



MIASS 241
Mathématiques (appliquées aux sciences sociales) 4
© El Hadj Touré, 2022

SEPT EXERCICES RÉCAPITULATIFS- Test t
(Solutionnaire)

1) Comparez le test du chi-carré et le test t en termes de similitude et de différence.

a) Au moins deux éléments de similitude

1. Le test du chi-carré et le test t sont tous des **tests d'hypothèses**. Ils consistent à formuler des hypothèses statistiques se rapportant à une population (paramètres) et à tenter de les valider à l'aide des données d'échantillon (statistiques).

2. Comme le test du chi-carré, le test t consiste concrètement à comparer les catégories de la variable indépendance, en termes de **différences**, eu égard à la variable dépendante. Avant d'utiliser le chi-carré ou le test t , il faut donc s'assurer qu'il y ait des différences, si minimales soient-elles, dans les données de l'échantillon (fréquences ou moyennes respectivement). L'utilisation des techniques tabulaires et graphiques peut aider à observer de telles différences dont on cherche ensuite à savoir 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.

NB : Le tableau des moyennes et le diagramme des moyennes sont au test t ce que le tableau bivarié et le diagramme en bâtons bivariés sont au test du chi-carré. Ces techniques tabulaires et graphiques révèlent des différences (n) que l'on tente de généraliser (N) à l'aide des tests d'hypothèses. 😊

b) Au moins deux éléments de différence

1. Contrairement au test du chi-carré, le test t est un test **paramétrique**. Non seulement il porte sur une *variable dépendante quantitative*, mais son utilisation requiert une *distribution paramétrée* (moyenne, écart-type) et donc la satisfaction d'hypothèses relatives à la normalité et l'égalité des variances dans la population.

2. Le test du chi-carré est un **test unilatéral supérieur**, les valeurs allant de 0 à plus l'infini du fait qu'il repose sur les écarts au carré entre fréquences observées et théoriques. Par contre, le test t peut être bilatéral (les deux côtés de la distribution sont considérés) ou unilatéral (inférieur ou supérieur) ; 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 t ?

Le test t approfondit l'analyse du tableau des moyennes en aidant à établir dans quelle mesure une différence est 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 t . 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 dichotomique** et ses valeurs possibles sont : **temps faible** (< à la médiane par exemple), **temps élevé** (\geq à la médiane par exemple).

La variable dépendante, « indice de masse corporelle », doit être de nature **quantitative** et ses valeurs possibles sont : **15, 16, 17 kg/m²**, etc.

NB : Lorsque les deux variables sont quantitatives, il faut dichotomiser les valeurs de la variable indépendante pour que la relation se prête au test t 😊

4) Un sociologue veut savoir si les fumeurs et les non-fumeurs (catégories de la variable indépendante) ont la même espérance de vie. Il décide alors de regrouper les valeurs de la variable dépendante quantitative, « espérance de vie en années », en deux catégories (faible, élevée) dans la perspective d'utiliser le test du chi-carré.

En tant qu'assistant-statisticien, vous conseillez au sociologue de ne pas transformer les valeurs de la variable dépendante (espérance de vie) en catégories, car vous estimez qu'un test de différence de deux moyennes 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 sociologue regroupe les scores de la variable « espérance de vie » en catégories, il **perd de l'information**. Non seulement il ne retrouvera plus les différences de scores entre les cas, 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 le tabagisme et l'espérance de vie. Il se peut que cette relation soit inhibée.

D'autre part, si le sociologue regroupe les scores de la variable « espérance de vie » 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 ordinaire. 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 médecin **perd ainsi en puissance statistique**. Streiner (2002)¹ estime l'efficacité de la dichotomisation d'une variable continue à 67%, soit une perte de 33%.

¹ David L. Streiner (2002). Breaking up is hard to do : the heartbreak of dichotomizing continuous data, *Can J Psychiatry*, 47 :262-266.

Bref, la perte d'information (dont le changement de la structure de la relation est la résultante) et la perte des propriétés mathématiques (qui occasionne une perte de puissance statistique) sont deux inconvénients liés au choix du sociologue. Pour éviter cela, il est approprié de conserver les valeurs métriques de l'espérance de vie dans la perspective d'effectuer un test t , lequel comparerait les moyennes des fumeurs et celles des non-fumeurs concernant l'espérance de vie.

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) Vous souhaitez comparer les dépenses moyennes en dollars dans les discothèques des célibataires et non-célibataires chez les étudiants. À partir d'un échantillon aléatoire, vous obtenez les résultats SPSS ci-dessous:

	Statut en terme de célibataire	n	Moyenne	Ecart-type
Dépense à la discothèque par mois	Célibataires	1837	26,17	45,623
	Non célibataires	357	13,44	40,297

	Test-t pour égalité des moyennes			
	t	dl	Différence entre les deux moyennes	Erreur-type de la différence
Dépense à la discothèque par mois	4,916	2192	12,738	2,591

a) Quel est le montant moyen dépensé dans les discothèques par les célibataires? Quel est le montant moyen dépensé dans les discothèques par les non-célibataires? Quelle est la différence entre ces deux moyennes?

Le montant moyen dépensé dans les discothèques par les 1837 célibataires est de 26 dollars. Le montant moyen dépensé dans les discothèques par les 357 non célibataires est de 13 dollars. La différence est donc de 13 dollars.

b) Spécifiez les hypothèses statistiques du test t .

Hypothèse nulle : les dépenses moyennes dans les discothèques des célibataires et des non-célibataires sont égales dans N. $H_0 : \mu_1 = \mu_2$

Hypothèse alternative bilatérale : les dépenses moyennes dans les discothèques des célibataires et des non-célibataires ne sont pas égales dans N. $H_a : \mu_1 \neq \mu_2$

NB : L'hypothèse alternative postulant simplement l'existence d'une différence dans les dépenses moyennes, le test d'hypothèse sera bilatéral.

c) Que vaut la statistique de t de Student ? Que signifie cette valeur ?

Elle vaut **4,916**. Cette valeur signifie que la différence entre les moyennes des célibataires et des non-célibataires se situe à 4,9 écarts-types (erreurs-types précisément) par rapport à 0.

d) Que vaut l'erreur-type de la différence $s_{\bar{x}_1 - \bar{x}_2}$? Que signifie cette valeur ?

Elle vaut **2,591**. Cette valeur signifie que si l'on calculait les différences entre les moyennes des célibataires et des non-célibataires pour tous les échantillons possibles, l'**écart-type de ces différences de moyennes** serait de 2,591. Précisément, environ deux tiers (68%) des échantillons se situeraient à plus ou moins 2,591 dollars de la différence des deux moyennes : 12,738.

e) Divisez la différence des moyennes par l'erreur-type de la différence. Retrouve-t-on la valeur calculée du t ?

12,738/2,591= 4,916. Oui, on retrouve la valeur calculée du t .

f) Sur la base de la table du t , quelle est la valeur critique du t au seuil de 0,05 ?

Pour un test bilatéral, elle est de 1,961 au seuil 0,05 avec 2192 degrés de liberté.

g) Comparez les valeurs calculée et critique du t et prenez une décision d'acceptation ou de rejet de l'hypothèse nulle

La valeur calculée du t (4,916) étant supérieure à la valeur critique du t (1,96), on rejette l'hypothèse nulle pour accepter l'hypothèse alternative au seuil de 0,05.

h) Que pouvez-vous conclure sur la relation étudiée?

Il y a au moins 95% de chances qu'il y ait une relation statistiquement significative entre le statut matrimonial et les dépenses dans les discothèques chez les étudiants. La différence dans les dépenses entre les célibataires et les non-célibataires n'est probablement pas due à une erreur d'échantillonnage. Les célibataires consacrent plus de dépenses dans les discothèques que les non-célibataires.

6) Je veux savoir s'il y a une différence significative dans les salaires entre les hommes et les femmes. Je présente ainsi les résultats :

L'analyse des données de l'enquête sur la dynamique du travail au Canada (2006) indique que les 16 128 hommes gagnent en moyenne 51 760\$, alors que le revenu moyen des 13 526 femmes est de 36 348\$. La différence entre les deux groupes est de 15 413\$. Elle s'avère statistiquement significative au seuil de 99,9% ($t = 30,09$, $dl = 29 652$). Par conséquent, on est sûr au moins à 99,9% que le sexe influe sur le revenu chez l'ensemble des Canadiens actifs.

a) Variables indépendante et dépendante ainsi que leurs valeurs:

Variable indépendante : sexe
Variable dépendante : revenu en dollars

b) Population étudiée :

Elle réfère à l'ensemble des Canadiens actifs en 2006.

c) Taille de l'échantillon :

L'échantillon est constitué de 29 654 répondants (16 128 + 13 526).

d) Exprimez en des termes sémantiques, c.-à-d. non techniques, la signification de la valeur du « t ».

La valeur du t (30,09) signifie que la différence entre le revenu des femmes et celui des hommes se situe à 30 écarts-types (erreurs-types) par rapport à 0.

e) Exprimez en des termes sémantiques, c.-à-d. non techniques, la signification de la valeur de « dl ».

La valeur de « dl » (29 652) signifie le nombre de répondants dont le revenu varie librement ou de façon aléatoire, sachant qu'au sein de chacun des groupes relatifs aux deux sexes le revenu d'un répondant est déterminé d'avance (29 654-2).

f) Exprimez en des termes sémantiques, c.-à-d. non techniques, la signification de « 99,9% ».

La valeur 99,9% (ou 0,99) est le niveau de confiance du test d'hypothèse t . Elle signifie qu'il y a au moins 99,9% de chances de ne pas nous tromper si nous concluons à l'existence d'une différence entre les hommes et les femmes quant à leur revenu (pour un test bilatéral).

7) Les garçons sont-ils meilleurs que les filles en statistiques sociales ? Pour répondre à cette question, un test t est effectué à partir des notes finales en % obtenues par les étudiants du SOL1020-A11 (je me suis fié aux prénoms pour classer les étudiants : les risques d'erreur sont donc réels puisque je n'appartiens pas à la même culture que tous les étudiants). D'après ce classement, voici les résultats obtenus à l'aide du logiciel Excel.

Test d'égalité des espérances: deux observations de variances égales 95%

	Femmes	Hommes
Moyenne	77,312	74,912
Observations (n)	67	40
Différence hypothétique des moyennes	0	
Degré de liberté	105	
Statistique t	0,888	
Valeur critique de t (unilatéral)	1,659	
Valeur critique de t (bilatéral)	1,983	

a) Variables indépendante et dépendante ainsi que leurs valeurs:

Variable indépendante : sexe

Variable dépendante : notes finales en % en statistiques sociales

b) Taille de l'échantillon des femmes et celle de l'échantillon des hommes :

Taille de l'échantillon des femmes : 67 cas

Taille de l'échantillon des hommes : 40 cas

c) Taille de l'échantillon total :

Au total, l'échantillon est constitué de 107 étudiants.

d) La différence entre les moyennes des femmes et des hommes vous paraît-elle importante?

La différence entre les notes moyennes des femmes et des hommes étant de 2,4% (77,3-74,9), elle nous paraît relativement faible. Reste à savoir si elle est statistiquement significative.

e) Quelle est la valeur calculée du t ? Que signifie-t-elle ?

La valeur calculée du t est de 0,888, soit l'écart-type qui sépare la différence des notes moyennes des femmes et des hommes à la valeur 0.

f) Quelle est la valeur critique du t pour le test bilatéral ? Que signifie-t-elle ?

La valeur critique du t pour le test bilatéral est de 1,98, soit la valeur minimale nécessaire au rejet de l'hypothèse nulle. Précisément, si l'on obtenait tous les échantillons possibles (distribution d'échantillonnage), la valeur maximale du t équivaldrait à 1,98 s'il n'y avait pas de relation entre le sexe et la performance, au seuil de signification 0,05 et avec 105 degrés de liberté.

La valeur calculée du t doit excéder cette valeur critique pour que le test bilatéral soit significatif : différence entre les moyennes des femmes et des hommes (\neq).

g) Quelle est la valeur critique du t pour le test unilatéral ? Que signifie-t-elle ?

La valeur critique du t pour le test unilatéral est de 1,659, soit la valeur minimale nécessaire au rejet de l'hypothèse nulle. Précisément, si l'on obtenait tous les échantillons possibles (distribution d'échantillonnage), la valeur maximale du t équivaldrait à 1,659 s'il n'y avait pas de relation entre le sexe et la performance, au seuil de signification 0,05 et avec 105 degrés de liberté.

La valeur calculée du t doit excéder cette valeur critique pour que le test unilatéral soit significatif : moyenne des femmes supérieure à celle des hommes ($>$).

h) Comparez les deux valeurs critiques du t pour le test bilatéral et le test unilatéral ? Justifiez.

La valeur critique du t pour le test bilatéral (1,98) est plus grande que la valeur critique du t pour le test unilatéral (1,659). Cette différence est liée au fait que le test unilatéral est associé à un seuil de signification deux fois plus petite que le test bilatéral (0,025 contre 0,05) pour une même valeur critique, puisqu'il ne comporte qu'une seule direction ou zone de rejet de l'hypothèse nulle.

i) Selon le test bilatéral, doit-on rejeter ou accepter l'hypothèse nulle ? Concluez.

Avec 105 degrés de liberté et au seuil de signification 0,05 pour le test bilatéral, la valeur calculée du t (0,888) est plus petite que la valeur critique du t (1,98). On accepte l'hypothèse nulle. Par conséquent, il y a 95% de chances que la note moyenne des femmes soit égale à celle des hommes dans la population.

- j) Selon le test unilatéral, doit-on rejeter ou accepter l'hypothèse nulle ?
Concluez.

Avec 105 degrés de liberté et au seuil de signification 0,025 pour le test unilatéral, la valeur calculée du t (0,888) est toujours plus petite que la valeur critique du t (1,659). On accepte l'hypothèse nulle. Par conséquent, il y a 95% de chances que la note moyenne des femmes soit égale à celle des hommes dans la population.

- k) Qu'est-ce qui détermine le choix d'un test bilatéral ou d'un test unilatéral dans le cadre du test t ?

La façon dont l'hypothèse de recherche est formulée détermine le choix d'un test plutôt qu'un autre. Si l'hypothèse de recherche ne précise pas la direction de la différence entre les deux groupes (\neq), le test bilatéral est approprié. Par contre, si l'hypothèse de recherche précise la direction de la différence dans un sens ou dans l'autre ($<$ ou $>$), le test unilatéral est approprié.

NB : Les chercheurs préfèrent utiliser le test bilatéral puisqu'il est plus exigeant en termes de rejet de l'hypothèse nulle et donc plus conservateur en termes d'acceptation de l'hypothèse nulle. Car, s'il est plus exigeant de rejeter l'hypothèse nulle et qu'on y arrive malgré tout, nos résultats seront plus crédibles lorsqu'on les infère à toute la population dont provient l'échantillon 😊