L'UTILISATION PRATIQUE DES TRAITEMENTS STATISTIQUES SOUS « EXCEL »

Adel SIDI-YAKHLEF

Par Alain Mouchès (*Maître de Conférences à l'Institut de Psychologie et de Sociologie Appliquées, U.C.O, Angers.)*

1.Généralités :

Un travail de recherche permet d'analyser et interpréter nos données, pour vérifier nos hypothèses. Et cette validation des conclusions expérimentales est intimement liée à l'emploi de la statistique. Mais le choix des tests appropriés est souvent délicat.
Le document proposé n'est pas un abrégé de statistiques, mais simplement une aide concrète vous permettant d'acquérir un « savoir-faire » des principaux tests statistiques.
Toutefois il est utile de vous souvenir de vos cours de statistiques, ou au besoin d'avoir sous la main un ouvrage de statistiques pour suivre ces exercices.

On parle souvent en Sciences Humaines, de "variable dépendante" et de "variables indépendantes"... Rappelons que la variable définit les caractéristiques de la mesure que l'on utilise pour prélever l'information.
La variable dépendante = les données : se poser une question-problème, et décrire la conduite étudiée.
La variable indépendante = source de variations, conditions manipulées par l'observateur.

Petit rappel (*avec un exemple tout à fait absurde, je le précise !*):
Supposons que je veux étudier la consommation de chocolat chez les étudiants (= *Variable dépendante*), et plus précisément, je veux savoir si les Littéraires mangent plus (ou moins) de chocolat que les Scientifiques (*Variable indépendante*).

Première difficulté: la typologie des variables.
Pour évaluer un traitement à partie des données opérées, il faut déterminer le type d'échelle de mesure utilisé.
Généralement, il existe trois niveaux de mesure. Ma variable dépendante est-elle: ordinale? nominale? d'intervalle ?
Nominal = classe d'équivalence, ordinal = plus grand que..., et intervalle = grandeur des intervalles entre les objets d'une échelle ordinale.
Si je décide de noter simplement l'existence ou l'absence de chocolat selon les individus (*je note « oui », ou « non », sans considérer la quantité*) alors la variable dépendante est devenue «nominale ».
Par contre, si je décide de comptabiliser le poids consommé de chocolat (en gramme) par jour, et par individus, dans ce cas nous avons affaire à une variable ordinale (continue). De même si l'on demande à notre consommateur de chocolat d'estimer de façon numérique sa dépendance (par ex . en utilisant une échelle de type « Likert » : 0= pas du tout, 1= un peu, 2 = assez souvent, etc.), c'est encore une échelle ordinale. Toutefois certains auteurs préfèrent parler d'échelle d'intervalle… Et j'avoue ne pas saisir toutes ces subtilités !

Disons qu'il existe des échelles « d'intervalles », c'est à dire sous forme de valeurs numériques particulières.
Par exemple on peut estimer le temps mis pour manger toute une tablette de chocolat.
Ou encore on obtient un score après épreuve qui indique l'état du consommateur, après ingestion de toute la tablette. (calcul par cumul des événements psychophysiologiques – nausée, anxiété, etc.-cités dans un questionnaire).

En tout cas selon les différentes échelles, on utilisera des tests appropriés.
Mais il existe un autre problème. Certains tests peuvent être « paramétriques », et d'autres « non-paramétrique ».
Que signifie cette différence entre tests ?

Si ma variable est ordinale, et si la population des étudiants est importante, on peut supposer que la distribution suit la loi normale (loi de Gauss).
En effet la consommation de chocolat varie selon les individus : quelques personnes ont une consommation nulle, ou très faible et au contraire quelques personnes trop gourmandes mangent toute une plaquette, et la majorité des individus auront une consommation plus raisonnable… Donc les échantillons suivent une distribution normale, c'est à dire un distribution « en forme de cloche ».
Si ma variable suit la loi de probabilité de Gauss, j'ai "le droit" d'utiliser les tests paramétriques. Je pourrai par exemple utiliser un « test de moyennes », tel que le « t de Student »
Cependant pour compliquer encore, on peut avoir des variables qui ne suivent pas vraiment la loi normale… Dans ce cas, on préférera les tests « non-paramétriques ».

En réalité, c'est parfois difficile de choisir les tests employés...
En effet, en particulier dans le cas des petits échantillons, certains histogrammes obtenus sont plus "ordinales" que "nominales", mais pourtant sont très loin d'une distribution dite « normale ».
Dans certains cas, les tests non-paramétriques sont plus adaptés. Et de fait, il existe des méthodes non-paramétriques qui traitent aussi des variables ordinales, et qui sont très adaptables à des cas particuliers.
Cependant beaucoup de chercheurs en Sciences humaines préfèrent utiliser les tests "paramétriques"… C'est une affaire de choix ! (*ou de flemme ?*).
Je vous signale néanmoins que certains nostalgiques des tests non-paramétriques ont réalisé des logiciels « free » permettant de calculer ces tests.

Dans tous les cas, le logiciel Excel (ainsi que ce logicien « free » trouvé par Internet) va vous permettre de réaliser très facilement la plupart des traitements statistiques, paramétriques ou non-paramétriques.

Mais auparavant, quelques « astuces » pour traiter facilement vos données sous Excel

A. Une première astuce : le « collage spécial » *(attention cette information concerne exclusivement la version « ancienne » d'Excel, et non la version Excel2007 !)*

Mes données que je vais tester sont en « ligne », et je veux qu'ils soient en « colonne »…
Que faire ? Réponse : si vous devez changer vos données de « ligne» en « colonne »-ou

inversement-: copiez vos données, et sélectionnez un emplacement, puis dans « **Edition** », choisir « **Collage spécial** », puis « **Transposé** », et cliquez **OK**.

B. Une deuxième astuce : « le filtrage »

Un exemple: vous venez de saisir les résultats d'un questionnaire…

SUJET	AGE	TEST 1	TEST 2	SEXE
1	enfant	25	10	homme
2	adulte	26	11	femme
3	adolescent	42	14	homme
4	adolescent	36	10	homme
5	adulte	21	9	homme
6	adulte	20	8	femme
7	enfant	32	12	femme
8	adulte	31	14	homme
//	…	...		femme
268	…etc.	…	…	…..

Vous possédez une foule d'informations, mais si vous devez comparer manuellement vos résultats aux différentes modalités (homme ou femme, grand moyen ou petit, enfant ou adulte etc…), votre analyse sera bien complexe !

Mais Excel possède un outil très efficace : le « filtrage », très pratique pour traiter vos données.
Procédure : dans « **Données** », cherchez « **filtre** ». Sélectionnez une cellule (par exemple dans « sujet », ou « sexe », ou « âge » etc.), et cliquez sur la commande « **filtrage automatique** ».
Ensuite vous pouvez très facilement séparer vos groupes soit en « hommes », soit en « femmes », ou encore vous pouvez analyser uniquement les « hommes-adultes », etc.

C. Où trouver les analyses statistiques intéressantes, sous Excel ?

C'est paradoxal, mais vous ne trouverez pas beaucoup de tests statistiques intéressants dans la fonction « statistiques » d'Excel !
Dans les versions « anciennes) d'Excel, l faut plutôt chercher dans les « macros », et plus précisément dans « **Utilitaire d'analyse** ».
Comment peut-on trouver ce précieux « macro » ? Dans **« Outils »,** cherchez « **Utilitaire d'analyse** », (et si vous ne le trouvez pas, cherchez dans « **macros complémentaire** », et cochez « Utilitaire d'analyse »…)
(remarque : dans la version Excel 2007, il faut cliquer le bouton Microsoft Office, et activer (en bas) « Option Excel », puis « Compléments », « Gérer », « complément Excel ». Et dans les Macros complémentaires disponibles, il faut activer la case « Analysis ToolPak ») Ensuite vous trouverez « l'utilitaire d'analyse » dans « Données »… Ouf !
Dans le cas des tests non-paramétriques, nous avons utilisé le logiciel "Astro Research" de Mr H. Delboy, médecin, statisticien, astrologue, musicologue, etc… Ce scientifique passionné d'astrologie, alchimie et d'autres bizarreries ésotériques a réalisé un logiciel remarquable et gratuit, qui fonctionne sous Excel. (*adresse :**hdelboy.club.fr/Nonparam.htm***)

2. Calculs statistiques paramétriques:

Ces quelques pages vous expliquent la marche à suivre des calculs les plus utilisés, en donnant des exemples.

A. L'enregistrement des observations:

1- Calculer la moyenne, l'écart-type, analyser la dispersion, etc…
(Visitez vos anciens cours de statistiques, SVP…)

Procédure : dans « **Utilitaire d'analyse** », cliquez « **Statistiques descriptives** », et cochez « **Rapport détaillé** ».

Entrez vos données dans « plage d'entrée » (en sélectionnant avec la souris la zone choisie), précisez si les données sont en colonnes, ou en lignes, et faites OK.

Vous trouvez aussitôt *la **moyenne**, **l'erreur-type** (*Erreur-type : $s_x = S/\sqrt{n}$), *la **médiane**, le **mode*** (= la valeur de l'observation associée à la fréquence la plus élevée) *, **l'écart-type*** (Ecart-type : $S = \sqrt{\sum (X - \overline{X})^2 / n-1}$*), la **variance** de l'échantillon (= le carré de l'écart-type S),, le coefficient d'aplatissement **Kurstosis**, le coefficient d'assymétrie, etc…*

2- Réalisation d'une distribution de fréquence : création d'un histogramme de données quantitatives groupées.

Exemple: un enseignant vient de corriger 20 copies d'examen. Les notes vont de 2 à 18/20, et il souhaite connaître la distribution.

Notes :

10	**9**	**8**	**7,5**	**17**	**18**	**12**	**13**	**7**	**6**	**4,5**	**11**	**13**	**10**	**8**	**8**	**11**	**13**	**2**	**6**	**11**

Cet enseignant décide d'utiliser des intervalles de notes pour réaliser un graphique plus représentatif.
Il détermine 9 classes, correspondant à l'intervalle de partition :
(1à 3), (3-5), (5-8), … (18-20)

Tableau de 9 classes :

1	**3**	**5**	**8**	**10**	**12**	**14**	**16**	**18**

Procédure : dans **« Utilitaire d'analyse », cliquez « Histogramme** ».
Rentrez les notes dans « plage d'entrée», et les 9 classes dans « plage des classes».
Vous pouvez cocher également « **représentation graphique** », puis « **OK** »… Et vous aurez aussitôt un résultat indiquant les classes, la fréquence des résultats, (*et en prime, un joli histogramme...*) Vous pouvez d'ailleurs transformer cet histogramme tout à loisir dans l' « Assistant graphique » d'Excel.

Remarque : si vous souhaitez créer une distribution de fréquence avec des données « non-groupées », il ne faut plus utiliser l'outil « histogramme » de l'Utilitaire d'analyse, mais à l'aide du « **Tableau croisé dynamique** » qui se trouve dans le menu « **Données** ».

Dans notre cas, cliquez sur « suivant », indiquez vos notes dans « plage », et cliquez sur **« disposition »**…

Ensuite glissez simplement le champ des « notes » sur le rectangle « ligne », puis glissez à nouveau sur « données ». Ensuite, cliquez « Terminer »… Là, vous allez vous sentir un peu bête car vous n'obtenez pas de « Fréquence », mais une banale « Somme » ! C'est normal, ne paniquez pas… Cliquez deux fois sur « somme », et vous tombez dans un « Champ dynamique », plein de merveilles : somme, moyenne, écart-type, produit, etc. Ici, choisissez « Nb » (qui signifie le nombre d'occurrence, ce qui correspond tout à fait !)

Le **tableau croisé dynamique** est également très intéressant pour réaliser un questionnaire, des tableaux, des analyses croisées, etc. Amusez-vous à vous exercer en glissant les différents boutons proposés, et bientôt vous allez devenir un « accro » d'Excel…

LES TESTS STATISTIQUES POUR UN, DEUX, OU *K* ECHANTILLONS

La plupart des tests sera un comparaison de moyennes ou de fréquences…
Mais il faut tout d'abord identifier la (ou les) variables. Comment est formée ma variable dépendante ? Quel type d'échelle faut-il employer? La variable est-elle « ordinale », ou alors « nominale »?

Trois possibilités : nous voulons analyser
-un seul échantillon à tester,
-deux échantillons,
-ou k échantillons…

Par exemple, si je compare simplement les étudiants qui consomment (ou non) du chocolat, c'est une variable à 1 échantillon. Si je veux analyser la comparaison Littéraire/Scientifique, et la consommation du chocolat, alors c' est une variable indépendante à 2 échantillons…
Et si je veux analyser la comparaison Littéraire/Scientifique des accros du chocolat, en considérant le sexe des individus, alors c'est une variable indépendante à 4 échantillons... Je vous conseille de regarder le tableau récapitulatif qui se trouve à la dernière page de ce document.

B. Les tests statistiques pour un, ou deux échantillons

Il faut d'abord préciser ce qu'on cherche: soit mon hypothèse suppose une indépendance (c'est à dire une absence de relation), ou au contraire mon hypothèse suppose une liaison (c'est à dire une association « corrélée »)?

B.1 **: Les tests d'indépendance:**

1-le test de Student, comparaison d'une moyenne :

Formule $t = \dfrac{\overline{X} - \mu}{s / \sqrt{n-1}}$

Exemple : d'après un rapport, on trouve que les hommes de plus de 30 ans regardent la télévision en moyenne 25 h par semaine. Nous voulons comparer cette moyenne à une population d'étudiants. Onze étudiants ont comptabilisé leur temps passé devant la télévision, par semaine :

Résultats

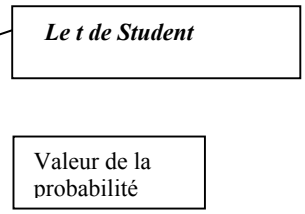
Etudiants	10	8	15	28	20	19	13	20	9	14	38

Procédure : dans « **Utilitaire d'analyse** », cliquez « **Test d'égalité des espérances : observations pairées ». Par un copier-coller (en colonnes, SVP [1]),** rentrez les échantillons observés dans « plage pour la variable 1 », et dans « plage pour la variable 2 » répétez simplement n.fois la moyenne théorique (ici, 25) :

Etudiants	10	8	15	28	20	19	13	20	9	14	38
théorique	25	25	25	25	25	25	25	25	25	25	25

Puis, faites **OK** : nous obtenons un tableau tout à fait clair, avec plusieurs informations:

Test d'égalité des espérances: observations pairées		
	Variable 1	*Variable 2*
Moyenne	17,6363636	25
Variance	80,2545455	0
Observations	11	11
Différence hypothétique des moyennes	0	
Degré de liberté	10	
Statistique t	-2,72617579	
P(T<=t) unilatéral	0,01066649	
Valeur critique de t (unilatéral)	1,81246151	
P(T<=t) bilatéral	0,02133298	
Valeur critique de t (bilatéral)	2,22813924	



Notez la moyenne des échantillons (17,63..), leur variance (80,25) les ddl (11), la probabilité (uni, ou bilatéral) etc.
Vous constatez que la moyenne des échantillons-étudiants est plus faible que celle de la population générale. Il y a une différence significative (p = .01).
Nous rejetons donc l'hypothèse nulle : les étudiants regardent moins la télévision que les adultes de plus de 30 ans.

2- **Le rapport de variance** : test de F de Fischer-Snedecor. Ce test permet de vérifier l'existence significative de différences entre les moyennes de 2 groupes. Et plus exactement, il permet de tester l'hypothèse de l'égalité des variances des 2 populations. On va estimer la dispersion des valeurs entre les deux distributions, en définissant les valeurs du rapport des deux variances.
(Formule: $F = S_1^2/S_2^2$) C'est à dire : rapport des 2 variances observées (en pratique, rapport de la plus grande valeur à la plus petite) . Selon les tables de Snedecor, si F est supérieur à 2,27, il y a 5 chances sur 100 pour que la différence observée soit significative.

[1] Pour passer de « ligne» en « colonne » sous Excel, copiez vos données, et sélectionnez un emplacement, puis dans « Edition », choisir « Collage spécial », puis « Transposé », et cliquez OK.

Procédure : dans « **Utilitaire d'analyse** », cliquez « **Test d'égalité des variances ».** Rentrez les deux échantillons dans « plage pour la variable 1 », et « plage pour la variable 2 », et faites « **OK**».
Et bien sûr, si vous constatez que la valeur du F est non-significative (cela veut dire que les deux distributions ne diffèrent pas du point de vue de la dispersion de leurs valeurs), alors dans ce cas, vous pouvez comparer les deux moyennes.

3- **le test de Student**: test de **deux moyennes** d'échantillons **appariés**. *(ou échantillons dépendants)* : Formule du **t** de Student: $t = \frac{\overline{D}}{S_{d\,/\,\sqrt{n-1}}}$

Avec $\overline{D}$ = moyenne de la différence des 2 moyennes
S_d = écart-type (de la différence … etc.) N = taille de l'échantillon

Exemple : nos 11 étudiants, (*apparemment passionnés par les expériences !*) passent un test d'anxiété, puis sont invité à participer à un entraînement à la relaxation. Ensuite, ils repassent le test d'anxiété… On veut évidemment estimer l'efficacité d'une formation à la relaxation.

Résultats

avant	30	38	45	28	20	19	23	40	29	34	38
après	10	21	16	16	11	22	23	26	18	32	28

Procédure : dans « **Utilitaire d'analyse** », cliquez « **Test d'égalité des espérances : observations pairées** ». **Par un coller-copier (en colonnes, SVP),** rentrez les deux échantillons dans « plage pour la variable 1 », et « plage pour la variable 2 », et faites « OK »

Un tableau s'affiche aussitôt :

Test d'égalité des espérances: observations pairées		
	Variable 1	*Variable 2*
Moyenne	31,2727273	20,2727273
Variance	72,6181818	47,4181818
Observations	11	11
Coefficient de corrélation de Pearson	0,29512579	
Différence hypothétique des moyennes	0	
Degré de liberté	10	
Statistique t	3,94784499	
P(T<=t) unilatéral	0,00136992	
Valeur critique de t (unilatéral)	1,81246151	
P(T<=t) bilatéral	0,00273983	
Valeur critique de t (bilatéral)	2,22813924	

Il indique plusieurs informations : moyenne, variance, etc., et même le coefficient **r** de Pearson (qui indique s'il y a une corrélation, ou non, entre les deux variables…)
Dans notre cas, nous allons nous s'intéresser à la valeur du **t** de Student, qui est indiqué dans la ligne « Statistique t » = 3,94. Le résultat est hautement significatif (probabilité unilatérale alpha de .001) (*Mais quel dommage, ce ne sont ici que des chiffres totalement inventés*…)

4- **le test de Student pour des échantillons indépendants** : il faut dans ce cas prendre le « **test d'égalité des espérances** » (*vous avez le choix entre « variances égales », ou « variances différentes »)*…

En théorie, le test t sur des échantillons indépendant suppose que les variances sont inconnues, mais égales. Mais parfois lorsqu'on suppose que les variances sont inégales –par exemple dans le cas des tailles d'échantillons trop réduites-, Excel utilise un autre calcul appelé la procédure de Welch-Aspin… *(Personnellement, je préfère utiliser dans ce cas un test non-paramétrique ...*)
En tout cas, dans une situation « normale » d'un test de Student à variances égales, la formule du **t** de Student, comparaison de deux moyennes est: $t = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}}$

(avec s² =variance commune aux deux échantillons).

Un exemple: nous avons choisi au hasard 8 garçons et 9 filles qui ont passé un concours de mathématiques. Les résultats sont indiqués dans ce tableau.

HOMMES	FEMMES
56	40
54	30
25	60
65	65
45	24
58	52
45	50
48	36
	30

En utilisant ce « **test d'égalité des espérances** », vous n'avez qu'à placer (dans les Paramètres d'entrée » les résultats des garçons (« *plage pour la variable 1 »*), et le résultat des filles *(« plage pour la variable 2 »)* et vous faites « OK ». On obtient aussitôt ce tableau :

Test d'égalité des espérances: deux observations de variances égales		
	Variable 1	*Variable 2*
Moyenne	49,5	43
Variance	145,428571	207,5
Observations	8	9
Variance pondérée	178,533333	
Différence hypothétique des moyennes	0	
Degré de liberté	15	
Statistique t	1,00114155	
P(T<=t) unilatéral	0,16631795	
Valeur critique de t (unilatéral)	1,75305104	
P(T<=t) bilatéral	0,33263591	
Valeur critique de t (bilatéral)	2,13145086	

Vous avez ici un résultat qui n'est pas significatif (t = 1,001 inférieur à la valeur critique de **t**, avec ddl :15, et un probabilité alpha de 0,166). Les garçons ne sont pas meilleurs en Maths que les filles.

5- **le test « z »** de deux moyennes (dans le cas des grands échantillons).

Procédure : dans « **Utilitaire d'analyse** », cliquez « test de la différence significative minimale ». *Attention : il faut d'abord calculer les 2 variances (voir « statistiques descriptives », par exemple)…* Puis, rentrez les données, et faites OK.

B.2 : **Les tests de corrélation** : ou la « force » d'une liaison entre deux, ou plusieurs séries de données.

1. Le test « r » de Bravais-Pearson

Formule : $r = \frac{\sum (x_i - m_x)(y_i - m_y)}{\sqrt{\sum (x_i - m_x)^2 \sum (y_i - m_y)^2}}$

(avec xi= valeurs échantillon 1, yi= valeurs échantillon 2, et m= moyenne échantillon)

On peut calculer très facilement le « r » de Bravais-Pearson sous Excel:

Procédure 1: dans « **Utilitaire d'analyse** », cliquez « analyse de corrélation »… Non seulement vous aurez un tableau dans lequel se trouve la corrélation, mais vous pouvez également calculer plusieurs corrélations en fonction des séries d'observations testées… (cf coefficient de corrélation partielle entre X et Z pour y1z1, y1z2, y3z2, etc…)

Procédure 2 : chercher (sur les boutons d'Excel) l'icône ***fx*** *(= « Coller une fonction »), puis cherchez la fonction statistiques, puis « coefficient.corrélation », (ou encore « Pearson », c'est le-même calcul…). Collez vos données dans « matrice 1 », puis dans « matrice 2 », et faites OK : la corrélation est aussitôt indiquée*

*Procédure 3 : ou éventuellement en cliquant « **Test d'égalité des espérances : observations pairées** » ! En effet nous avons vu que ce test de deux moyennes va calculer non seulement les moyennes et variances, mais également l'analyse de corrélation entre les deux variables.*

Attention, l'utilisation des corrélations reste délicate car certaines variables peuvent influencer les autres, et on peut parfois trouver des résultats absurdes.
Par exemple, en testant une population de lycéens, des étudiants ont obtenu une corrélation surprenante : en croisant les résultats du saut en hauteur, et le poids des sujets, ils ont trouvé une corrélation significative (r =.60) : conclusion, plus vous êtes gros, plus vous sautez haut !
Bien évidemment ces étudiants avaient oublié une variable importante : celle de l'âge… Bien sûr, les petits collégiens de 12 ans sautent généralement moins haut que les grands lycéens de Terminale, et donc la corrélation apparente entre test et poids disparaît si l'on considère l'âge constant !
Dans ce cas, il faut alors réaliser une **corrélation partielle** en éliminant l'effet de certaines variables.
Revenons à notre exemple : nous trouvons r =.60 pour la corrélation A : saut/poids, mais il faut également calculer la corrélation B: saut/âge (ici, r = .69), et bien sûr la corrélation C : poids/âge (r = .88).

	Test saut	Poids	Age
Test saut	-	-	-
Poids	rA = .60	-	-
Age	rB = .69	rC =.88	-

Le calcul de corrélation partielle est $\frac{rA - (rB.rC)}{\sqrt{(1 - r^2B).(1 - r^2C)}}$.

Ce qui correspond *(en « traduction Excel»)* à cette formule un peu bizarre :

=(A1-(B1*C1))/((1-B1^2)*(1-C1^2))^0,5

Vous n'avez qu'à copier cette formule, et la coller sous Excel.
Il faut au préalable placer les chiffres dans les cases indiqués (A1, B1, C1).
Ainsi dans notre exemple on écrit 0,60 dans A1, 0,69 dans B1, et 0,88 dans C1, puis vous collez la formule dans une case quelconque : le calcul est aussitôt réalisé. La corrélation partielle donne r = -.02, c'est-à-dire une corrélation parfaitement nulle !

C. Les tests statistiques pour **k** échantillons :

La comparaison de plusieurs moyennes : La VD est ordinale, et nous voulons analyser k échantillons.

C.1 **- les tests d'indépendance** :

Il faut utiliser les **analyses de variances** (ANOVA), en analysant le croisement d'une, deux variables (*ou même plusieurs variables, avec le risque de devenir fou !*). Il existe des logiciels très adaptés (SPSS, Var3, Sphinx ou autres...) Mais attention à la "pêche à la ligne" des comparaisons multiples ! Les ANOVA multiples sont évidemment intéressantes, mais en comparer sans discernement plusieurs échantillons, on peut trouver des résultats totalement absurdes!

En tout cas, l'utilitaire d'analyse d'Excel permet aussi de calculer des ANOVA.
Sans être aussi performant que certains logiciels statistiques, il est suffisant pour la plupart des cas...
Plus exactement, Excel propose d'effectuer :

- **une analyse de variance entre groupes pour 1 facteur**
- **une analyse de variance entre groupes pour un facteur, avec des mesures répétitives**
- **une analyse de variance entre groupes pour deux facteurs**.

1- **Analyse de variance à un facteur,** constitué de k modalités (*Plan : Sn<An>*)

Exemple : un chercheur veut savoir si la musique peut jouer sur l'apprentissage… Pour cela, notre chercheur fait apprendre des listes de mots à 4 groupes d'étudiants qui entendent des styles de musiques nettement différents : de l'opéra, du flamenco, du piano classique, et du free-jazz…On note le nombre de mots mémorisés après apprentissage…

opéra	flamenco	piano	jazz
13	15	12	16
15	12	13	12
13	12	10	13
14	15	12	14
15	14	14	10
10	11	12	11
16	15	16	15
14	15	15	12
15	14	12	12
13	15		16
14			17

Procédure : dans « **Utilitaire d'analyse** », cliquez « **analyse de variance : un facteur** », et comparez (en précisant colonnes, ou lignes) ces résultats en faisant OK.

Résultats : on trouve sur Excel ce tableau :

Analyse de variance: un facteur

RAPPORT DÉTAILLÉ

Groupes	*Nombre d'échantillons*	*Somme*	*Moyenne*	*Variance*
Colonne 1	11	152	13,8181818	2,56363636
Colonne 2	10	138	13,8	2,4
Colonne 3	9	116	12,8888889	3,36111111
Colonne 4	11	148	13,4545455	5,27272727

ANALYSE DE VARIANCE

Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*	*F*	*Probabilité*
Entre Groupes	5,39137719	3	1,79712573	0,52418075	0,66836743
A l'intérieur des groupes	126,852525	37	3,42844663		
Total	132,243902	40			

Vous constatez que dans notre expérience, la musique ne semble pas affecter l'apprentissage, car les moyennes sont très proches, et l'analyse de variance (F=0,524) n'est pas significative…(p = .669)

Explication et calcul de cette analyse de la variance entre groupes, pour un facteur :

F = **MC** « entre » / **MC** « inter » : soit le rapport entre la moyenne des carrés entre les groupes, et la moyenne des carrés à l'intérieur des groupes

Moyenne des carrés « entre groupes » = (somme des carrés / degré de liberté) entre les groupes
Moyenne des carrés « à l'intérieur » = (somme des carrés / degré de liberté) à l'intérieur des groupes

2- **Analyse de variance sur des moyennes d'échantillons appariés** (*Plan : Sn*Ap*)

Exemple : 11 sujets ont des troubles du sommeil importants, et acceptent de tester 3 traitements pharmacologiques différents. Chaque individu va utiliser un médicament durant une semaine. On comptabilise le nombre d'heures de sommeil, par nuit :

sujets	Médicament 1	Médicament 2	Médicament 3
1	2	0	3
2	4	1	4
3	2	1	3
4	2	2	4
5	1	0	1
6	3	2	5
7	4	2	11
8	4	2	10
9	10	3	9
10	8	6	14
11	2	2	5

Procédure : dans « Utilitaire d'analyse », cliquez «**analyse de variance : deux facteurs, sans répétition d'expérience** », et comparez…

Tableau affiché :

Analyse de variance: deux facteurs sans répétition d'expérience				
RAPPORT DÉTAILLÉ	*Nombre d'échantillons*	*Somme*	*Moyenne*	*Variance*
Ligne 1	4	6	1,5	1,66666667
Ligne 2	4	11	2,75	2,25
Ligne 3	4	9	2,25	0,91666667
Ligne 4	4	12	3	1,33333333
Ligne 5	4	7	1,75	4,91666667
Ligne 6	4	16	4	3,33333333
Ligne 7	4	24	6	15,3333333
Ligne 8	4	24	6	13,3333333
Ligne 9	4	31	7,75	10,25
Ligne 10	4	38	9,5	11,6666667
Ligne 11	4	20	5	18
Colonne 1	11	66	6	11
Colonne 2	11	42	3,81818182	7,76363636
Colonne 3	11	21	1,90909091	2,69090909
Colonne 4	11	69	6,27272727	16,6181818

ANALYSE DE VARIANCE						
Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*	*F*	*Probabilité*	*Valeur critique pour F*
Lignes	270	10	27	7,315270949	8,046E-06	2,16457963
Colonnes	138,272727	3	46,0909091	12,48768471	7,977E-05	2,92227753
Erreur	110,727273	30	3,69090909			
Total	519	43				

Vous constatez qu'il y a bien une différences entre les 3 médicaments et le F et très significatif (12,48)… (*et à mon avis, le médicament 2 était probablement un placebo !*)

Explication et calcul de cette analyse de variance pour deux facteurs, avec des mesures répétitives :
F= MC « *traitement* **» / MC «** *erreur* **»**
Dans cette situation « pairée », il y a trois calculs de sommes des carrés : la SC « sujets »(= « lignes »), la SC « traitements » (= « colonnes »), et la SC « interactions » (= « erreurs » dans l'interaction sujets **x** traitements)

Moyenne des carrés « traitement » = (somme des carrés / degré de liberté) des traitements
Moyenne des carrés « erreur » = (somme des carrés / degré de liberté) des erreurs

3- **Analyse de variance, à plan factoriel 2 x 2 (ou 2 x 3, etc…)** : (*plan : Sn<Ap x Bq>*)

Procédure : dans « Utilitaire d'analyse », cliquez « **analyse de variance : deux facteurs, avec répétition d'expérience** » (*Attention, c'est un peu plus compliqué : suivez bien les consignes !*).
En pratique, vous devez réaliser un tableau du type 2 x 2, placez les échantillons dans le tableau, mettez ce tableau dans « plage d'entrée) et indiquez le nombre d'échantillons par case dans « **nombre de lignes par échantillons** »...

<u>Limite</u> : ce calcul est possible uniquement dans le cadre des « plans équilibrés ». En pratique, cela veut dire que les cases doivent avoir le même nombre d'échantillons (= nombre de lignes par échantillons)…

***Exemple** : 36 futurs policiers, hommes ou femmes, ont tous passé des tests psychologiques avant de commencer leur stage. Après le première semaine d'effort, ils doivent décider s'ils arrêtent ou continuent leur stage. Exactement la moitié du groupe (hommes et femmes) décident d'arrêt. Nous avons donc 4 groupes équilibrés de 9 personnes, et le psychologue veut savoir si la décision est liée : 1er : au sexe, 2ème : à l'anxiété (en utilisant les résultats du test d'anxiété).*

	arrêt	poursuite
hommes	12	19
	19	18
	25	15
	21	18
	18	17
	22	15
	12	14
	20	17
	18	14
femmes	21	21
	20	12
	14	14
	15	16
	21	14
	21	10
	18	15
	20	10
	18	8

Résultat sur le tableau Excel :

Analyse de variance: deux facteurs avec répétition d'expérience

RAPPORT DÉTAILLÉ	arrêt	poursuite	Total
hommes			
Nombre d'échantillons	9	9	18
Somme	167	147	314
Moyenne	18,5555556	16,3333333	17,4444444
Variance	18,5277778	3,5	11,6732026
femmes			
Nombre d'échantillons	9	9	18
Somme	168	120	288
Moyenne	18,6666667	13,3333333	16
Variance	7	15,25	18
Total			
Nombre d'échantillons	18	18	
Somme	335	267	
Moyenne	18,6111111	14,8333333	
Variance	12,0163399	11,2058824	

ANALYSE DE VARIANCE						
Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*	*F*	*Probabilité*	*Valeur critique pour F*
Échantillon	18,7777778	1	18,7777778	1,69636136	0,20206088	4,14908641
Colonnes	128,444444	1	128,444444	11,6035132	0,0017919	4,14908641
Interaction	21,7777778	1	21,7777778	1,96737767	0,17035434	4,14908641
A l'intérieur du groupe	354,222222	32	11,0694444			
Total	523,222222	35				

Explication et calcul de cette analyse de variance pour deux facteurs, entre deux groupes :

Trois tests F sont proposées dans ce cas : le F de l'effet principal du facteur A (ici, le facteur « sexe »), le F de l'effet principal du facteur B (facteur « décision »), et le F de l'interaction A x B.
F_A = **MC** « A » / **MC** « *intérieur* »
F_B = **MC** « B » / **MC** « *intérieur* »
F_{AxB} = **MC** « A x B » / **MC** « *intérieur* »

Avec Moyenne de carrés « intérieur » = (somme des carrés / degré de liberté) des variations à l'intérieur du groupe. Notons que sous Excel, les 3 moyennes de carrés **MC** « A », **MC** « B » et **MC** « A x B » sont appelés respectivement « échantillon », « colonnes », et « interaction »

Dans cet exemple, on constate donc que **le facteur « décision » est dépendant au niveau d'anxiété**. Par contre les autres facteurs ne sont pas significatifs : il n'y a pas de différence entre les deux sexes…

C.2 **- les tests de corrélation :**

Un autre type d'analyse multivariée : **la régression linéaire**.

L'analyse de régression linéaire utilise la méthode des « moindres carrés » pour tracer une droite sur l'ensemble d'observations, et analyse l'incidence des variables indépendantes sur la variable dépendante unique. (*Par exemple, vous voulez savoir si le poids des individus varie en fonction de la taille, et de l'âge, etc*…)
Dans le cas d'une régression à deux variables, l'équation est donnée par $\hat{Y} = a + bX$
Avec Y = la variable de critère, X = la variable « de prédiction », a = la **constante** de régression, et b = la **pente**.
Si $\overline{X}$ correspond à la moyenne de X, ***r*** correspond au coefficient de corrélation, et *Sy* l'écart-type de Y, la constante a = $\overline{Y} - b\overline{X}$, et la pente b = $\frac{Sy}{Sx}$, cela donne finalement une formule pas trop complexe : $\hat{Y} = \overline{Y} - r\frac{Sy}{Sx}.\overline{X}$
(*Mais cela se complique beaucoup dans le cas d'une régression multiple, puisque l'équation de régression devient* $\hat{Y}$ *= b0 + b1X1+ b2 X2+…+BnXn ! Eh bien, sachez que le brave Excel peut prendre en compte jusqu'à 16 variables de prédiction…*)

Procédure : dans « Utilitaire d'analyse », cliquez « régression linéaire ». Indiquez les données pour la variable Y, et pour la (ou les) variable(s) X, et faites OK…
Les résultats affichés sont :
- le ***coefficient de détermination multiple*** (dans le cas à deux variables, cela correspond simplement au coefficient r de corrélation de Pearson)

- le ***coefficient de détermination*** **R^2** (*indiqué bizarrement en R^2 : voir les symboles de calcul d'Excel…* ***** *=multiplication, ^=puissance, etc.*) : il donne une idée du % de variabilité de la variable à modéliser, et plus le coefficient R^2 est proche de 1, plus il y a une corrélation et meilleur est le modèle… (*et le coefficient de détermination* R^2 ***ajusté*** *reflète, d'une façon plus fidèle, le degré de cette relation linéaire à la population…*)
- ***l'analyse de la variance*** : elle indique la régression (= le modèle) en indiquant le F de Fischer, et les « résidus ». Par exemple la régression correspond à la variation de « taille » qui s'explique par sa relation avec « le poids ». Et au contraire, les résidus (ou variation résiduelle) représente la variation de la « taille » qui ne peut s'expliquer par « le poids ».
Attention, cette ANOVA est particulière : elle teste si la moyenne de la variable à modéliser (le poids, par ex.) suffit à décrire les résultats obtenus… Bref, les variables explicatives apportent (ou non…) une quantité d'information significative au modèle. Si F est significatif, cela veut dire que la pente de la droite de régression diffère de 0, et donc nous admettons qu'il existe une relation linéaire significative entre le 2 (ou plus) variables.

-l'écart-type et le test de Student : pour la (ou les) variables X (le poids, et la taille, par ex.) en lien avec le modèle. Il faut considérer non le « t » de la constante, mais plutôt le « t » des variables X (. S'intéresser également aux limites (supérieures et inférieures) pour un seuil de confiance de 95%