Exercices de biostatistique

Rappel: pour visualiser la formule associée aux résultats obtenus, il vous suffit d'aller cliquer sur la case concernée(uniquement dans excel et non avec "Adobe Acrobat") !!

Analyse de variance à deux critères (ANOVA 2)

Exercice 1

Dans trois fermes de vaches laitières, on a comparé la calcémie (en mgr %) des bêtes lactantes en fonction de l'époque et du niveau de production. On a obtenu les résultats suivants:

Niveau de production	**Epoque**				
	Février	**Avril**	**Juillet**	**Octobre**	**Décembre**
de 0 à 8 L.	**104**	**113**	**116**	**110**	**98**
de 8 à 15 L.	**104**	**115**	**117**	**115**	**94**
> 15 L.	**104**	**116**	**118**	**113**	**97**

Quelle est l'influence respective du facteur "époque" et du facteur "niveau de production" ?

H0: 1) Pas de différence entre les mois.
2) Pas de différence entre les niveux de production.

Février	**Avril**	**Juillet**	**Octobre**	**Décembre**	Somme	Moyenne
104	**113**	**116**	**110**	**98**	541	108,2
104	**115**	**117**	**115**	**94**	545	109
104	**116**	**118**	**113**	**97**	548	109,6
312	344	351	338	289	1634	
104	114,6666667	117	112,6666667	96,33333333		108,9333333

Somme des carrés totaux.

24,33777778	16,53777778	49,93777778	1,137777778	119,5377778
24,33777778	36,80444444	65,07111111	36,80444444	223,0044444
24,33777778	49,93777778	82,20444444	16,53777778	142,4044444

SCT= 912,9333333

Somme des carrés "époque".

24,33777778	32,87111111	65,07111111	13,93777778	158,76
24,33777778	32,87111111	65,07111111	13,93777778	158,76
24,33777778	32,87111111	65,07111111	13,93777778	158,76

SC"époque"= 884,9333333

Somme des carrés "niveau de production".

0,537777778	0,537777778	0,537777778	0,537777778	0,537777778
0,004444444	0,004444444	0,004444444	0,004444444	0,004444444
0,444444444	0,444444444	0,444444444	0,444444444	0,444444444

SC"niveau"= 4,933333333

Somme des carrés erreur.

SCE= 23,06666667

Table d'analyse de variance.

Source	SC	Ddl	Variance	F	P(>F)
Epoque	884,9333333	4	221,2333333	76,7283237	2,04009E-06
Niveau	4,933333333	2	2,466666667	0,855491329	0,460582289
Erreur	23,06666667	8	2,883333333		
Total	912,9333333	14			

Vu que seule la valeur du F pour "l'époque" est significativement différente (en effet, une telle valeur de F a une probabilité de 2,04009E-06), nous refusons l'hypothése nulle relative à l'époque; par contre, celle concernant le niveau de production est bel et bien acceptée (la valeur de F est probabble dans 46,0582289% des cas.

Exercice 2

On compare les effets de 5 régimes sur la croissance de rats, pendant les 4 semaines qui suivent le sevrage. On dispose de 8 nichées, chaque nichée comprenant autant de sujets qu'il y a de régimes à comparer. Existe-t-il des différences entre les régimes, entre les nichées?

	Traitements					
	A	B	C	D	E	Somme
Nichées						
I	57	64,8	70,7	68,3	76	336,8
II	55	66,6	59,4	67,1	74,5	322,6
III	62,1	69,5	64,5	69,1	76,5	341,7
IV	74,5	61,1	74	72,7	86,6	368,9
V	86,7	91,8	78,5	90,6	94,7	442,3
VI	42	51,8	55,8	44,3	43,2	237,1
VII	71,9	69,2	63	53,8	61,1	319
VIII	51,5	48,6	48,1	40,9	54,4	243,5
Somme	500,7	523,4	514	506,8	567	2611,9

H0: 1) Pas de différence entre les régimes.
2) Pas de différence entre les nichées.

Données					Somme	**Moyenne**
57	64,8	70,7	68,3	76	336,8	**67,36**
55	66,6	59,4	67,1	74,5	322,6	**64,52**
62,1	69,5	64,5	69,1	76,5	341,7	**68,34**
74,5	61,1	74	72,7	86,6	368,9	**73,78**
86,7	91,8	78,5	90,6	94,7	442,3	**88,46**
42	51,8	55,8	44,3	43,2	237,1	**47,42**
71,9	69,2	63	53,8	61,1	319	**63,8**
51,5	48,6	48,1	40,9	54,4	243,5	**48,7**
500,7	523,4	514	506,8	567	2611,9	
62,5875	**65,425**	**64,25**	**63,35**	**70,875**	**326,4875**	**65,2975**

Somme des carrés totaux.

68,84850625	0,24750625	29,18700625	9,01500625	114,5435063
106,0385063	1,69650625	34,78050625	3,24900625	84,68600625
10,22400625	17,66100625	0,63600625	14,45900625	125,4960063
84,68600625	17,61900625	75,73350625	54,79700625	453,7965063

458,0670063	702,3825063	174,3060063	640,2165063	864,5070063
542,7735062	182,1825063	90,20250625	440,8950062	488,2995062
43,59300625	15,22950625	5,27850625	132,1925063	17,61900625
190,3710063	278,8065062	295,7540062	595,2380062	118,7555063

SCT= 7584,06975

Somme des carrés "traitements".

7,3441	0,01625625	1,09725625	3,79275625	31,10850625
7,3441	0,01625625	1,09725625	3,79275625	31,10850625
7,3441	0,01625625	1,09725625	3,79275625	31,10850625
7,3441	0,01625625	1,09725625	3,79275625	31,10850625
7,3441	0,01625625	1,09725625	3,79275625	31,10850625
7,3441	0,01625625	1,09725625	3,79275625	31,10850625
7,3441	0,01625625	1,09725625	3,79275625	31,10850625
7,3441	0,01625625	1,09725625	3,79275625	31,10850625

SCTrait.= 346,871

Somme des carrés "nichées".

4,25390625	4,25390625	4,25390625	4,25390625	4,25390625
0,60450625	0,60450625	0,60450625	0,60450625	0,60450625
9,25680625	9,25680625	9,25680625	9,25680625	9,25680625
71,95280625	71,95280625	71,95280625	71,95280625	71,95280625
536,5014063	536,5014063	536,5014063	536,5014063	536,5014063
319,6050063	319,6050063	319,6050063	319,6050063	319,6050063
2,24250625	2,24250625	2,24250625	2,24250625	2,24250625
275,4770062	275,4770062	275,4770062	275,4770062	275,4770062

SCNichées= 6099,46975

Somme des carrés erreur(ou résiduelle).

SCE= 1137,729

Table d'analyse de la variance.

Source	SC	Ddl	Variances	F	P(>F)
Traitements	346,871	4	86,71775	2,134161123	0,102930811
Nichées	6099,46975	7	871,3528214	21,44436768	1,11957E-09
Erreur	1137,729	28	40,63317857		
Total	7584,06975	39			

Il n'ya donc pas de différence entre les traitements (une telle valeur de F est possible dans 10,2930811% des cas, avec 4 et 28 ddl), mais il y a une différence entre les nichées (seulement possible dans 1,11957E-07% des cas!).

Autre résolution par package

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	5	336,8	67,36	50,143
Row 2	5	322,6	64,52	56,857
Row 3	5	341,7	68,34	30,548

Row 4	5	368,9	73,78	81,717
Row 5	5	442,3	88,46	39,243
Row 6	5	237,1	47,42	36,582
Row 7	5	319	63,8	50,675
Row 8	5	243,5	48,7	25,385
Column 1	8	500,7	62,5875	206,5498214
Column 2	8	523,4	65,425	173,6707143
Column 3	8	514	64,25	99,58571429
Column 4	8	506,8	63,35	265,6742857
Column 5	8	567	70,875	288,405

ANOVA

Source of Varia	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	6099,46975	7	871,3528214	21,44436768	1,11957E-09	2,359257678
Columns	346,871	4	86,71775	2,134161123	0,102930811	2,714074299
Error	1137,729	28	40,63317857			
Total	7584,06975	39				

SC Colonnes = 8*((62,5875-65,2975)²+…) = 346,871
SC Rangées = 5*((67,36-65,2975)²+…) = 6099,46975
SC Erreur = (57+65,2975-67,36-62,5875)²+… = 1137,729

58,5225	7,22265625	19,25015625	8,33765625	9,37890625
46,3761	3,81225625	16,58525625	20,49825625	19,38200625
12,4609	1,06605625	7,79805625	7,33055625	6,66930625
11,7649	164,0320563	1,60655625	0,75255625	52,45380625
0,9025	10,32015625	79,43265625	16,70765625	0,43890625
7,3441	18,08375625	88,87775625	1,37475625	95,99100625
116,8561	27,79925625	0,06125625	64,84275625	68,51700625
30,3601	0,05175625	0,20025625	34,25175625	0,01500625

Erreur standard d'une moyenne traitement = déviation standard d'une moyenne dans le trait.
=> es = s/sqrt(r => e.s.(trait) = sqrt(var_e/n) = 2,253696369

Erreur standard d'une différence = somme de deux erreurs standard = 4,507392737

Interaction: l'effet d'un traitement n'est pas le même dans chaque nichée, et vice-versa

Exercice 3

Un expérimentateur, étudiant les valeurs normales de constantes sanguines chez le beagle, a dosé les protéines totales (en mgr/100ml) sur le sérum sanguin d'une part et sur le plasma d'autre part, d'un échantillon de 7 chiens, âgés d'un an et vivant dans les mêmes conditions de milieu et d'alimentation. Il a obtenu les résultats suivants:

Sérum	**Plasma**
5,8	6,1
5,6	5,9
6,1	6,3
6,3	6,4
6,1	6,3

7	7,7
6,2	6,5

Y-a-t'il des différences entre les chiens et la nature du prélèvement?

H0: 1) Pas de différence entre les chiens.
2) Pas de différence entre la nature du prélèvement.

Les données sont constituées de couples d'observations faites sur 7 beagles. Il s'agit clairement de données pairées. La première possibilité est donc d'employer un test de t pairé.

	Sérum	Plasma	d
Beagle 1	5,8	6,1	-0,3
Beagle 2	5,6	5,9	-0,3
Beagle 3	6,1	6,3	-0,2
Beagle 4	6,3	6,4	-0,1
Beagle 5	6,1	6,3	-0,2
Beagle 6	7	7,7	-0,7
Beagle 7	6,2	6,5	-0,3
Totaux	43,1	45,2	-2,1
Moyennes	6,157142857	6,457142857	-0,3
Dév. Std.	0,442933941	0,582686716	0,191485422

La déviation standard des moyennes de 7 différences est égale à la déviation standard des différences divisée par la racine de 7.

Soit, Sdbarre = 0,072374686

La statistique t = dbarre/Sdbarre vaut: t = -4,145095678
avec (7-1)=6 degrés de liberté

Dans la table de t, cette valeur est significative au seuil 1%.
En fait , on a que: P(t6 < -4,14509568) = 0,003022052

Comme le sens de la différence n'était pas postulé à priori, il faut utiliser un test bilatéral.
La probabilité est donc: 0,006044103

La seconde manière d'aborder le problème est de considérer que deux effets distincts affectent potentiellement les données: l'origine du prélèvement d'une part, et l'individu d'autre part.
Seul l'origine nous intéresse, mais l'effet individu peut être éliminé (effet de nuisance), ce qui augmente la puissance du test de comparaison simple des moyennes 'Sérum' et 'Plasma' (qui aurait été réalisé par un test de t non pairé ou par une Anova I).

	Sérum	Plasma	
Beagle 1	5,8	6,1	5,95
Beagle 2	5,6	5,9	5,75
Beagle 3	6,1	6,3	6,2
Beagle 4	6,3	6,4	6,35
Beagle 5	6,1	6,3	6,2
Beagle 6	7	7,7	7,35
Beagle 7	6,2	6,5	6,35
Moyennes	6,157142857	6,457142857	6,307142857

A) Contributions à la somme des carrés totaux

Les contributions proviennent des écarts entre les données individuelles et la moyenne générale élevés au carré, ce qui donne le tableau suivant:

0,257193878	0,042908163
0,50005102	0,165765306
0,042908163	5,10204E-05
5,10204E-05	0,008622449
0,042908163	5,10204E-05
0,48005102	1,94005102
0,011479592	0,037193878

la somme des carrés totaux est la somme de ces contributions, soit: 3,529285714

B) Contributions des moyennes 'Origine'

Ces contributions proviennent des écarts au carré des moyennes des origines par rapport à la moyenne générale. On obtient le tableau suivant:

0,0225	0,0225
0,0225	0,0225
0,0225	0,0225
0,0225	0,0225
0,0225	0,0225
0,0225	0,0225
0,0225	0,0225

la somme des carrés 'origine' est la somme de ces contributions, soit: 0,315

C) Contributions des moyennes 'Individu'

Ces contributions proviennent des écarts au carré des moyennes des individus par rapport à la moyenne générale. On obtient le tableau suivant:

0,12755102	0,12755102
0,310408163	0,310408163
0,011479592	0,011479592
0,001836735	0,001836735
0,011479592	0,011479592
1,08755102	1,08755102
0,001836735	0,001836735

la somme des carrés 'individu' est la somme de ces contributions: 3,104285714

D) Contributions de l'erreur

Ces contributions peuvent se calculer par (Yijk-Yi..-Y.j.+Y…)^2 ou la somme des carrés peut se calculer par différence (SCE=SCT-SCO-SCI)

0	0
7,88861E-31	0
0,0025	0,0025
0,01	0,01
0,0025	0,0025
0,04	0,04
0	0

la somme des carrés 'erreur' est la somme de ces contributions: 0,11

Vérification: SCO + SCI + SCE = SCT
3,529285714 3,529285714

Table d'analyse de la variance

Source	SC	DDL	Carrés moy.	F	P(>F)
Origine	0,315	1	0,315	17,18181818	0,006044103
Individu	3,104285714	6	0,517380952	28,22077922	0,000380505
Erreur	0,11	6	0,018333333		
Total	3,529285714	13	0,271483516		

La valeur qui est drectement d'intérêt est la valeur relative à l'origine. On vérifie que cette valeur a une probabilité de 0.006 (comme plus haut): il y a donc bien un effet origine.
On peut également vérifier que $t^2=F$.
Les deux hypothèses nulles sont donc rejetées, il y a des différences entre les chiens et la nature du prélèvement(sérum et plasma).

La remarque sur l'âge et les conditions similaires permettrait s'il était nécessaire de postuler que les individus proviennent d'une population homogène, avec en particulier des variances dans les groupes qui sont comparables. Dans le cas de données pairées, cette contrainte disparaît: il n'y a donc pas d'intérêt particulier, pour le test statistique, que les échantillons soient homosédastiques.

Autre résolution par package

t-Test: Two-Sample Assuming Equal Variances

	Sérum:	*Plasma:*
Mean	6,157142857	6,457142857
Variance	0,196190476	0,33952381
Observations	7	7
Pooled Varianc	0,267857143	
Hypothesized M	0	
df	12	
t Stat	-1,08443534	
P(T<=t) one-ta	0,149739389	
t Critical one-ta	1,782286745	
P(T<=t) two-tai	0,299478777	
t Critical two-ta	2,178812792	

t-Test: Paired Two Sample for Means

	Sérum:	*Plasma:*
Mean	6,157142857	6,457142857
Variance	0,196190476	0,33952381
Observations	7	7
Pearson Corre	0,966802832	
Hypothesized M	0	
df	6	
t Stat	-4,14509568	
P(T<=t) one-ta	0,003022052	
t Critical one-ta	1,943180905	
P(T<=t) two-tai	0,006044103	
t Critical two-ta	2,446913641	

F-Test Two-Sample for Variances

	Sérum:	*Plasma:*
Mean	6,157142857	6,457142857
Variance	0,196190476	0,33952381
Observations	7	7
df	6	6
F	0,577840112	
P(F<=f) one-ta	0,260880351	
F Critical one-t	0,233434605	

Exercice 4

Un groupe de chatons subit des tests à 1 mois et à 6 mois pour mesurer leur agressivité.
Les données sont les suivantes:

	1 mois	**6 mois**
Chat 1	12	10
Chat 2	18	18
Chat 3	20	22
Chat 4	16	20

Montrez s'il y a des différences significatives entre les mesures aux différentes âges et entre les chats?

H0: 1) Pas de différence entre les mesures aux différents âges.
2) Pas de différence entre les chats.

	1 mois	**6 mois**	Somme	**Moyenne**
Chat 1	12	10	22	11
Chat 2	18	18	36	18
Chat 3	20	22	42	21
Chat 4	16	20	36	18
Somme	66	70		
Moyenne	16,5	17,5		17

Somme des carrés totaux.

25	49
1	1
9	25
1	9

SCT= 120

Somme des carrés "mesures aux différents âges".

0,25	0,25
0,25	0,25
0,25	0,25
0,25	0,25

SCMesures= 2

Somme des carrés "chats".

36	36
1	1
16	16
1	1

SCChiens= 108

Somme des carrés erreur.

SCE= 10

Table d'analyse de variance.

Source	SC	Ddl	Variances	F	P(>F)
Chats	108	3	36	10,8	0,040800387
Mesures	2	1	2	0,6	0,495025346
Erreur	10	3	3,333333333		
Total	120	7			

L'hypothèse nulle est donc acceptée au seuil 5 % pour les mesures et rejetée pour les chats.

Exercice 5

On a comparé le temps de coagulation (en minutes) du sang de 8 individus traités par 4 anticoagulants différents. Les échantillons ont été soumis à ces quatre traitements dans un ordre au hasard. Les données sont les suivantes:

	Traitements			
Individus	1	2	3	4
1	8,4	9,4	9,8	12,2
2	12,8	15,2	12,9	14,4
3	9,6	9,1	11,2	9,8
4	9,8	8,8	9,9	12
5	8,4	8,2	8,5	8,5
6	8,6	9,9	9,8	10,9
7	8,9	9	9,2	10,4
8	7,9	8,1	8,2	10

Quelles sont les influences respectives des facteurs "traitement" et individu" ? Comparez les traitements 1 et 4 par un test de t et en utilisant l'erreur standard calculée dans l'analyse de variance.

Encore une fois, le traitement simultané des effets 'Traitement' et 'Individu' va permettre d'augmenter la puissance du test qui nous intéresse ('Traitement'). On peut à nouveau modéliser ces données par une anova II:

$$Yij = \mu + Ti + Ij + \varepsilon ij$$

L'effet traitement i se mesure par l'écart entre la moyenne du traitement i et la moyenne globale.
L'effet individu j se mesure par l'écart entre la moyenne de l'individu j et la moyenne globale.

	Traitements

Individus	1	2	3	4	Moyennes
1	8,4	9,4	9,8	12,2	9,95
2	12,8	15,2	12,9	14,4	13,825
3	9,6	9,1	11,2	9,8	9,925
4	9,8	8,8	9,9	12	10,125
5	8,4	8,2	8,5	8,5	8,4
6	8,6	9,9	9,8	10,9	9,8
7	8,9	9	9,2	10,4	9,375
8	7,9	8,1	8,2	10	8,55
Moyennes	9,3	9,7125	9,9375	11,025	9,99375

On peut donc procéder à nouveau en calculant les différentes contributions.

A) Effet 'Traitement'

	Traitements			
Individus	1	2	3	4
1	0,481289062	0,079101563	0,003164062	1,063476563
2	0,481289062	0,079101563	0,003164062	1,063476563
3	0,481289062	0,079101563	0,003164062	1,063476563
4	0,481289062	0,079101563	0,003164062	1,063476563
5	0,481289062	0,079101563	0,003164062	1,063476563
6	0,481289062	0,079101563	0,003164062	1,063476563
7	0,481289062	0,079101563	0,003164062	1,063476563
8	0,481289062	0,079101563	0,003164062	1,063476563

La somme des contributions (Ti^2) vaut donc: 13,01625

B) Effet 'Individu'

	Traitements			
Individus	1	2	3	4
1	0,001914063	0,001914063	0,001914063	0,001914063
2	14,67847656	14,67847656	14,67847656	14,67847656
3	0,004726562	0,004726562	0,004726562	0,004726562
4	0,017226562	0,017226562	0,017226562	0,017226562
5	2,540039063	2,540039063	2,540039063	2,540039063
6	0,037539062	0,037539062	0,037539062	0,037539062
7	0,382851563	0,382851563	0,382851563	0,382851563
8	2,084414063	2,084414063	2,084414063	2,084414063

La somme des contributions (Ii^2) vaut donc: 78,98875

C) Erreur

	Traitements			
Individus	1	2	3	4
1	0,733164063	0,072226562	0,008789062	1,485351563
2	0,109726562	2,743164063	0,754726563	0,208164063
3	0,135976562	0,295664062	1,772226563	1,336914063
4	0,135976563	1,089414062	0,028476563	0,711914062
5	0,481289062	0,006601563	0,024414062	0,867226563
6	0,256289063	0,145351563	0,003164063	0,004726562
7	0,047851562	0,008789063	0,014101563	3,90625E-05
8	0,001914062	0,028476562	0,086289063	0,175351562

La somme des contributions (Ii^2) vaut donc: 13,77375

D) La somme des carrés totaux

	Traitements			
Individus	1	2	3	4
1	2,540039063	0,352539063	0,037539062	4,867539063
2	7,875039063	27,10503906	8,446289063	19,41503906
3	0,155039063	0,798789063	1,455039063	0,037539062
4	0,037539062	1,425039063	0,008789063	4,025039063
5	2,540039063	3,217539063	2,231289063	2,231289063
6	1,942539063	0,008789063	0,037539062	0,821289063
7	1,196289063	0,987539063	0,630039063	0,165039063
8	4,383789063	3,586289063	3,217539063	3,90625E-05

La somme des contributions (Ii^2) vaut donc: 105,77875

L'information est synthétisée dans la table:

Source	SC	DDL	Carrés moy.	F	P(>F)
Traitements	13,01625	3	4,33875	6,615028587	0,002549665
Individu	78,98875	7	11,28410714	17,20419276	2,1968E-07
Erreur	13,77375	21	0,655892857		
Total	105,77875	31	3,412217742		

Vérifications:

SCT = SCTr + SCInd + SCE
DdlT = DdlTr + DdlInd + DdlE

La conclusion est que l'effet (de nuisance) Individu est très significatif, et qu'il est par conséquent conseillé de supprimer cet effet avant de procéder à l'analyse, ce qui justifie à postériori l'utilisation de données pairées. L'effet traitement est également très significatif globalement, signifiant que **certains** traitements sont statistiquement différents de **certains** autres.
A remarquer que l'analyse de variance, si elle permet de voir s'il y a des différences, ne précise pas lesquelles.

Pour comparer deux moyennes par un test de t, on utilise la différence de moyennes d'échantillons au numérateur et la variance de la différence (c'est à dire la somme des variances) de ces moyennes au dénominateur, ce qui aboutit au test de t classique. La variance à laquelle il est fait référence est estimée par la variance erreur de l'ANOVA. L'effet de corriger pour les effets de nuisance est de réduire cette variance, ce qui a un effet bénéfique sur la puissance du test.
On peut donc comparer les traitements 1 et 4 par ce canal:

A) Différence de moyennes = 1,725
B) Variance de la différence = 2*Variance erreur de l'anova
=> Variance de la différence de moyennes = 2*Variance erreur de l'Anova/8
=> Erreur standard de la différence = 0,404936062
C) Degrés de liberté = ddl erreur dans l'Anova = 21

Par conséquent, le test à utiliser est: t(21) = 4,259931782
P(>t) = 0,000174491
Un test bilatéral nécessite de doubler cette valeur, soit P = 0,000348981

La différence est donc très significative.

Une autre manière de procéder serait de calculer la somme des carrés due à la différence, soit

SC(T4-T1) = 8*(Xb1-Xb4)^2 =	23,805

Le nombre de degrés de liberté associés à cette différence est 1.

Par ailleurs, la variance de la différence est égale à 2 fois la variance erreur, soit:

VARE(T4-T1) = 2*VARE =	1,311785714

Le nombre de degrés de liberté reste 21.

Par conséquent, le rapport entre ces deux résultats permet de tester (avec un test de F) la différence des moyennes des traitements 1 et 4.

On obtient:	$F = t^2$ =	18,14701879