
Moyenne	4	2,6
Variance	2	1,3
Observations	6	5
Degré de liberté	5	4
F	1,53846154	
P (F ≤ f) unilatéral	0,3486768	
F critique (unilatéral)	6,2560565	

Interprétation du test :

H_0 = {Les variances sont identiques} ; contre
H_1 = {Les variances ne sont pas identiques}.

Etant donné que la p-value calculée (0.3487) est supérieure au niveau de signification seuil α = 0.05, on ne peut pas rejeter l'hypothèse nulle H_0.

Conclusion : les variances sont égales ; et donc on peut procéder au test d'égalité des moyennes.

c) *Test d'égalité des moyennes* avec le tableau ANOVA à un facteur

- Substance adjuvante : facteur (variable indépendante qualitative) de 3 niveaux
- Quantité d'anticorps : variable expliquée quantitative dépendante.

Statistiques descriptives : **n = 16** $\bar{\bar{x}}$ **= 2.8125** **x.. = 45**

Variable	Obs.	Minimum	Maximum	Moyenne $\bar{x}_{i.}$	Somme **x.**	Ecart-type s	Variance s^2
Sans Adjuv	5	0	3	1,6	8	1,341640786	1.8
Alumine	6	2	6	4	24	1,414213562	2
Phosphate	5	1	4	2,6	13	1,140175425	1.3

Tableau de l'ANOVA d'un facteur :

Source	*SS*	*DDL*	*MS*	*F*	*P-Value*	*F critique*
Inter groupes	16,0375	2	8,01875	4,65374	0,0299	3,80556525
Intra groupes	22,4	13	1,723076923			
Total	38,4375	15				

Interprétation du test :

H_0 : les moyennes égales ; contre H_1 : Au moins l'une des trois est différente d'une autre.

Etant donné que la p-value calculée (2.99 %) est inférieure au niveau de signification seuil α = 0.05, on doit rejeter l'hypothèse nulle H_0.

Le risque de rejeter l'hypothèse nulle H_0 alors qu'elle est vraie est de 2,99 %.

Conclusion : les moyennes ne sont pas égales. Il y a lieu de considérer la 2ème partie de la question c) du problème pour tester s'il y a égalité des deux moyennes concernant les adjuvants (alumine et phosphate).

Utilisation manuelle des formules de l'analyse de la variance vues au cours :

$SCE_{fa} = \sum n_{i.}(\bar{X}_{i.} - \bar{\bar{x}})^2 = 5(1.6 - 2.8125)^2 + 6(4 - 2.8125)^2 + 5(2.6 - 2.8125)^2 =$ ***16 .0375***

$SCE_{fa} = \sum \frac{x_{i.}^2}{n_{i.}} - \frac{x_{..}^2}{N} = \frac{8^2}{5} + \frac{24^2}{6} + \frac{13^2}{5} - \frac{45^2}{16} = 142.6 - 126.5625 =$ ***16.0375***

$SCE_r = \sum_{i,j} x_{ij}^2 - \sum \frac{x_{i.}^2}{n_{i.}} = 5597 - (\frac{8^2}{5} + \frac{24^2}{6} + \frac{13^2}{5}) = 165 - 142.6 =$ ***22.4***

$SCE_t = \sum_{i,j}(x_{ij} - \bar{\bar{x}})^2 = \sum_{i,j} x_{ij}^2 - \frac{x_{..}^2}{N} = 165 - 126.5625 =$ ***38.4375***

La fameuse équation de l'analyse de la variance :

$SCE_t = SCE_{fa} + SCE_r = 16.0375 + 22.4 = 38.4375$

2ème partie question c)

Test de comparaison de deux moyennes : *Formulation des hypothèses* :

$H_0 = \{\mu_1 - \mu_2 = 0\}$ contre $H_1 = \{\mu_1 - \mu_2 \neq 0\}$ (Egalité contre inégalité)

Le seuil critique : $t_{(0.05\,;\,9)} = 2.2622$ si $To \in [-t_{(0.05\,;\,9)}\,;\,+t_{(0.05\,;\,9)}] \Rightarrow$ on ne peut pas rejeter H_0.

	Alumine	*Phosphate*
Moyenne	4	2,6
Variance	2	1,3
Observations	6	5
Variance pondérée	1,68888889	
Différence hypothétique des moyennes	0	
Degré de liberté	9	
Statistique de test observée t	1,77906486	
P(T ≤ t) unilatéral	0,05446847	
Valeur critique de t (unilatéral)	1,83311293	
P(T ≤ t) bilatéral	0,10893695	
Valeur critique de t (bilatéral)	2,26215716	

Interprétation du test :

H_0 : les moyennes égales ; contre H_1 : Les moyennes ne sont pas égales.

Etant donné que la p-value calculée (10.89 %) est supérieure au niveau de signification seuil $\alpha = 0.05$, on ne doit pas rejeter l'hypothèse nulle H_0.

Le risque de rejeter l'hypothèse nulle H_0 alors qu'elle est vraie est de 10,89 %.

Conclusion : les moyennes sont égales.

Exercice 2 **:**

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

Traitement A	Traitement B	Traitement C
26. 27. 35. 36. 38. 38. 41. 42 45. 50. 65	29. 42. 42. 44. 45. 48. 48. 52 56. 56. 58. 58. 60. 61. 63. 63. 69	26. 26. 30. 30. 33. 36. 38. 38 39. 46. 47. 51. 51. 56. 75

a) Tester l'égalité des variances.
b) Pouvons-nous conclure que les traitements ont une efficacité différente pour le critère « temps séparant une crise à la prochaine ».

Solution :

a) Après avoir calculé les variances de 3 échantillons, on teste l'égalité des deux variances (la plus grande et la plus petite en utilisant le test F) :

	Trait_C	*Trait_B*
Moyenne	41,46666667	52,58823529
Variance	174,4095238	103,0073529
Observations	15	17
Degré de liberté	14	16
F	1,693175476	
P ($F \leq f$) unilatéral	0,155668941	
Valeur critique pour F (unilatéral)	2,373318231	

Interprétation du test :

H_0 = {Les variances sont identiques} ; contre

H_1 = {Les variances ne sont pas identiques}.

Etant donné que la p-value calculée (0.155668941) est supérieure au niveau de signification seuil $\alpha = 0.05$, on ne peut pas rejeter l'hypothèse nulle H_0, donc il y a égalité des variances.

b) Analyse de la variance : un facteur

- Traitement : facteur à 3 niveaux (variable qualitative indépendante).
- Temps en jours (entre le début de l'étude et la prochaine crise) : variable quantitative expliquée dépendante.

Statistiques descriptives : **n = 43** $\overline{\overline{x}}$ **= 45.55813953** **x.. = 1959**

Groupes	Nb échant	Somme	Moyenne	Variance	Ecart-type
Trait_A	11	443	40,27272727	116,8181818	10.808246
Trait_B	17	894	52,58823529	103,0073529	10.149254
Trait_C	15	622	41,46666667	174,4095238	13.206420

ANOVA : un facteur

Source	*SS*	*DDL*	*MS*	*F*	*P-Valeur*	*F critique*
Inter groupes	1398,571853	2	699,285926	5,31975325	0,0089415	3,23172699
Intra groupes	5258,032799	40	131,45082			
Total	6656,604651	42				

Interprétation du test :

H_0 : moyennes égales ; contre H_1 : Au moins l'une des trois est différente d'une autre.

Etant donné que la p-value calculée est inférieure au niveau de signification seuil $\alpha = 0.05$, on doit rejeter l'hypothèse nulle H_0.

Conclusion : Il y a une différence significative entre les moyennes et que les traitements ont une efficacité différente.

Utilisation manuelle des formules de l'analyse de la variance vues au cours :

$SCE_{fa} = \sum n_{i.}(\bar{X}_{i.} - \bar{\bar{x}})^2 = 11(40.27273 - 45.55814)^2 + 17(52.58824 - 45.55814)^2 + 15(41.46667 - 45.55814)^2 =$ ***1398.57185***

$SCE_{fa} = \sum \frac{x_{i.}^2}{n_{i.}} - \frac{x_{..}^2}{N} = \frac{443^2}{11} + \frac{894^2}{17} + \frac{622^2}{15} - \frac{1959^2}{43} = 90\,646.9672 - 89\,248.39535 =$ ***1398.57185***

$SCE_r = \sum_{i,j} x_{ij}^2 - \sum \frac{x_{i.}^2}{n_{i.}} = 95\,905 - (\frac{443^2}{11} + \frac{770^2}{15} + \frac{491^2}{13}) = 95\,905 - 90\,646.9672 =$ ***5258.03280***

$SCE_t = \sum_{i,j}(x_{ij} - \bar{\bar{x}})^2 = \sum_{i,j} x_{ij}^2 - \frac{x_{..}^2}{N} = 95\,905 - 89\,248.39535 =$ ***6656.60465***

La fameuse équation de l'analyse de la variance :

$$\mathbf{SCE_t = SCE_{fa} + SCE_r = 1398.57185 + 5258.03280 = 6656.60465}$$

Exercice 3 :

On étudie l'activité d'un enzyme sérique, noté PDE, en fonction de différents facteurs dans l'espèce humaine. Les résultats sont exprimés en unités internationales par litres de sérum. On admettra que les populations considérées sont gaussiennes.

a) Chez deux groupes de femmes, enceintes et non enceintes, on obtient les résultats suivants

Enceintes 4.2 ; 5.5 ; 4.6 ; 5.4 ; 3.9 ; 5.4 ; 2.7 ; 3.9 ; 4.1 ; 4.1 ; 4.6 ; 3.9 ; 3.5

Non Enceintes 1.5 ; 1.6 ; 1.4 ; 2.9 ; 2.2 ; 1.8 ; 2.7 ; 1.9 ; 2.2 ; 2.8 ; 2.1 ; 1.8 ; 3.7 ; 1.8 ; 3.1

La grossesse a-t-elle une influence significative sur l'activité de la PDE ?

b) Afin d'évaluer la précocité de l'augmentation d'activité enzymatique lors de la grossesse, on pratique les dosages chez des femmes enceintes à différentes semaines d'aménorrhée. (On suppose que les conditions de validité du test sont satisfaites). On obtient sur des échantillons indépendants les résultats suivants :

4 semaines	5 semaines	6 semaines	7 semaines	8 semaines
7.2	4.9	10.4	4.6	6.1
4.3	4.8	4.6	5.6	11.4
5.5	4.7	8.4	8.3	8.2
4.5	5.4	6.1	6.9	5.7
4.7	4.7	8.1	4.5	6.6
5.5	4.7	5.4	4.7	6.6
6.6	6.2	6.7	6.7	6.3
5.3	5.6	7.5	4.8	5.9
5.4	3.2	6.4	5.0	5.8
3.9	6.1	5.6	5.0	4.8
5.5	6.7	6.3	5.3	9.1
2.7	5.5	7.7	7.8	13.2

L'âge de la grossesse a-t-il une influence sur l'activité de l'enzyme ?

Solution :

a) On doit tester l'égalité des variances

	Enceinte	*Non Enceint*
Moyenne	4,292307692	2,233333333
Variance	0,650769231	0,443809524
Observations	13	15
Degré de liberté	12	14
F	1,46632552	
P (F ≤ f) unilatéral	0,244925378	
Valeur critique pour F (unilatéral)	2,534243253	

Interprétation du test :

H_0 = {Les variances sont identiques} ; contre H_1 = {Les variances ne sont pas identiques}.

Etant donné que la p-value calculée (0.2449225378) est supérieure au niveau de signification seuil α = 0.05, on ne peut pas rejeter l'hypothèse nulle H_0.

Test de comparaison de deux moyennes les conditions de validité sont satisfaites.

On doit tester à présent l'égalité des moyennes sachant que les variances sont égales

	Enceintes	*Non Enceintes*
Moyenne	4,292307692	2,233333333
Variance s^2	0,650769231	0,443809524
Observations	13	15
Variance pondérée	0,539329389	
Différence hypothétique des moyennes	0	
Degré de liberté	26	
Statistique de test observé	7,398815288	
P(T ≤ t) unilatéral	3,70567E-08	
Valeur critique de t (unilatéral)	1,70561792	
P(T ≤ t) bilatéral	7,41133E-08	
Valeur critique de t (bilatéral)	2,055529439	

Interprétation du test :

H_0 : les moyennes égales ; contre H_1 : les moyennes ne sont pas égales.

Etant donné que la p-value calculée (presque nulle) est inférieure au niveau de signification seuil $\alpha = 0.05$, on ne peut pas accepter l'hypothèse nulle H_0.

Conclusion : Il y a une différence significative entre les moyennes ; c'est-à-dire que la grossesse a une influence hautement significative sur l'activité de la PDE.

b) Evaluation de la précocité de l'augmentation d'activité enzymatique lors de la grossesse : test sur les moyennes en utilisant ANOVA.

Rapport détaillé :

Groupes	*Obs*	*Somme*	*Moyenne*	*Variance*
4_semaine	12	61,1	5,091666667	1,420833333
5_semaine	12	62,5	5,208333333	0,849924242
6_semaine	12	83,2	6,933333333	2,495151515
7_semaine	12	69,2	5,766666667	1,742424242
8_semaine	12	89,7	7,475	6,522045455
Total	60	365.7	6.095	

Analyse de la variance : un facteur

- Âge de grossesse : facteur à 5 niveaux (variable qualitative indépendante).
- Dosage de l'enzyme : variable quantitative expliquée dépendante.

Source	*SS*	*DDL*	*MS*	*F*	*P-Valeur*	*F critique*
Inter groupes : SCEfa	54,0943333	4	13,52358333	5,189251807	0,001279507	2,53968863
Intra groupes : SCEr	143,334167	55	2,606075758			
Total : SCEt	197,4285	59				

Interprétation du test :

H_0 : les moyennes égales ; contre H_1 : les moyennes ne sont pas égales.

Etant donné que la p-value calculée (presque nulle = 0.13 %) est inférieure au niveau de signification seuil $\alpha = 0.05 = 5$ %, on ne peut pas accepter l'hypothèse nulle H_0.

Conclusion : Il y a une différence significative entre les moyennes c'est-à-dire qu'il y a une influence de la grossesse sur l'activité de l'enzyme.

Exercice 4 **:**

On a ensemencé des boites de Pétri avec des spores du genre Penicillium. Quatre milieux nutritifs ont été utilisés. Au bout d'un temps identique on a mesuré le diamètre des colonies (tableau ci-dessous). On admet que le diamètre des colonies suit une loi normale et on suppose que les cultures sont faites de manière indépendante. La nature du milieu nutritif a-t-il un effet sur la taille des colonies de spores ?

A	B	C	D
9.5	8	7.5	14
11.5	6.5	5	11.5
9	10	6	12
12	7	5.5	11
11.5	11.5	8.5	13

10 11	9.5 10.5 10	6.5 9	15

Solution :

RAPPORT DETAILLE :

Groupes	*échantillons*	*Somme*	*Moyenne*	*Variance*
A	7	74,5	10,6428571	1,30952381
B	8	73	9,125	3,125
C	7	48	6,85714286	2,30952381
D	6	76,5	12,75	2,375

Test sur l'égalité des variances : On considère la plus grande et la plus petite variance. Comparaison de ces deux variances par la distribution F.

H_0 = {variances égales} contre H_1 = {variances inégales}

$S^2_{max} = S^2_B = 3.125$ $S^2_{min} = S^2_A = 1.30952381$ $F_{max} = \frac{S^2_B}{S^2_A} = \frac{3.125}{1.3095} = 2.386$

$F_{(7\ ;\ 6\ ;\ 0.05)} = 4.215 \Rightarrow$ comme la statistique de test observée :

$To = F_{max} = 2.386 < F_{(7\ ;\ 6\ ;\ 0.05)} = 4.215 \Rightarrow$

On ne peut pas rejeter H_0 et donc on conclut que les variances sont égales.

Test de comparaison de 4 moyennes (ANOVA avec un facteur) **:**

- Milieu nutritif : Facteur (variable qualitative indépendante) à 4 niveaux.
- Taille des colonies : Variable quantitative expliquée dépendante.

Analyse de variance un facteur

Source	*SS*	*DDL*	*MS*	*F*	*P-Valeur*	*F critique*
Inter groupes	121,25	3	40,4166667	17,4887315	3,1032E-06	3,00878657
Intra groupes	55,4642857	24	2,3110119			
Total	176,714286	27				

Interprétation du test :

H_0 : les moyennes égales ; contre H_1 : les moyennes ne sont pas égales.

Décision : $\boldsymbol{T}o = 17.489 > \boldsymbol{F}_{(3;24;0.95)} = 3.009$

Etant donné que la p-value calculée (presque nulle = 0.000003 %) est inférieure au niveau de signification seuil $\alpha = 0.05 = 5$ %, on ne peut pas accepter l'hypothèse nulle H_0 ; alors au moins une moyenne est différente.

Conclusion : Il y a une différence significative entre les moyennes c'est-à-dire que le milieu nutritif a un effet sur le diamètre des colonies.