UNIVERSITE GASTON BERGER
L'excellence au service du développement

UFR des Lettres et Sciences humaines
Section de sociologie

MIASS 241
Mathématiques (appliquées aux sciences sociales) 4

SEPT EXERCICES RÉCAPITULATIFS- ANOVA (Solutionnaire)

1) Comparez le test *t* et le test ANOVA en termes de similitude et de différence ?

a) Au moins deux éléments de similitude

1. Le test ***t*** et le test ANOVA sont des **tests de comparaison de moyennes**. Ils consistent à comparer des moyennes de groupes d'observations afin de savoir si leurs différences sont significatives au niveau de la population, c'est-à-dire si elles sont suffisamment importantes pour ne pas être dues à une erreur d'échantillonnage. L'échantillon doit donc être de grande taille et de type probabiliste.

2. Le test ***t*** et le test ANOVA sont des **tests paramétriques**. Non seulement ils portent sur une variable dépendante quantitative, mais leur utilisation requiert une distribution paramétrée de la VD (selon la moyenne et l'écart-type) et donc la satisfaction d'hypothèses relatives à la normalité et l'égalité des variances dans N.

b) Au moins deux éléments de différence

1. Contrairement au test ***t***, le test ANOVA ne compare pas seulement les moyennes des groupes (différences ou variation intergroupes). Il va plus loin en les comparant aux scores individuels à l'intérieur de chaque groupe (variation intra-groupe).

2. Le test ANOVA est un **test unilatéral droite** ou **supérieur**, les valeurs allant de 0 à plus l'infini du fait qu'il repose sur la variance (variance intergroupes/variance intra-groupe). Par contre, le test ***t*** peut être **bilatéral** (les deux côtés de la distribution sont considérés) ou **unilatéral** (inférieur ou supérieur). En fait, de trois choses l'une : soit on considère la moyenne 1 comme différente de la moyenne 2, soit la moyenne 1 est inférieure à la moyenne 2, soit la moyenne 1 est supérieure à la moyenne 2.

2) Quel est le lien qui existe entre le tableau des moyennes et le test ANOVA?

Le test ***ANOVA*** approfondit l'analyse du tableau des moyennes en aidant à établir dans quelle mesure ***au moins une différence*** dans les moyennes des groupes est statistiquement significative au niveau de la population.

3) Un chercheur étudie la relation entre le temps passé à pratiquer du sport (variable indépendante) et l'indice de masse corporelle (variable

dépendante). Il veut utiliser le test d'hypothèse *ANOVA*. Précisez la nature et les valeurs possibles des variables (indépendante et dépendante) que le chercheur doit obtenir pour que le test en question soit approprié.

La variable indépendante, « temps de pratique de sport », doit être de nature **qualitative non dichotomique** et ses valeurs possibles sont : **temps faible**, **temps moyen**, **temps élevé**.

La variable dépendante, « indice de masse corporelle », doit être de nature **quantitative** et ses valeurs possibles sont : **15, 16, 17, 18, 19 kg/m^2,** etc.

NB : **Lorsque deux variables sont quantitatives, il faut catégoriser (« trichotomiser » par exemple) les valeurs de la variable indépendante pour que la relation se prête au test ANOVA** ☺

4) **Afin d'asseoir sa position, un consultant dans une organisation internationale veut montrer que l'indice du développement humain (échelle de 0 à 1) varie selon la région économique (Afrique, Europe, Amérique latine). Il décide alors de regrouper les valeurs de la variable dépendante quantitative, « indice du développement humain», en trois catégories (faible, moyen, élevé) dans la perspective d'utiliser le test du chi-carré.**

En tant qu'assistant de recherche, vous suggérez au consultant de ne pas transformer les valeurs de la variable dépendante (IDH) en catégories, car vous estimez qu'un test de différence de trois moyennes (analyse de variance) serait plus intéressant comme choix. Donnez au moins deux raisons qui militent en faveur d'un tel choix, tout en prenant soin de bien les expliciter ?

D'une part, si le consultant regroupe les scores de la variable « Indice du développement humain » en catégories, il **perd de l'information**. Non seulement il ne retrouvera plus les différences de scores entre les pays, mais il regroupera dans une même catégorie des scores qui sont pourtant différents les uns des autres. Par conséquent, le risque est grand que la catégorisation ***change complètement la structure de la relation*** entre la région économique et le développement humain. Il se peut que cette relation soit tout simplement éclipsée à cause de la catégorisation.

D'autre part, si le consultant regroupe les scores de la variable « Indice du développement humain » en catégories, il en **perd les propriétés mathématiques.** En effet, la catégorisation réduit le niveau de mesure de la variable : d'échelle de mesure d'intervalles/ratio, elle se transforme en échelle de mesure ordinale. Or, les valeurs métriques permettent de calculer la moyenne, la variance, l'écart-type, qui sont des mesures beaucoup plus intéressantes que les fréquences en statistiques, inférentielles notamment. Le chercheur ***perd ainsi en puissance statistique***.

Pour éviter ces inconvénients, il est approprié de conserver les valeurs métriques de l'IDH en vue d'effectuer un test ANOVA, lequel comparerait les indices moyens de développement humain des pays africains, européens et latino-américains.

NB : **Lorsqu'on décide de catégoriser les scores d'une variable afin d'adapter l'analyse à des techniques statistiques non paramétriques comme le chi-carré, on doit être conscient des inconvénients qu'implique un tel choix** ☺

5) **On s'intéresse à la relation entre le groupe d'appartenance (VI) et la**

performance à un test (VD). L'analyse de variance à un facteur fixe, à l'aide d'Excel, donne les résultats ci-dessous :

Analyse de variance: un facteur

RAPPORT DÉTAILLÉ

Groupes	*Nombre d'échantillons*	*Somme*	*Moyenne*	*Variance*
Groupe 1	20	66	3,3	2,12
Groupe2	20	52	2,6	5,52
Groupe 3	20	76	3,8	1,64

ANALYSE DE VARIANCE

Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*	*F*	*Valeur critique pour F*
Entre Groupes	14,53	2	7,27	2,35	3,16
A l'intérieur des groupes	176,20	57	3,09		
Total	190,73	59			

a) Quel est le facteur ici ?

Le facteur renvoie au « groupe d'appartenance », soit la variable explicative.

b) Quels sont les niveaux du facteur ?

Les niveaux du « groupe d'appartenance » sont : groupe 1, groupe 2, groupe 3.

c) Sommes-nous en présence d'un devis balancé ou non balancé ? Justifiez.

Nous sommes en présence d'un devis balancé, chacun des trois groupes ayant le même nombre de cas, soit 20.

d) Quelle est la taille de l'échantillon ?

L'échantillon est constitué de 60 cas (20+20+20).

e) Calculez la moyenne totale

Moyenne totale=3,23, soit (3,3+2,6+3,8)/3.

f) Indiquez les sources de la variation dans la performance au test et explicitez.

Il y a trois sources de variation dans la performance au test :

-Variation à l'intérieur de chacun des trois groupes (**variation intra-groupe**) : la performance varie d'un cas à l'autre à l'intérieur d'un même groupe.

-Variation entre les trois groupes (**variation intergroupes**) : la performance moyenne varie d'un groupe à l'autre.

-Variation dans l'ensemble de l'échantillon (**variation totale**) : la performance varie d'un cas à l'autre dans l'ensemble de l'échantillon, peu importe l'appartenance à un groupe.

g) Précisez la somme des carrés intra-groupe et la somme des carrés intergroupes. Qu'est-ce qu'elles mesurent précisément?

La somme des carrés intra-groupe est égale à 176,20. Elle mesure la déviation (écart), par rapport à la moyenne de leur groupe d'appartenance respectif, des scores individuels relatifs à la performance au test. Précisément, elle mesure la variation des scores des cas du groupe 1 par rapport à la moyenne de leur groupe, la variation des scores des cas du groupe 2 par rapport à la moyenne de leur groupe, la variation des scores des cas du groupe 3 par rapport à la moyenne de leur groupe.

La somme des carrés intergroupes est égale à 14,53. Elle mesure la déviation (écart), par rapport à la moyenne totale, des moyennes de groupe concernant la performance au test. Précisément, elle mesure la variation de la moyenne du groupe 1 par rapport à la moyenne totale, la variation de la moyenne du groupe 2 par rapport à la moyenne totale, la variation de la moyenne du groupe 3 par rapport à la moyenne totale.

h) Assurez-vous que l'addition de la somme des carrés intergroupes à la somme des carrés intra-groupes est égale à la somme des carrés totale. Que mesure-t-elle la somme des carrés totale ?

La somme des carrés totale est égale à 190,73, soit la somme des carrés intra-groupe (176,20) additionnée à la somme des carrés intergroupes (14,53). Elle mesure la variation, par rapport à la moyenne totale de l'échantillon, des scores individuels relatifs à la performance au test.

i) Précisez la moyenne des carrés intergroupes et la moyenne des carrés intra-groupe. Comment on les retrouve ? Qu'est-ce qu'elles mesurent ?

La moyenne des carrés intergroupes est de 7,27, soit la somme des carrés intergroupes relativisée par les degrés de liberté correspondants (14,53/2). Elle mesure la variance entre les trois groupes quant à la performance au test.

La moyenne des carrés intra-groupe est de 3,09, soit la somme des carrés intra-groupes divisée par les degrés de liberté correspondants (176,20/57). Elle mesure la variance à l'intérieur des trois groupes quant à la performance au test.

j) Précisez la valeur du F calculé. Comment on la retrouve ? Qu'est-ce qu'elle mesure ?

La valeur du F calculé est de 2,35, soit la variance intergroupes divisée par la variance intra-groupe (7,27/3,09). Elle mesure la variance entre les trois groupes quant à la performance au test, relativisée par la variance à l'intérieur des groupes.

6) Vous souhaitez étudier les dépenses consacrées à la discothèque par mois chez les étudiants en fonction du revenu en classes du père. Le tableau suivant présente les résultats obtenus à partir d'un sondage (supposément aléatoire) effectué auprès des étudiants du SOL1020 de 1986 à 2011:

Dépense à la discothèque par mois

Revenu	Moyenne	N	Ecart-type
Faible	18,51	385	30,348
Moyen	27,35	683	48,651
Élevé	24,25	1134	47,191

Total	24,21	2202	45,268

a) Analysez très brièvement le tableau des moyennes

Le tableau des moyennes montre que les 683 étudiants dont le statut socioéconomique du père est moyen dépensent plus par mois pour la discothèque (27,35$) que les 1134 étudiants dont le père a un statut socioéconomique élevé (24,25$) ou les 385 étudiants dont le statut socioéconomique du père est faible (18,51). Mais, existe-t-elle au moins une différence suffisamment significative pour ne pas être due à une erreur d'échantillonnage?

b) Calculez la somme des carrés totale (SC*totale*) en transposant la formule ci-dessous.

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1} = \frac{SCtotale}{n-1}$$

Si par définition :

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1} = \frac{SCtotale}{n-1}$$

Donc, en transposant :

$$SCtotale = s^2 * (n-1)$$
$$= 45{,}268^2 * 2201 = 4510271$$

c) Calculez la somme des carrés inter-groupes à l'aide de la formule ci-dessous.

$$SC\text{inter} = \sum N_G (\bar{x}_G - \bar{x}_T)^2$$

$$SC\text{inter} = \sum N_G (\bar{x}_G - \bar{x}_T)^2$$
$$= 385(18{,}51 - 24{,}21)^2 + 683(27{,}35 - 24{,}21)^2 + 1134(24{,}25 - 24{,}21)^2$$
$$= 12508{,}6 + 6734{,}1 + 1{,}81 = 19244{,}6$$

d) Calculez la somme des carrés intra-groupe connaissant SCtotale et SCinter

Puisque

$$SCtotale = SC\text{intra} + SC\text{inter}$$

On a

$$SC\text{intra} = SCtotale - SC\text{inter}$$
$$= 4510271 - 19244{,}6 = 4491026{,}4$$

e) Calculez la statistique F

Variance intergroupes= SCinter/k-1 = 19244,6/3-1 = 9622,3

Variance intra-groupe= SCintra/n-k = 4491026,4/2202-3 = 2042,3

$$F = \frac{\text{Variance intergroupes}}{\text{Variance intra - groupe}} = \frac{9622{,}3}{2042{,}3} = 4{,}71$$

f) Calculez les degrés de liberté pour les variations intergroupes et intra-groupe et déterminez la valeur critique du F au seuil de signification de 0,05

Pour la variation intergroupes, le nombre de dl est de 2 (k-1= 3-1)

Pour la variation intra-groupe, le nombre de dl est de 2199 (n-k= 2202-3)

Avec 2 et 2199 degrés de liberté, la valeur critique de F est de 3,00 au seuil de signification 0,05.

g) Présentez correctement le tableau ANOVA

Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés ou Variance*	*F*	*Valeur critique pour F*
Intergroupes	19244,6	2	9622,3	**4,71**	3,00
Intra-groupe	4491026,4	2199	2042,3		
Total	4510271	2201			

h) Prenez une décision et concluez sur la relation entre la classe sociale et les dépenses consacrées à la discothèque chez les étudiants.

Décision : La valeur calculée du F (4,71) étant supérieure à la valeur critique du F (3,00), avec 2 et 2199 dl et au seuil de 0,05, on rejette l'hypothèse nulle d'une absence de différence entre les dépenses moyennes.

Conclusion : Par conséquent, il existe une relation statistiquement significative entre la classe sociale du père et les dépenses dans les discothèques chez les étudiants. On est sûr au moins à 95% qu'il y a au moins une différence dans les dépenses qui n'est probablement pas due à une erreur d'échantillonnage.

i) Calculez la valeur de l'Êta-carré et interprétez-la.

$$E^2 = \frac{SC\text{inter}}{SC\text{totale}}$$

$$= \frac{19244{,}6}{4510271} = 0{,}0004$$

La valeur de l'êta-carré étant de 0,0004, on peut affirmer que l'effet de la classe sociale du père sur les dépenses moyennes dans les discothèques chez les étudiants est de faible taille, d'après les balises de Cohen (1988). La classe sociale n'explique que 0,04% de la variation dans les dépenses.

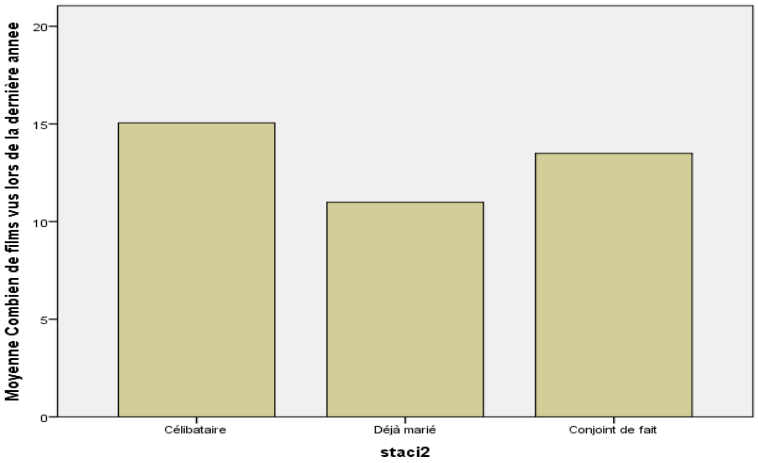
7) Vous souhaitez étudier le nombre de films visionnés au cours de la dernière année (VD) en fonction du statut matrimonial (VI) chez les étudiants du SOL1020 de 1986 à 2011. Pour y parvenir, sur la base des résultats SPSS, répondez aux questions suivantes :

a) Analysez globalement les différences entre les moyennes des trois groupes à l'aide du tableau des moyennes et des diagrammes. Les différences vous paraissent-elles importantes?

Comme le montrent le tableau des moyennes et les diagrammes, les 1872 célibataires ont regardé plus de films au cours de la dernière année (15,05 films) que les 188 étudiants ayant des conjoints de fait (13,49 films) ou les 181 étudiants déjà mariés (10,99 films). Les différences nous paraissent importantes. Mais, malgré la présence d'une pluralité de valeurs aberrantes, existe-t-elle au moins une différence assez significative pour ne pas être due à une erreur d'échantillonnage?

Combien de films vus lors de la dernière annee

Statciv_2	Moyenne	N	Ecart-type
Célibataire	15,05	1872	15,192
Déjà marié	10,99	181	14,425
Conjoint de fait	13,49	188	12,680
Total	14,59	2241	14,975



b) Testez s'il y a une relation statistique significative en

- formulant les hypothèses nulle et alternative,

Hypothèse nulle : Il n'y a pas de différence dans le nombre moyen de films vus au cours de la dernière année chez les étudiants célibataires, déjà mariés ou conjoints de fait. $H_0 : \mu_1 = \mu_2 = \mu_3$

Hypothèse alternative : Il y a au moins une différence dans le nombre moyen de films vus au cours de la dernière année chez les étudiants célibataires, déjà mariés ou conjoints de fait. H_a : Au moins une des moyennes μ_1, μ_2, μ_3 est différente des autres.

- prenant une décision d'acceptation ou de rejet de l'hypothèse nulle sur la base de la comparaison des valeurs de F (α= 0,01),

D'une part, la valeur calculée du F (6,663) est supérieure à la valeur critique du F (4,61), avec 2 et 2238 dl et au seuil de 0,01.

Donc, on rejette l'hypothèse nulle pour accepter l'hypothèse alternative.

- dégageant une conclusion sur la relation étudiée.

Par conséquent, il existe une relation statistiquement significative entre le statut matrimonial et le nombre moyen de films vus par les étudiants au cours de la dernière année. On est sûr au moins à 99% qu'il y a au moins une différence dans le nombre de films visionnés qui n'est probablement pas due à une erreur d'échantillonnage.

Tableau ANOVA

			Somme des carrés	df	Moyenne des carrés	F	
Combien de films vus lors de la dernière annee * staci2	Inter-groupes	Combiné	2973,188	2	1486,594	6,663	
	Intra-classe		499334,034	2238	223,116		
	Total		502307,222	2240			

c) Quelle est la taille de l'effet du statut civil sur le nombre de films visionnés. Exprimé autrement, quelle est la proportion de la variation expliquée?

La valeur de l'êta-carré étant de 0,006, on peut affirmer que l'effet du statut matrimonial sur le nombre moyen de films visionnés chez les étudiants est de taille faible, d'après les balises de Cohen (1988). Le statut matrimonial explique seulement 0,6% de la variation dans le nombre moyen de films visionnés.

Mesures des associations

	Eta	Eta carré
Combien de films vus lors de la dernière annee * staci2	,077	,006