Statistiques paramétriques et non paramétriques
Traitements sur des variables numériques

1 Travail sur des variables numériques avec Excel

1.1 Moyenne, écart type, moyennes par groupe

Fichier de travail : Conso-crustaces.xls

Moyenne, médiane, variance et écart type, maximum, minimum de la consommation, tous groupes confondus.

Fonctions à utiliser :

```
= MOYENNE(<plage de cellules>)
= MEDIANE(<plage de cellules>)
= ECARTYPE(<plage de cellules>)
= ECARTYPEP(<plage de cellules>)
= VAR(<plage de cellules>)
= VAR.P(<plage de cellules>)
= MAX(<plage de cellules>)
= MIN(<plage de cellules>)
```

Pour désigner une plage de cellules : par exemple : `A2:A513`

Pour obtenir les moyennes dans les différents groupes définis par la variable catégorisée SD Residence, par exemple, on peut utiliser les fonctions SOMME.SI et NB.SI dans des expressions telles que :

```
=SOMME.SI(B2:B513;"=1";A2:A513)/NB.SI(B2:B513;"=1")
```

N.B. Réfléchissez au comportement d'une telle formule vis-à-vis des valeurs manquantes.

1.2 Faire un regroupement en classes sur la variable Conso

1.2.1 Utiliser l'utilitaire d'analyse

L'utilitaire d'analyse permet d'obtenir un résultat tel que le suivant :

Classes Conso	Fréquence
5	238
10	108
20	91
40	47
60	13

80	6
ou plus...	7

On commence par saisir un tableau contenant les bornes supérieures des classes souhaitées, puis on choisit l'item Histogramme de l'utilitaire d'analyse.

1.2.2 Utiliser la fonction FREQUENCE

L'utilitaire d'analyse se sert en fait de la fonction FREQUENCE. Contrairement aux fonctions vues jusqu'à présent, cette fonction renvoie non pas un résultat, mais une plage de résultats. Dans Excel, une telle fonction est appelée *fonction matricielle*. La saisie se fait comme suit :

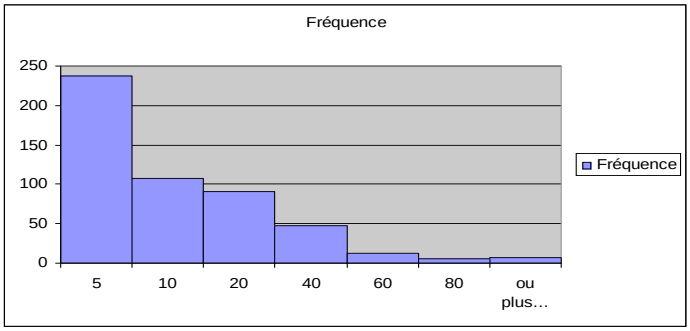
Comme précédemment, on saisit un tableau contenant les bornes supérieures des classes.
Par exemple, saisissons les bornes précédentes dans la plage I22 à I27.
On sélectionne ensuite une plage comportant une cellule de plus que la page des bornes de classes, et on saisit pour l'ensemble de ces cellules la formule :

```
=FREQUENCE(A2:A513;I22:I27)
```

On valide ensuite la saisie en appuyant simultanément sur Maj + Ctrl + Return.

1.2.3 Construire une représentation graphique des données

On peut facilement construire un diagramme à bandes à partir du tableau d'effectifs. Attention cependant : il ne s'agit pas d'un véritable histogramme puisqu'il n'est pas tenu compte de l'amplitude des classes.



2 Travail sur des variables numériques avec Statistica

Nous allons ici utiliser une partie des données fournies comme exemple avec Modalisa 4.6 : l'enquête "Habitudes alimentaires". Les deux seules variables numériques de cette enquête sont les variables 20 et 21 : dépense moyenne par personne lors d'un repas pris au Fast Food et nombre de repas pris au Fast Food par mois.

1 Nombre d'enfants
    1 UN enfant
    2 DEUX enfants

3 TROIS enfants et plus

2 Activité professionnelle
- 1 Avec activité profess.
- 2 Sans activité profess.

3 CSP
- 1 Prof. Lib. & Cadres sup.
- 2 Ouvriers

4 Origine géographique
- 1 Paris
- 2 Banlieue
- 3 Province
- 4 Autre

5 Age
- 1 de 25 à 34 ans
- 2 de 35 à 44 ans

6 Produit utilisé le plus souvent en weekend
- 1 Produits surgelés
- 2 Produits frais
- 3 Conserves

7 Produit utilisé le plus souvent en semaine
- 1 Produits surgelés
- 2 produits frais
- 3 conserves

8 Critère de choix pour les courses
- 1 Facilité de préparation
- 2 recherche d'équilibre
- 3 Fraîcheur ou durée de conservation

9 Repas du soir habituel avec les enfants
- 1 Oui
- 2 Non

10 Repas du soir tous les jours à la même heure
- 1 Oui
- 2 Non

11 Qualité principale d'une bonne alimentation
- 1 Nourrissante
- 2 Digeste
- 3 Légère
- 4 Savoureuse
- 5 Naturelle
- 6 Equilibrée

12 Changements perçus dans les habitudes alimentaires
- 1 Oui
- 2 Non
- 3 Pas tellement

13 Lieu de prise des Repas à la maison
- 1 Cuisine
- 2 Salle à Manger

14 Changement dans le contenu des repas
- 1 Oui
- 2 Non

15 Changement du nombre de plats
- 1 Plus de plats
- 2 Moins de plats

3 C'est égal

16 Changement dans l'abondance des portions
1 Plus copieux
2 Moins copieux
3 C'est pareil

17 Changement dans la durée des repas
1 Plus rapides
2 Moins rapides
3C'est égal

18 TV regardée pendant les repas
1 TV Oui
2 TV Non

19 Fréquentation des fast-food
1 Oui
2 Non

20 Si oui, dépense moyenne par personne
numérique, de 10 à 100

21 Nombre de repas pris en fast-food par mois
numérique de 0 à 50

2.1 Calcul des paramètres descriptifs

On peut calculer les paramètres descriptifs des deux variables 20 et 21 à l'aide du menu Statistiques - Statistiques élémentaires - Statistiques descriptives :

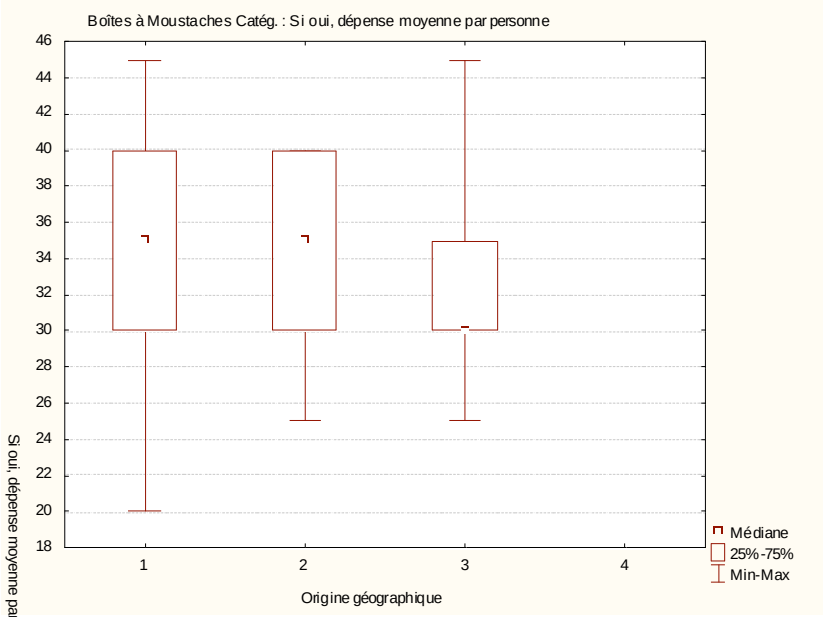
	Statistiques Descriptives (Habitudes-Alimentaires-Num dans Habitudes-					
Variable	N Actifs	Moyenne	Minimum	Maximum	Variance	Ecart-type
Si oui, dépense moyenne par personne	52	34,28846	20,00000	45,00000	36,91516	6,075785
Nombre de repas pris en fast-food par mois	51	5,94118	2,00000	15,00000	7,29647	2,701198

Pour obtenir ces paramètres pour chacun des groupes définis par les valeurs d'une variable catégorisée, on peut utiliser le menu : Statistiques - Statistiques élémentaires - Décompositions et ANOVA à un facteur, les onglets ANOVA puis Stats Descriptives. Par exemple :

Statistiques Descriptives par Groupes (Habitudes-Alimentaires-Num dans Habitudes-Alimentaires.stw) Plus petit N de ttes les vars51						
Origine géographique	Si oui, dépense moyenne par personne Moyennes	Si oui, dépense moyenne par personne N	Si oui, dépense moyenne par personne Ec-Type	Si oui, dépense moyenne par personne Variance	Si oui, dépense moyenne par personne Minimum	Si oui, dépense moyenne par personne Maximum
1	34,76667	30	6,452390	41,63333	20,00000	45,00000
2	34,06250	16	5,234103	27,39583	25,00000	40,00000
3	32,50000	6	6,892024	47,50000	25,00000	45,00000
4		0				
TsGrpes	34,28846	52	6,075785	36,91516	20,00000	45,00000

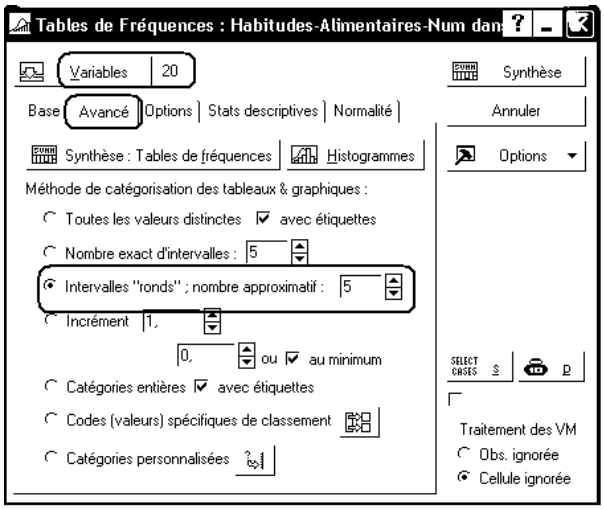
On peut également obtenir des graphiques de type "boîtes à moustaches" qui permettent assez bien de comparer les différents groupes du point de vue de l'une des variables numériques :

Par exemple, pour la dépense moyenne, selon l'origine géographique :



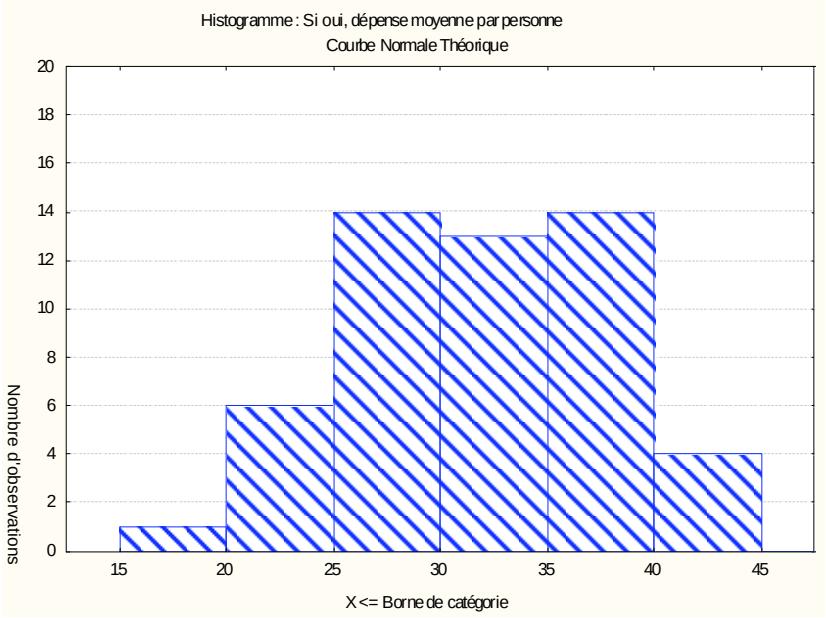
2.2 Répartition en classes d'une variable numérique

On peut utiliser le menu Statistiques - Statistiques élémentaires - Tables de fréquence. Par exemple, en complétant l'onglet Avancé comme suit :



on obtiendra les résultats suivants :

	Table de fréquences : Si oui, dépense moyenne par personne			
De... à...	Effectif	Effectifs Cumulés	%age	%age Cumulé
15,00000<x<=20,00000	1	1	0,37879	0,3788
20,00000<x<=25,00000	6	7	2,27273	2,6515
25,00000<x<=30,00000	14	21	5,30303	7,9545
30,00000<x<=35,00000	13	34	4,92424	12,8788
35,00000<x<=40,00000	14	48	5,30303	18,1818
40,00000<x<=45,00000	4	52	1,51515	19,6970
45,00000<x<=50,00000	0	52	0,00000	19,6970
50,00000<x<=55,00000	0	52	0,00000	19,6970
VM	212	264	80,30303	100,0000



2.3 Tests statistiques paramétriques de comparaison de moyennes

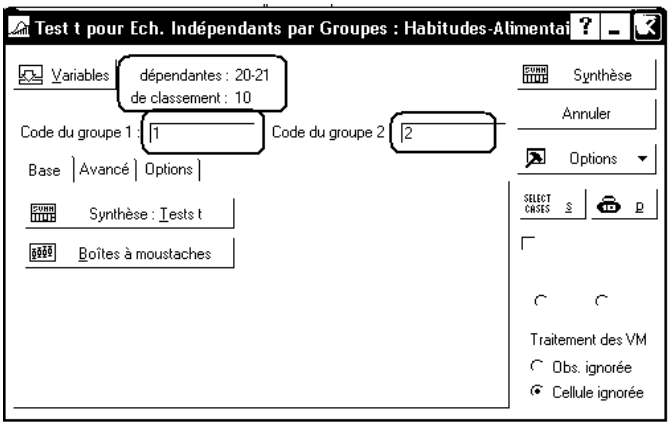
2.3.1 Comparaison de deux groupes indépendants

On considère les 52 questionnaires correspondant aux personnes ayant déclaré fréquenter les fast food. Ces 52 observations sont un échantillon de la population qui les fréquente. On souhaite étudier si les groupes définis par l'une ou l'autre des variables catégorisées ont des comportements différents du point de vue de la dépense par personne et de la fréquentation mensuelle, ou non.

Pour les variables catégorielles dichotomiques (CSP, Age, Repas du soir avec les enfants, Repas du soir tous les jours à la même heure, Lieu de prise de repas, TV regardée pendant les repas), on peut utiliser le **test T de Student sur des groupes indépendants**.

Exemple : on prend comme variable catégorisée "Repas du soir tous les jours à la même heure" et on effectue un test de Student sur chacune des deux variables numériques :

Menu Statistiques - Statistiques élémentaires - Test t pour échantillons indépendants (par groupe). Onglet : Avancé :



Variable	Tests t ; Classmt : Repas du soir tous les jours à la même heure Groupe1: 1 Groupe2: 2 Moyenne 1	Moyenne 2	Valeur t	dl	p	N Actifs 1	N Actifs 2	Ecart-Type 1	Ecart-Type 2
Si oui, dépense	35,30769	33,00000	1,36631	49	0,178079	26	25	5,026086	6,922187
Nombre de repa	5,08000	6,64000	-2,14417	48	0,037112	25	25	1,630951	3,251666

Lecture des résultats : Statistica a fait un test de Student pour chacune des deux variables numériques. La valeur de la statistique de test se trouve dans la colonne "Valeur t". La conclusion du test est obtenue à partir de la colonne "p" :

- pour la dépense par personne, la p-value est p = 0,17. La différence sur les moyennes n'est donc pas significative aux seuils traditionnels de 5% et de 1%.

- pour le nombre de repas par mois, la p-value est p=0,037=3,7%. La différence sur les moyennes des deux groupes est donc significative au seuil de 5%, sans l'être au seuil de 1%.

Dans un rapport d'analyse de l'enquête, on pourrait indiquer pour la première variable : *pas de différence selon la régularité des horaires de repas (t(49)=1,37 ; p>.10)* et pour la deuxième variable : *il semble que la fréquentation des fast food soit plus élevée dans les foyers où le repas du soir n'est pas pris à un horaire régulier (t(48)=-2,14 ; p=.037).*

Faites des tests analogues sur les autres variables dichotomiques. On constate que seule la variable relative à la régularité des horaires de repas semble produire des groupes ayant des comportements différents, et cela seulement pour la variable "Nombre de repas pris en Fast Food".

2.3.2 Principe général d'un test statistique

Un test statistique est une procédure permettant de faire un choix entre deux alternatives à partir des résultats observés sur un ou plusieurs échantillons. La démarche générale est la suivante :

1. Deux hypothèses, exclusives l'une de l'autre, sont formulées :
   - l'hypothèse nulle H0, qui exprime que les différences entre les groupes, ou entre les valeurs observées et des valeurs théoriques sont uniquement dues au hasard (fluctuations d'échantillonnage) ;
   - l'hypothèse alternative H1, qui exprime qu'au contraire, les différences observées sont significatives de "quelque chose".

2. On choisit ensuite un seuil de signification (traditionnellement : 5%, 1%, etc).

3. En fonction des données étudiées et de la forme des hypothèses, on choisit une statistique de test, c'est à dire une variable statistique, dont on peut calculer la valeur à partir des données observées sur les échantillons, et dont la loi statistique théorique sous l'hypothèse H0 est connue.

4. On calcule la valeur de cette statistique sur les données observées, ainsi que son niveau de significativité ou p-value, c'est-à-dire les chances que l'on avait, sous H0, d'obtenir une valeur au moins aussi extrême que celle observée.

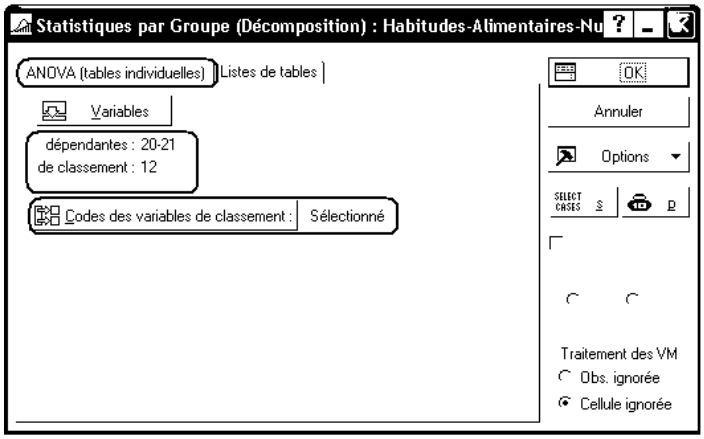
5. On compare ce niveau de significativité au seuil et on applique la règle de décision suivante :
   - Si la p-value est supérieure ou égale au seuil a, on retient l'hypothèse nulle ;
   - Si la p-value est inférieure au seuil a, on retient l'hypothèse alternative H1.

2.3.3 Comparaison de plusieurs groupes indépendants

Lorsqu'il s'agit de comparer les moyennes d'une variable numérique sur plus de deux groupes indépendants, le test paramétrique que l'on peut utiliser est l'analyse de variance à un facteur.

Par exemple, on veut comparer les moyennes de chacune des variables numériques dans les groupes définis par la variable "Origine géographique".

On peut utiliser le menu Statistiques - Statistiques élémentaires - Décompositions et ANOVA à 1 facteur en complétant le dialogue comme suit :



Il vaut mieux ne sélectionner que les origines géographiques 1, 2 et 3, aucune observation ne correspondant à l'origine géographique 4. Il est également pertinent d'afficher d'abord les statistiques par groupe, afin de s'assurer que les groupes ainsi définis ne sont pas d'effectifs trop faibles :

Statistiques Descriptives par Groupes (Habitudes-Alimentaires-Num dans Habitudes-Alimentaires.stw)
Plus petit N de ttes les vars51

Origine géographique	Si oui, dépense moyenne par personne Moyennes	Si oui, dépense moyenne par personne N	Si oui, dépense moyenne par personne Ec-Type	Nombre de repas pris en fast-food par mois Moyennes	Nombre de repas pris en fast-food par mois N	Nombre de repas pris en fast-food par mois Ec-Type
1	34,76667	30	6,452390	6,206897	29	2,820579
2	34,06250	16	5,234103	6,062500	16	2,694903
3	32,50000	6	6,892024	4,333333	6	1,751190
TsGrpes	34,28846	52	6,075785	5,941176	51	2,701198

Nous constatons que nous n'avons que 6 questionnaires correspondant à l'origine géographique 3, alors que nous en avons 29 ou 30 pour l'origine 1. Il aurait été préférable que les groupes soient à peu près équilibrés. Poursuivons cependant le test. On obtient (onglet ANOVA et Tests - bouton Analyse de variance) :

Analyse de la Variance (Habitudes-Alimentaires-Num dans Habitudes-Alimentaires.stw)
Effets significatifs marqués à p < ,05000

Variable	SC Effet	dl Effet	MC Effet	SC Erreur	dl Erreur	MC Erreur	F	p
Si oui, dépense moy	26,86891	2	13,43446	1855,804	49	37,87355	0,354719	0,703157
Nombre de repas pr	17,79408	2	8,89704	347,029	48	7,22978	1,230610	0,301165

La colonne "p" nous renseigne sur les résultats de ces deux tests : aucune différence significative entre les groupes n'a été mise en évidence.

2.4 Conditions d'application des tests paramétriques : normalité, homoscédasticité

Pour qu'il soit légitime d'utiliser les tests paramétriques tels que le T de Student ou l'ANOVA à un facteur, nos données doivent satisfaire certaines propriétés de régularité. Plus précisément :

- Les données observées dans les différents groupes doivent permettre de faire l'hypothèse que la variable numérique *est distribuée selon une loi normale* dans chacune des populations parentes ;

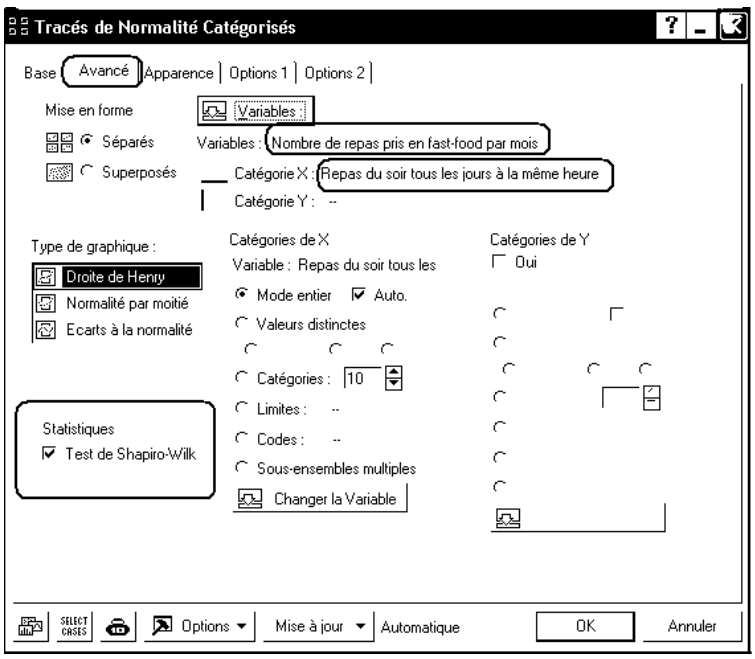
- Les données observées dans les différents groupes doivent permettre de faire l'hypothèse que la variable numérique a *la même variance dans chacune des populations parentes* (homoscédasticité des résidus).

On veut, par exemple, tester si ces conditions sont vérifiées pour les deux groupes définis par la variable catégorisée "Repas du soir pris toujours à la même heure".

Le menu "Statistiques" de Statistica ne permet pas de tester la normalité dans les différents groupes de manière simple. En revanche, le résultat peut être obtenu en annexe d'un graphique de normalité.

Pour étudier la normalité de la variable "Nombre de repas pris en fast-food par mois" dans les deux populations définies par les modalités de la variable "Repas du soir tous les jours à la même heure", on peut procéder de la manière suivante :

Menu : Graphiques - Graphiques catégorisés - Tracés de Normalité
Onglet Avancé ; indiquez les variables concernées et cochez la boîte "Test de Shapiro Wilk".



Le résultat est lu dans le cartouche en bas du graphique :

Repas du soir tous les jours à la même heure: 1 Nombre de repas pris en fast-food par mois: SW-W = 0,92; p = 0,0511
Repas du soir tous les jours à la même heure: 2 Nombre de repas pris en fast-food par mois: SW-W = 0,9099; p = 0,0303

Interprétation : l'hypothèse de normalité est l'hypothèse nulle H0. Au seuil de 5%, cette hypothèse peut être retenue dans le premier groupe (la p-value est de 5,11%), mais ne l'est pas dans le second groupe (p-value de 3%).

Il est plus facile d'obtenir un résultat concernant l'égalité des variances.

Pour un test de Student (comparaison de deux groupes), il suffit de cocher les boîtes "Test de Levene" et/ou "Test de Brown et Forsythe" dans l'onglet Options du dialogue relatif à ce test.

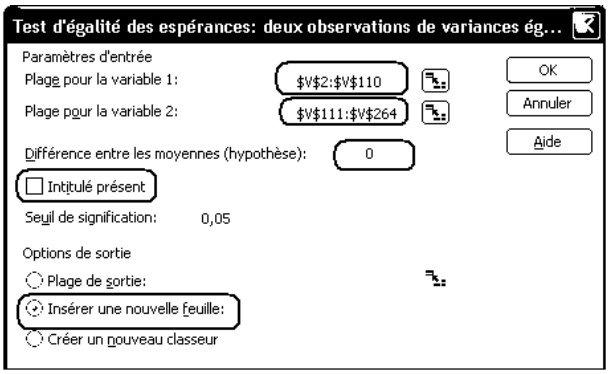
Pour une analyse de variance à un facteur (menu Statistiques - Décompositions et ANOVA à un facteur), il suffit d'utiliser les boutons "Test de Levene" et "Test de Brown et Forsythe". Comme dans le cas du test de normalité, c'est l'hypothèse H0 qui correspond à l'égalité des variances. Autrement dit, l'égalité des variances peut être retenue si la p-value renvoyée par le test est supérieure au seuil fixé (5% par exemple).

2.5 Tests de comparaison de moyennes avec Excel ou Modalisa

2.5.1 Tests de comparaison de moyennes avec Excel

Il est possible de réaliser un test de Student ou une analyse de variance avec Excel, bien que ce logiciel ne soit pas l'outil le mieux adapté.

Par exemple :
- Triez l'ensemble du tableau de données selon les valeurs de la variable "Repas du soir pris tous les jours à la même heure" (sélection des lignes de 2 à 265, puis menu Données - Trier).
- Menu Outils - Utilitaire d'analyse, puis l'item "Test d'égalité des espérances - Deux observations de variance égale" :



Le résultat s'affiche de la manière suivante :

Test d'égalité des espérances: deux observations de variances égales		
	Variable 1	Variable 2
Moyenne	5,08	6,769230769
Variance	2,66	10,58461538
Observations	25	26

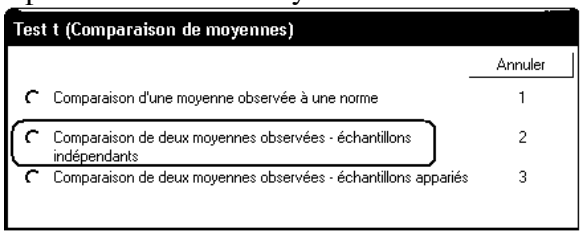
Variance pondérée	6,703171115	
Différence hypothétique des moyennes	0	
Degré de liberté	49	
Statistique t	-2,329273151	
P(T<=t) unilatéral	0,012005429	
Valeur critique de t (unilatéral)	1,676550893	
P(T<=t) bilatéral	0,024010858	
Valeur critique de t (bilatéral)	2,009575199	

Comme Statistica, Excel ignore les cellules vides (valeurs manquantes). Cependant, les résultats ne sont pas strictement identiques à ceux de Statistica, car une observation supplémentaire est prise en compte par Excel dans le groupe 2. Il faudrait déterminer l'origine de cette anomalie.

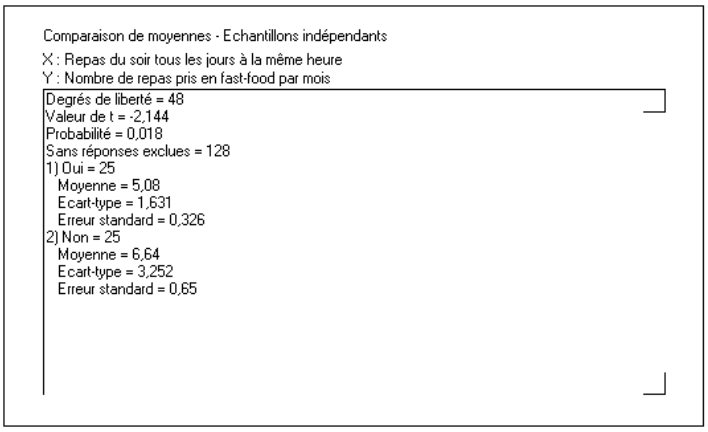
2.6 Tests de comparaison de moyennes avec Modalisa

L'enquête "Habitudes Alimentaires" est disponible comme exemple sur les versions de Modalisa installées dans nos salles.

Pour faire le test de Student précédent, on peut utiliser le menu : Analyse - Test t (comparaison de moyennes), puis l'item "Comparaison de deux moyennes observées - échantillons indépendants".



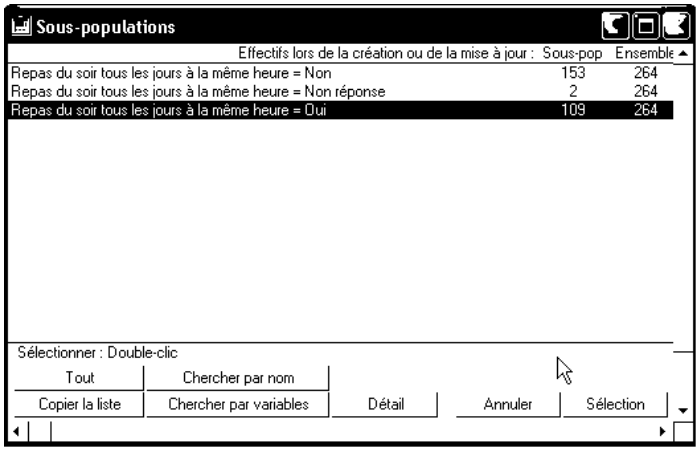
On sélectionne la question à réponses fermées n° 10, puis la question à réponses numériques n° 23 et on obtient comme résultat :



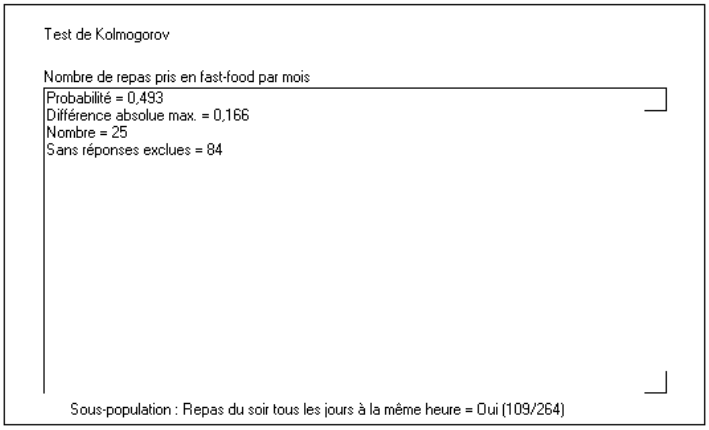
Ce résultat peut être enregistré dans un fichier texte à l'aide du menu Fichier - Export ASCII.

Modalisa permet également de tester la normalité des distributions parentes, à l'aide du menu Analyse - Test de Kolmogorov - Normalité. Cependant :
- L'utilisation directe du menu teste la normalité pour l'ensemble des données, ce qui présente peu d'intérêt ;
- Pour obtenir des tests de normalité sur les différents groupes définis par l'une des variables catégorisées, il faut définir des sous-populations (menu Classement - sous-populations), et

sélectionner l'une de ces sous-populations avant de réaliser le test. Par exemple, on sélectionne la sous-population correspondant à la réponse "oui" à la question 10 :



Le menu "Test de Kolmogorov - Normalité" donne alors, pour la variable "Nombre de repas pris en fast food " :



Notez que les résultats diffèrent ici sensiblement de ceux fournis par Statistica. Il semble en effet que Modalisa fasse un test de normalité de Kolmogorov-Smirnov, test qui n'est pas approprié au données que nous étudions ici.