Licence de Psychologie - Semestre N° 5 - TD n° 2
Intervalles de confiance, lois de distribution classiques et tests paramétriques avec Statistica

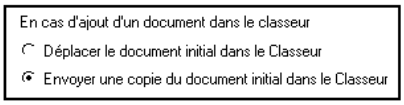
10 Organiser son espace de travail avec Statistica (complément)

10.1 Copier - coller entre classeurs, entre un classeur et un objet Statistica

Pour déplacer un objet d'un classeur à un autre, il suffit de déplacer son icône depuis le volet gauche du premier classeur dans le volet gauche du second. On peut également utiliser les menus locaux Copier et Coller obtenus à l'aide d'un clic droit dans le volet gauche de chaque classeur.

Le menu local "Insérer" du volet gauche d'un classeur permet également d'insérer dans ce classeur un document contenu dans une fenêtre indépendante. Il suffit de choisir les options : Document Statistica - Créer à partir d'une fenêtre.

L'opération faite par Statistica est soit une copie (l'original de l'objet est conservé) soit un déplacement (l'original de l'objet n'est pas conservé) selon le paramétrage choisi dans le menu Outils - Options - Onglet Classeurs - Item "En cas d'ajout d'un document dans le classeur".



10.2 Supprimer un objet d'un classeur

Il est également possible de supprimer un objet d'un classeur, à l'aide d'un clic droit et de l'item de menu Supprimer. Cela permet notamment de ne garder, pour un traitement donné, que le résultat le plus abouti. Attention cependant : lorsque l'on supprime un objet qui n'est pas une feuille de la hiérarchie, on supprime en même temps tous les objets qui en dépendent.

11 Travail sur des données recensées : statistiques descriptives sur un tableau d'effectifs

11.1 Introduire une pondération dans la feuille de données

On considère l'exemple suivant :
Dans le cadre d'une analyse médicale, deux méthodes de dosage peuvent être utilisées. A partir d'un même prélèvement, on répète 25 fois la méthode A et 30 fois avec la méthode B. Les résultats sont rassemblés dans les tableaux rassemblés dans le classeur Dosages.stw.

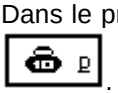
Ouvrez le classeur Dosages.stw et observez la façon dont les données ont été saisies dans les feuilles Méthode A, Méthode B et Ensemble.

Contrairement aux exemples traités précédemment, les données sont ici présentées sous la forme de tableaux recensés : les observations ont fait l'objet d'un tri à plat. Par exemple, la valeur 42 a été obtenue 7 fois comme résultat de mesure par la méthode A.

Comment effectuer les traitements de Statistiques descriptives sur des données structurées sous cette forme, par exemple dans la feuille "Méthode A" ?

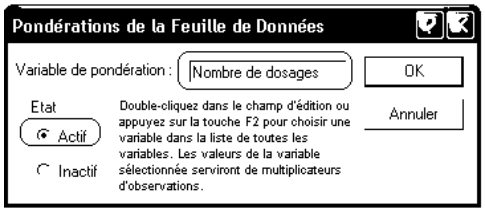
Il faut indiquer à Statistica que la colonne "Nombre de dosages" est une colonne d'effectifs ou **pondérations** des données.

Les pondérations peuvent aussi bien être définies comme propriété de la feuille elle-même que comme propriété de l'une des analyses.

Dans le premier cas, on affiche la feuille de données et on utilise le bouton "pondérations" de la barre d'outils :  . Les pondérations s'appliquent alors à toutes les analyses utilisant cette feuille.

Dans le deuxième cas, on utilise l'un des items du menu Statistiques et on clique sur le bouton "pondérations"  de la fenêtre de dialogue. Les pondérations ne concerneront alors que l'analyse en cours.

Ici, rendez active la feuille "Méthode A" et indiquez que la variable 2 (Nombre de dosages) est la variable de pondération :

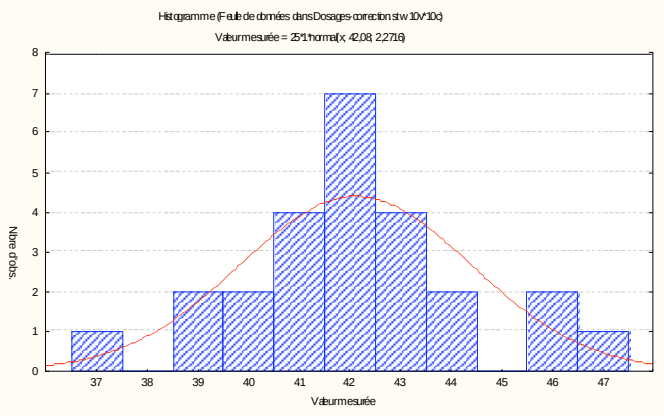


Le reste du traitement peut alors être réalisé de la même façon que pour un tableau-protocole. On obtient par exemple:

	Statistiques Descriptives (Méthode A dans Dosages.stw)				
Variable	N Actifs	Moyenne	Minimum	Maximum	Ecart-type
Valeur mesurée	25	42,08000	37,00000	47,00000	2,271563

Remarquez que le nombre d'observations pris en compte n'est pas égal à 9 (nombre de lignes du fichier) mais à 25 (somme des effectifs contenus dans la deuxième colonne).

De même, réalisez à l'aide du menu Graphiques, l'histogramme suivant :

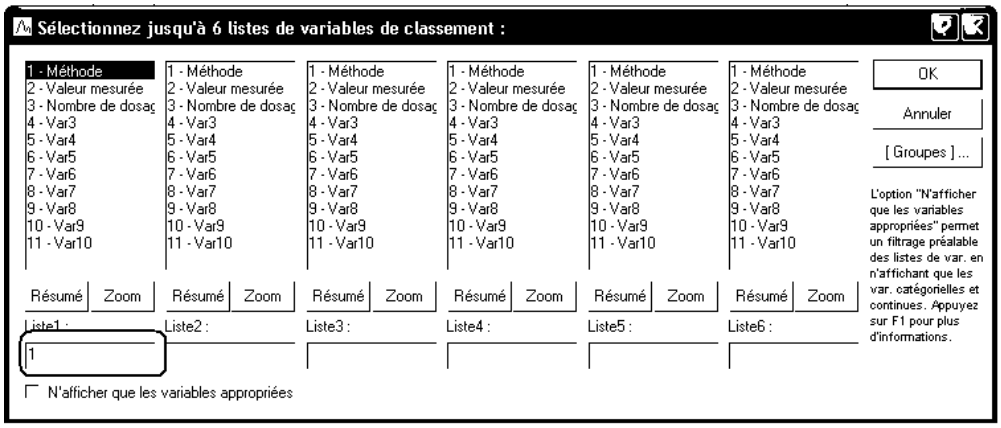


11.2 Calculer des paramètres de statistiques descriptives pour des données structurées "par groupe"

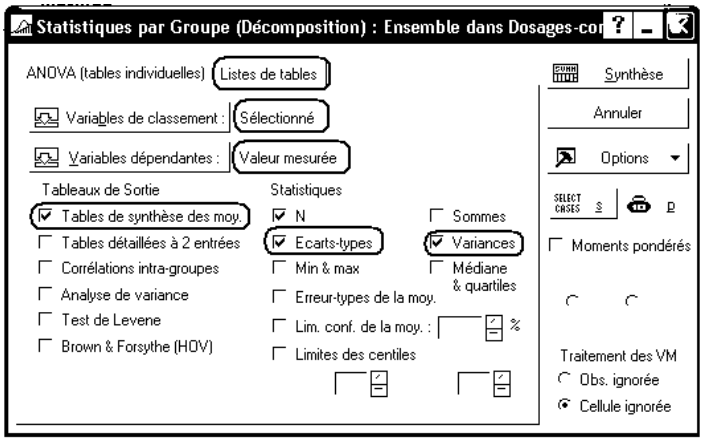
Nous souhaitons calculer la moyenne, la variance et l'écart type de la variable "Valeur mesurée" pour chacune des méthodes, et obtenir les résultats dans une même feuille de résultats.

Rendez active la feuille "Ensemble" et définissez la variable "Nombre de dosages" comme variable de pondération.

Utilisez ensuite le menu Statistiques - Statistiques Elémentaires - Décompositions & ANOVA à un facteur. Activez l'onglet "Listes de tables". Choisissez la variable "Méthode" comme variable de classement :



Choisissez ensuite la variable "Valeur mesurée" comme variable dépendante :



En cliquant sur le bouton Synthèse, on obtient les résultats suivants :

Statistiques Descriptives par Groupes (Ensemble)
N=55 (pas de VM dans les vars dépendante

Valeur mesurée Moyennes	Valeur mesurée N	Valeur mesurée Ec-Type	Valeur mesurée Variance
42,08000	25	2,271563	5,160000
42,10000	30	1,398275	1,955172
42,09091	55	1,828506	3,343434

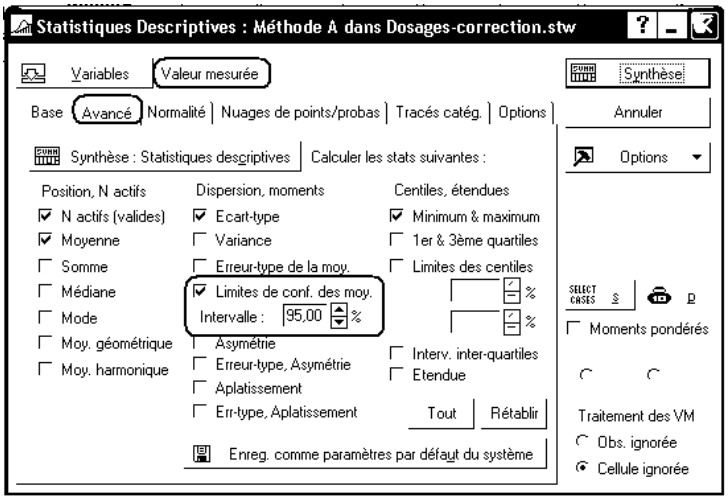
12 Intervalles de confiance

12.1 Intervalle de confiance pour une moyenne

Les menus Statistiques - Statistiques Elémentaires - Statistiques descriptives et Statistiques - Statistiques Elémentaires - Décompositions & ANOVA à un facteur permettent également d'obtenir une estimation de la moyenne d'une variable numérique par un intervalle de confiance avec un degré de confiance donné.

Ainsi, à partir de l'échantillon de mesures réalisées par la méthode A, quel intervalle estimant la "vraie valeur" de la substance dosée peut-on donner avec un degré de confiance de 95% ?

Rendez active la feuille "Méthode A" et utilisez le menu Statistiques - Statistiques Elémentaires - Statistiques descriptives en complétant la fenêtre de dialogue comme ci-dessous :



Vous devriez obtenir les résultats suivants :

	Statistiques Descriptives (Méthode A dans Dosages-correction.stw)						
Variable	N Actifs	Moyenne	Confiance -95,000%	Confiance +95,000%	Minimum	Maximum	Ecart-type
Valeur mesurée	25	42,08	41,14	43,02	37,00	47,00	2,27

Autrement dit, on estime, avec un degré de confiance de 95%, que la vraie valeur de la quantité à doser est comprise entre 41,14 et 43,02.

Exercice : De même, comparez les intervalles de confiance obtenus par les deux méthodes en utilisant la feuille "Ensemble" et le menu Statistiques - Statistiques Elémentaires - Décompositions & ANOVA à un facteur. Vous deviez obtenir :

Statistiques Descriptives par Groupes (Ensemble dans Dosages-correction.stw) N=55 (pas de VM dans les vars dépendante				
Valeur mesurée Moyennes	Confiance -95,000%	Confiance +95,000%	Valeur mesurée N	Valeur mesurée Ec-Type
42,08000	41,14234	43,01766	25	2,271563
42,10000	41,57788	42,62212	30	1,398275
42,09091	41,59659	42,58522	55	1,828506

Enregistrez le classeur Dosages.stw et refermez-le.

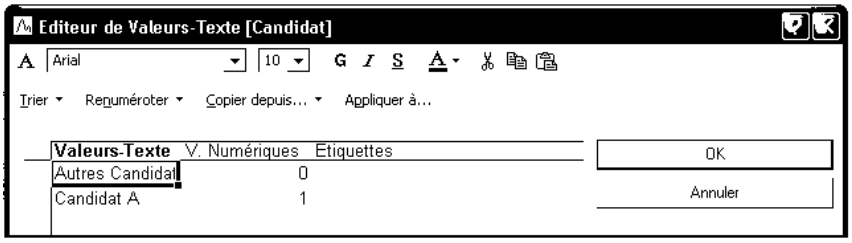
12.2 Intervalle de confiance pour une proportion

Lors d'un sondage électoral, on interroge au hasard 1000 personnes. 535 personnes déclarent vouloir voter pour le candidat A pendant que 465 déclarent vouloir voter pour un autre candidat. Quel intervalle de confiance, avec un degré de confiance de 95%, peut-on donner concernant le score du candidat A ?

Pour répondre à cette question, on peut :
- Créer la feuille de données suivante :

	1 Candidat	2 Suffrages
1	1	535
2	0	465

- Au besoin définir des étiquettes de texte pour la variable "Candidat", de façon que la feuille affiche "Candidat A" et "Autres Candidats". Pour cela, faire un double-clic sur la tête de la colonne "Candidat", puis utiliser le bouton "Valeurs-Texte" :



	1 Candidat	2 Suffrages
1	Candidat A	535
2	Autres Candidats	465

- Définir la variable "Suffrages" comme variable de pondération.

- Déterminer l'intervalle de confiance comme précédemment :

	Statistiques Descriptives			
Variable	N Actifs	Moyenne	Confiance -95,000%	Confiance +95,000%
Candidat	1000	0,5350	0,5040	0,5660

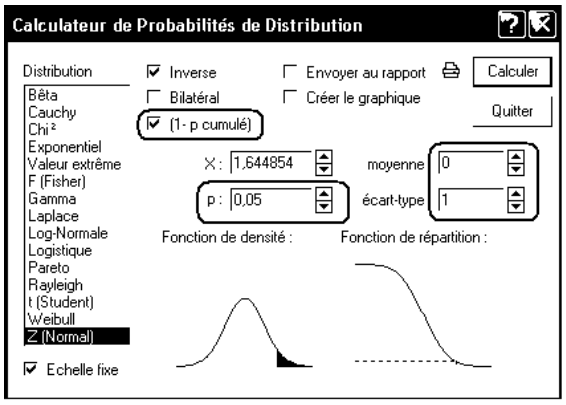
Au vu de l'intervalle trouvé, il semblerait que l'on puisse affirmer, avec un degré de confiance de 95%, que le candidat A sera élu...

13 Lois statistiques classiques

Le menu Statistiques - Calculateur de Probabilités - Distributions permet d'une part de trouver une valeur critique ou un niveau de significativité pour les lois statistiques continues usuelles, soit de réaliser des représentations graphiques de la densité ou de la fonction de répartition de ces lois.

13.1 La loi normale centrée réduite

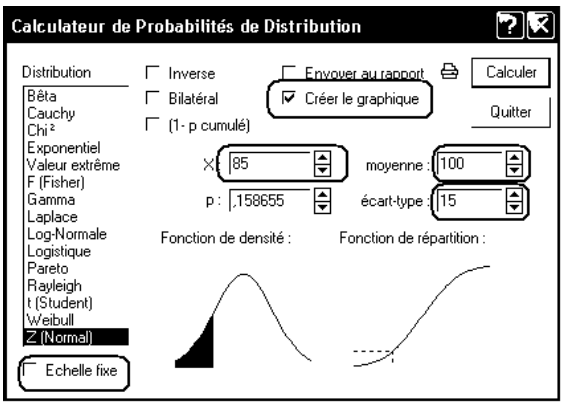
Quelle est la valeur de Zcritique pour un test unilatéral à 5% ?
Compléter le dialogue comme suit :



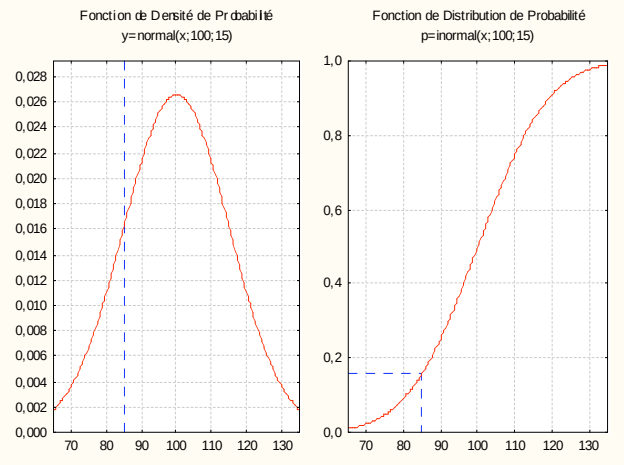
La réponse doit être lue dans la zone d'édition "X: _____". C'est ici : Zc=1,644854.

13.2 Représenter graphiquement la densité d'une loi normale quelconque

On veut représenter la densité de la loi normale de paramètres m = 100 et s=15 pour X compris entre 70 et 130
Complétez la fenêtre de dialogue comme suit, en veillant à ce que la boite "Echelle Fixe" ne soit pas cochée :

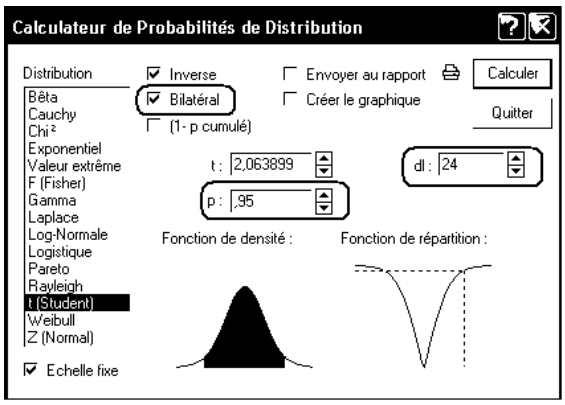


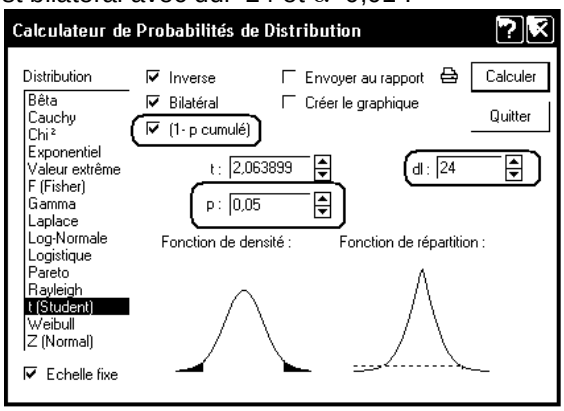
Remarquez que Statistica exige que l'un des deux champs "X:" ou "p:" soit complété, et repère la valeur correspondante sur le graphique. On obtient ainsi un graphique ayant l'allure suivante :



13.3 Loi de Student

Calculez la valeur critique de la loi de Student pour un test bilatéral avec ddl=24 et α=0,01 :





Remarquez que les paramètres peuvent être entrés de deux façons différentes :

- p = 0,95 si l'on ne coche pas la boîte "(1-p cumulé)
- p = 0,05 si l'on coche cette boîte.

Lecture du résultat : dans les deux cas, on obtient t_{crit}=2,063899.

On a réalisé un test de Student, avec un nombre de degrés de liberté égal à 58. La valeur observée de la statistique est t_{obs}=2,54. Quel est le niveau de significativité du résultat obtenu pour un test unilatéral ? pour un test bilatéral ?
Réponses : p=1,38% pour un test bilatéral et p=0,69% pour un test unilatéral.

13.4 Loi du khi-2

Selon les habitudes américaines, cette loi est désignée par Chi^2.
Déterminez la valeur critique du khi-2 pour un seuil de 5% et 6 ddl.
Vous devriez trouver : Khi-2 = 12,59.

Réaliser un graphique de la densité de la loi du khi-2 à 1 degré de liberté, en ajustant les échelles de manière à faire apparaître clairement la valeur critique correspondant à un seuil de 5% : 3,94.
Là, le choix d'échelle fait par Statistica est plus que discutable (que l'on ait, ou non, coché la case "Echelle Fixe"). Pour imposer notre choix d'échelle :

- Cliquez sur le graphique avec le bouton droit de la souris
- Sélectionnez l'item de menu Propriétés du Graphique (toutes options)
- Affichez l'onglet Axe -Echelle
- Indiquez pour l'axe X une plage de variation de 0,00015 à 5
- Indiquez pour l'axe Y une plage de variation de 0 à 1,5 (par exemple).

14 Tests paramétriques classiques

14.1 Test de comparaison d'une moyenne à une norme

Ouvrez le classeur ADD.stw et consultez la présentation des données qui se trouve dans le rapport contenu dans ce classeur.

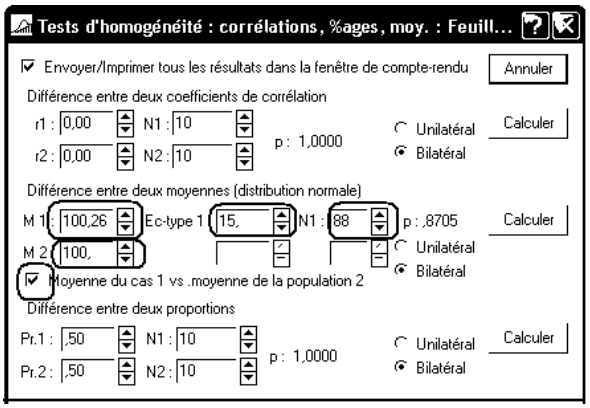
On se pose la question suivante : la population étudiée diffère-t-elle significativement de la population générale du point de vue du QI ?

Utilisez le menu Statistiques - Statistiques Elémentaires, puis l'item Comparer une moyenne à un standard.
Sélectionnez la variable IQ et indiquez 100 comme valeur de référence. Vous devriez obtenir le résultat suivant :

	Comparaison de moyennes à un standard (constante) (ADD.sta)							
Variable	Moyenne	Ec-Type	N	Erreur-T	Valeur de Référence	Valeur t	dl	p
IQ	100,2614	12,98496	88	1,384201	100,0000	0,188819	87	0,850674

Lecture du résultat :
La moyenne observée sur l'échantillon de 88 sujets est de 100,16. Statistica compare cette valeur à la valeur de référence (100), à l'aide d'un test de Student, avec 87 degrés de liberté. La statistique de test vaut t=0,189, ce qui correspond à un niveau de significativité de 85%. Autrement dit, rien n'indique une différence entre la population étudiée et la population générale du point de vue du QI.

Remarque : Dans le test ci-dessus, Statistica utilise l'estimation de l'écart type des QI faite à partir de l'échantillon, et non l'écart type de la distribution des QI dans la population (15). Il est possible d'introduire l'écart type de la population, à condition d'utiliser le menu Statistiques - Statistiques Elémentaires - Tests d'homogénéité et de remplir la fenêtre de dialogue comme suit :



14.2 Test de comparaison de deux moyennes sur des groupes indépendants

14.2.1 Données saisies "par sujet"

Lorsque la saisie a été faite correctement, la feuille de données Statistica rassemble sur une même ligne les observations relative à un même individu statistique. Ainsi, la saisie des observations relatives à un plan S<A> comportera au moins 2 colonnes :

- Une colonne "Groupe" ou "Condition expérimentale", avec, comme variable nominale, les différents niveaux du facteur A

- Une colonne "Variable dépendante".

	1 Groupe	2 VD
1	1	5,7
2	1	4,8
3	2	7,6
4	2	6,5

Dans ce cas, on utilisera le menu Statistiques - Statistiques élémentaires - Test t pour échantillons indépendants - par groupes.

Exemple : Le score ADDSC est-il significativement différent pour les garçons et les filles dans la population étudiée ?

Utilisez le menu Statistiques - Statistiques élémentaires - Test t pour échantillons indépendants - par groupes. Indiquez ADDSC comme variable dépendante et GENDER comme variable de classement.
Statistica devrait produire le résultat suivant :

	Tests t ; Classmt : GENDER (ADD.sta) Groupe1: 1 Groupe2: 2										
Variable	Moyenne 1	Moyenne 2	Valeur t	dl	p	N Actifs 1	N Actifs 2	Ecart-Type 1	Ecart-Type 2	Ratio F Variances	p Variances
ADDSC	54,29091	49,78788	1,662895	86	,099974	55	33	12,90230	11,20479	1,325949	0,395292

Lecture des résultats :
Statistica fait un test t de Student. La valeur observée de la statistique t est t=1,66, ce qui correspond à un niveau de significativité de presque 10%. Bien que Statistica ne le mentionne pas, il fait ici un test bilatéral. Autrement dit, les différences ne sont pas significatives au seuil de 5% bilatéral.

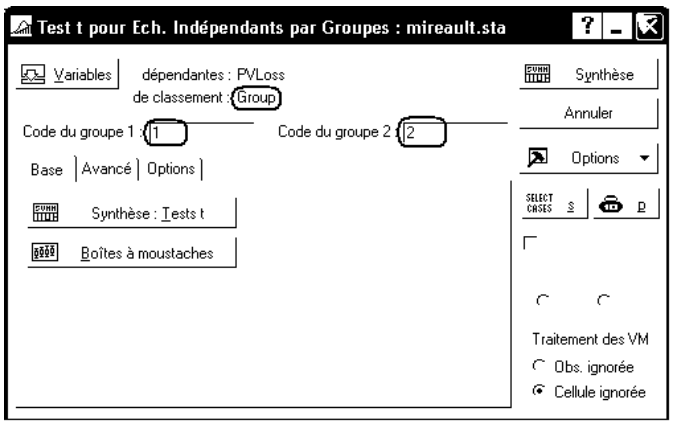
Insérez si besoin les fenêtres de résultats dans le classeur, enregistrez le classeur ADD.stw sur votre compte et refermez-le.

14.2.2 Données saisies "par sujet" : test sur une partie des observations

Lorsque le facteur A comporte plus de deux niveaux, il peut être utile de comparer les observations faites pour deux niveaux particuliers.

Reprenons, par exemple, les données Mireault. La variable Group comporte 3 modalités (codées 1, 2 et 3). Nous souhaitons comparer les scores de la variable PVLoss sur les groupes 1 et 2.

Chargez le classeur Mireault.stw, puis utilisez le menu Statistiques - Statistiques élémentaires - Test t pour échantillons indépendants - par groupes. Indiquez PVLoss comme variable dépendante, et Group comme variable de classement. Indiquez ensuite les deux niveaux choisis pour la variable groupe, comme ci-dessous :



Vous devriez obtenir le résultat suivant :

Variable	Tests t ; Classmt : Group (mireault.sta) Groupe1: 2 Groupe2: 1 Moyenne 2	Moyenne 1	Valeur t	dl	p	N Actifs 2	N Actifs 1	Ecart-Type 2	Ecart-Type 1	Ratio F Variances	p Variances
PVLoss	17,79	22,21	-6,44	320	0,000000	182	140	5,64	6,68	1,40	0,03

Remarque : Le test de Fisher sur l'égalité des variances conclut ici sur une différence significative des deux variances. On pourra donc recommencer le test en utilisant la boîte à cocher "Estimation séparée des variances".

Enregistrez le classeur sur votre compte et refermez-le.

14.2.3 Données saisies par variable

Statistica permet également de réaliser un test de comparaison de moyennes sur deux groupes indépendants lorsque les observations de la VD ont été saisies dans deux colonnes différentes.

Exemple : On reprend l'énoncé suivant :
Lors d'une expérience pédagogique, on s'intéresse à l'effet comparé de deux pédagogies des mathématiques chez deux groupes de 10 sujets:

- pédagogie traditionnelle : Gr1
- pédagogie moderne : Gr2.

On note la performance à une épreuve de combinatoire.
Ces données expérimentales permettent-elles d'affirmer que la pédagogie a un effet sur les résultats à l'épreuve de combinatoire?

Définissez un nouveau classeur contenant une feuille de calcul et saisissez dans cette feuille les données suivantes :

	1 VD-Gr1	2 VD-Gr2
1	5	4
2	4	5,5
3	1,5	4,5
4	6	6,5
5	3	4,5
6	3,5	5,5
7	3	1
8	2,5	2
9	1,5	4,5
10	2,5	4,5

Utilisez ensuite le menu Statistiques - Statistiques Elémentaires - Test t pour échantillons indépendants, par variables. Vous devriez obtenir le résultat suivant :

	Test t pour des Echantillons Indépendants (Feuille de données35) Note : Variables traitées comme des échantillons indépendants								
Groupe1 vs. Groupe2	Moyenne Groupe 1	Moyenne Groupe 2	valeur t	dl	p	N Actifs Groupe 1	N Actifs Groupe 2	Ec-Type Groupe 1	Ec-Type Groupe 2
VD-Gr1 vs. VD-Gr2	3,25	4,25	-1,45	18	0,16	10	10	1,438556	1,637240

Enregistrez le classeur contenant la feuille de données et les résultats du traitement.

14.2.4 Construire une variable calculée pour définir les deux groupes

On reprend le classeur ADD.stw

La médiane de la variable ADDSC est égale à 50. On souhaiterait définir deux groupes en utilisant la position de l'observation par rapport à la médiane, et comparer ces deux groupes du point de vue de la variable GPA.

Ajoutez une colonne supplémentaire au tableau de données et définissez une variable calculée à l'aide de la formule :

```
=iif(v2 >=50;1 ; 0)
```

ou, de manière équivalente, mais supportant mieux d'éventuelles modifications ultérieures :

```
=iif('ADDSC' >= 50; 1 ; 0)
```

N.B. iif est une fonction ("si immédiat") de Statistica, analogue à la fonction SI d'Excel.
Réalisez ensuite le test sur la variable GPA pour les deux groupes ainsi définis.

15 Test de comparaison de deux moyennes sur des groupes appareillés

On reprend l'exemple ADD.dat.
Les variables ENGG et GPA représentent les résultats obtenus en Anglais d'une part, et la moyenne des points obtenus en 9è année d'autre part. Ces résultats utilisent la même échelle de notation : le score varie de 0 à 4. Nous souhaiterions savoir si, dans la population étudiée, les scores en anglais sont égaux à la moyenne des scores.

Utilisez ensuite le menu Statistiques - Statistiques Elémentaires - Test t pour des échantillons appariés.
Sélectionnez ENGG comme variable pour constituer la première liste, et GPA comme variable pour constituer la seconde. Vous devriez obtenir le résultat suivant :

	Test t pour des Echantillons Appariés (ADD.sta) Différences significatives marquées à p < ,05000							
Variable	Moyenne	Ec-Type	N	Différ.	Ec-Type Différ.	t	dl	p
ENGG	2,6591	0,9455						
GPA	2,4562	0,8614	88	0,2028	0,5188	3,6677	87	0,0004

Enregistrez le classeur ADD.stw sur votre compte puis refermez-le.

16 Tests sur les proportions

16.1 Test de comparaison de deux proportions sur des groupes indépendants

16.1.1 Variables dichotomiques données par des effectifs

Deux échantillons provenant de deux populations différentes ont passé un test commun.
Dans le premier groupe, d'effectif 150, le taux de succès a atteint 68%. Autrement dit, 102 sujets ont passé le test avec succès, et 48 ont échoué.
Dans le deuxième groupe, d'effectif 180, le taux de succès a atteint 55,5%. Autrement dit, 100 sujets ont passé le test avec succès et 80 ont échoué.

Peut-on dire que la seconde population réussit l'épreuve moins facilement que la première ?

Statistica ne comporte pas de module spécifiquement destiné à traiter ce genre de situation. Cependant, la comparaison de deux proportions n'est qu'un cas particulier de la comparaison de deux moyennes.
Nous allons donc saisir nos données dans une feuille de calcul Statistica, de la façon suivante :

	1 Groupe	2 Resultat	3 Effectifs	V
1	A	Succès	102	
2	A	Echec	48	
3	B	Succès	100	
4	B	Echec	80	
5				

Pour faciliter l'interprétation des résultats, il sera commode de personnaliser le codage de la variable "Resultat" Succès sera codé 1 tandis que Echec sera codé 0. Vous allez donc définir "Succès" comme valeur-texte correspondant à la valeur numérique 1 et "Echec" comme valeur-texte correspondant à la valeur numérique 0. Pour cela, double-cliquez sur la tête de la colonne "Resultat", puis utilisez le bouton "Valeurs-Texte". Lors de la saisie, saisissez de préférence 1 et 0 plutôt que "Succès" et "Echec" dans cette colonne.

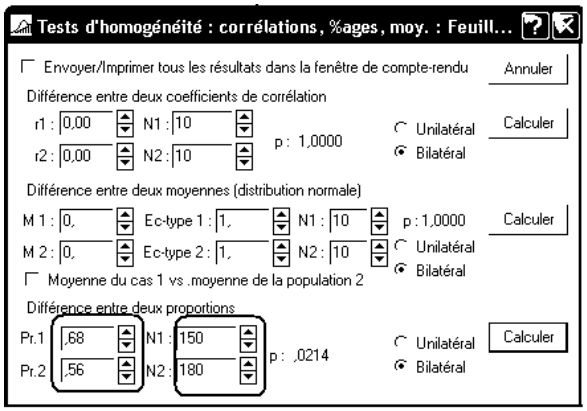
N'oubliez pas d'indiquer à Statistica que la colonne Effectifs est une colonne de pondérations.

Statistica fournit les résultats suivants :

	Tests t ; Classmt : Groupe (Feuille de données1) Groupe1: A Groupe2: B				
Variable	Moyenne A	Moyenne B	Valeur t	dl	p
Resultat	0,680000	0,555556	2,321956	328	0,020847

On retrouve ainsi les valeurs des proportions observées (68% et 55,55%). La valeur observée de la statistique de test est 2,32. La formule de calcul n'est pas exactement la même que celle donnée en cours, mais la différence est très faible (2,3219 ici, alors que la formule du cours donnerait 2,3201). Statistica utilise une loi de Student avec 328 ddl. Vu le nombre de degrés de liberté, cette loi est très proche d'une loi normale, et le résultat produit est tout à fait correct.

On peut aussi utiliser le menu Statistiques - Statistiques Elémentaires - Tests d'homogénéité. On complète alors la fenêtre comme suit :



Rappel des formules du cours :

$$p = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

$$Z = \frac{f_1 - f_2}{E} \text{ avec } E^2 = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Exercice. 1) Faites varier les taux de succès dans les deux groupes. Que devient le résultat du test lorsque les taux varient ?
2) Faites varier les effectifs dans les deux groupes. Avec les taux de succès indiqués, quelles sont les tailles minimales des échantillons permettant d'obtenir un résultat significatif à 5% ?

16.1.2 Variables dichotomiques données par leur protocole

Dans l'exemple traité précédemment, le protocole est un tableau de données à 2 colonnes et 330 lignes. Il est en fait assez facile de constituer à l'aide de Statistica un tableau répétitif de ce type.
- Ouvrez une nouvelle feuille de données de 330 observations, pour 2 variables.
- Pour la première variable, spécifiez le type "texte".
- Pour la seconde variable, spécifiez le type "Numérique", avec, comme valeurs-texte, "Succès" pour la valeur numérique 1, et Echec pour la valeur numérique 0.
- Dans la première variable, saisissez l'identifiant du premier groupe ("A" par exemple) en ligne 1, puis utilisez ensuite le menu Edition - Remplir / Centrer-Réduire le bloc - Remplir/Copier vers le bas pour recopier cette valeur jusqu'en ligne 150.
- Saisissez l'identifiant du second groupe en ligne 151, puis recopiez de même jusqu'en ligne 330
- Dans la seconde variable, saisissez la valeur 1, ou le texte Succès en ligne 1 et recopiez jusqu'en ligne 102.
- De même pour le texte Echec de la ligne 103 à la ligne 150.
- Continuez en saisissant "Succès" de la ligne 151 à la ligne 250, puis "Echec" de la ligne 251 à la ligne 330.

Réalisez ensuite un test de comparaison de moyennes en utilisant ces deux variables. Vous devriez retrouver les résultats précédents.