UNIVERSITE GASTON BERGER
L'excellence au service du développement

UFR des Lettres et Sciences humaines
Section de sociologie
Semestre 1, 2024

SOCIO 532 EC2
STATISTIQUES & INFORMATIQUE APPLIQUÉES AUX SCIENCES SOCIALES

© El Hadj Touré, Ph.D.

LABO SPSS #1
Mesures de tendance centrale et de variation, estimation par intervalle de confiance

Dans le labo précédent, nous avions appris à procéder à l'analyse descriptive à l'aide des distributions de fréquences et de pourcentages afin d'apprécier l'ampleur d'un phénomène social. Ces statistiques descriptives sont surtout appropriées pour des variables qualitatives ou des variables quantitatives dont les valeurs sont transformées en classes ou catégories. Que se passe-t-il lorsque l'on veut décrire les données d'une variable quantitative pour laquelle on souhaite conserver les valeurs métriques ? Dans cette situation, on utiliser les mesures de tendance centrale et de variation.

L'analyse descriptive est très utile à des fins exploratoires. Il en est ainsi lorsqu'on explore un phénomène nouveau ou peu étudié comme la Covid-19. Ou lorsqu'il s'agit de décrire les caractéristiques d'un problème social. Toutefois, les études quantitatives uniquement descriptives sont plutôt rarement rencontrées en sciences sociales. Dans bien des cas, après avoir réduit les données en informations claires, on cherche à généraliser ces informations à la population dont provient l'échantillon. C'est l'objet de l'analyse inférentielle. Elle consiste à inférer à la population de référence les résultats obtenus. L'inférence statistique obéit donc au principe de généralisation des informations, et cherche à évaluer l'incertitude associée à cette généralisation. Par exemple, il faut s'assurer que le revenu moyen des 1000 adultes sénégalais sondés est statistiquement significatif au niveau de toute la population étudiée en utilisant l'estimation par intervalle de confiance. Il s'agit là d'une *analyse inférentielle univariée*.

Dans ce labo, nous verrons comment faire calculer et interpréter :

- Les mesures de tendance centrale, pour rendre compte de la représentativité d'un phénomèn social;
- Les mesures de variation, qui informent sur la variabilité d'un phénomène social;
- L'intervalle de confiance, pour estimer avec un certain degré de certitude l'intervalle à l'intérieur duquel se situe la vraie moyenne d'une population.

Des exercices d'application aideront à atteindre ces objectifs d'apprentissage.

1. Mesures de tendance centrale

Si les distributions de fréquences et de pourcentages sont adéquates pour décrire les données d'une variable de façon à apprécier l'ampleur d'un phénomène, les mesures de tendance centrale sont encore plus appropriées lorsqu'il s'agit d'obtenir un résumé quantitatif précis représentatif de l'ensemble des données. Il en est ainsi du mode, de la médiane et de la moyenne. Précisément, quelle est la valeur centrale autour de laquelle sont agglomérés les scores d'une distribution ? ***Quel est le score typique, le plus représentatif ou commun d'une distribution*** ? Pour répondre à ces questions, on calcule la tendance centrale des données.

1.1. Les étapes d'une analyse descriptive à l'aide de la tendance centrale

Procédons en quatre étapes à l'aide de SPSS :

1ère étape : examiner la variable à analyser

2e étape : faire exécuter les calculs

3e étape : sauvegarder ou reconfigurer les résultats

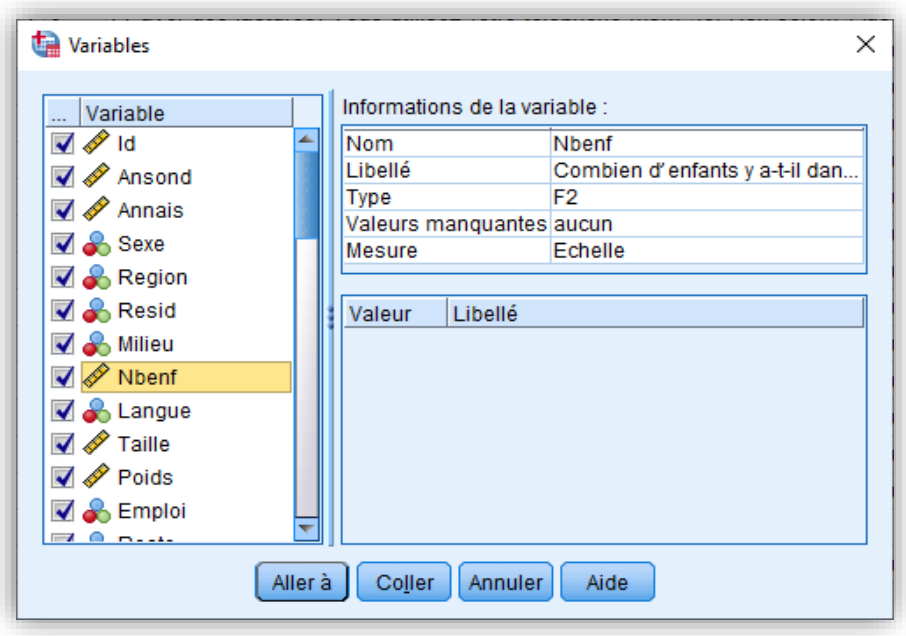
4e étape : analyser/interpréter les résultats

Pour illustrer ces étapes, intéressons-nous à la taille des familles des étudiants (**nbenf**).

1.1.1. Vérification de la variable à analyser

Utilitaires

Variables



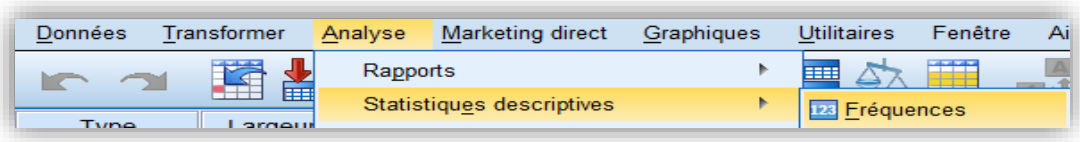
Nous pouvons observer l'étiquette ou le libellé de la variable **Nbenf**, ainsi que son niveau de mesure (échelle ou quantitative).

1.1.2. Exécution des opérations d'analyse statistique

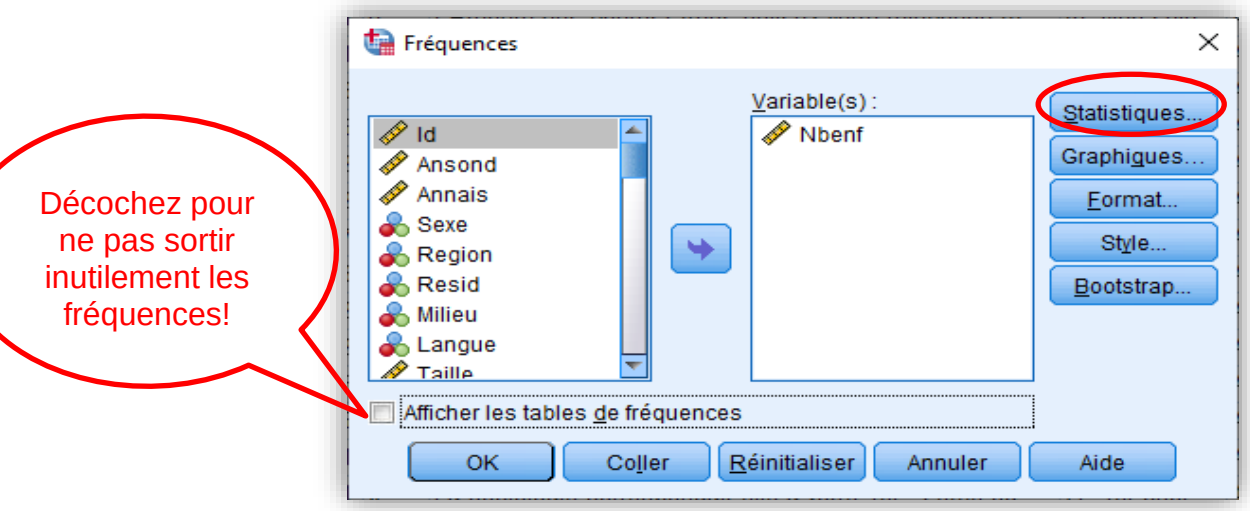
Une fois les informations sur le nombre d'enfant des familles des étudiants connues et

validées, nous pouvons faire exécuter des opérations de calcul pour connaître la tendance centrale, la variable étant quantitative. Voici la procédure :

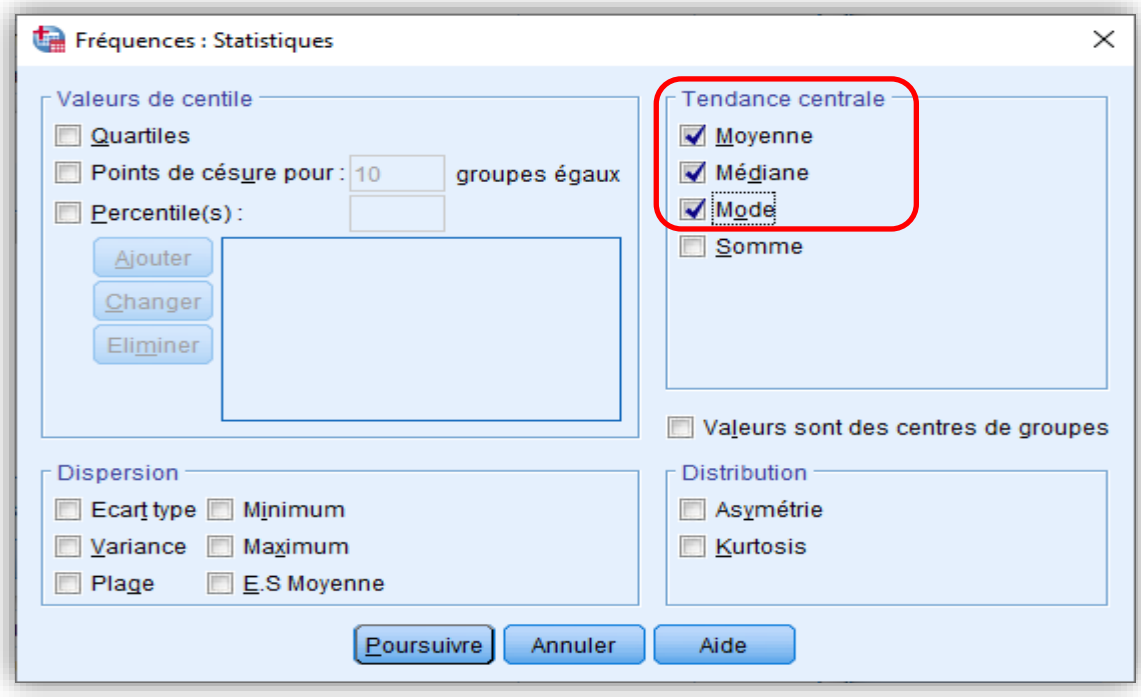
Analyse
Statistiques descriptives
Fréquences



Cliquez sur Fréquences pour obtenir la fenêtre ci-dessous :



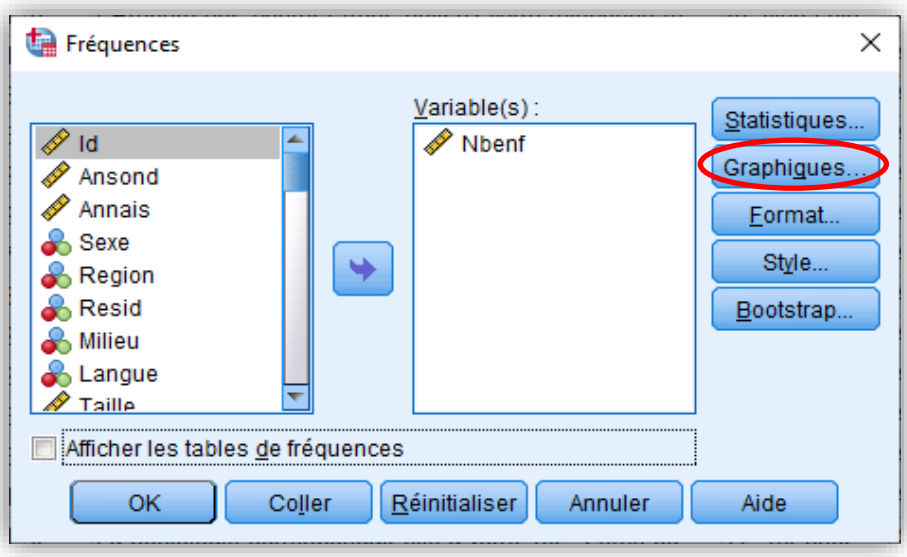
Cliquez sur Statistiques pour obtenir l'écran de dialogue ci-dessous :



Vous pouvez maintenant sélectionner les mesures de tendance centrale : mode,

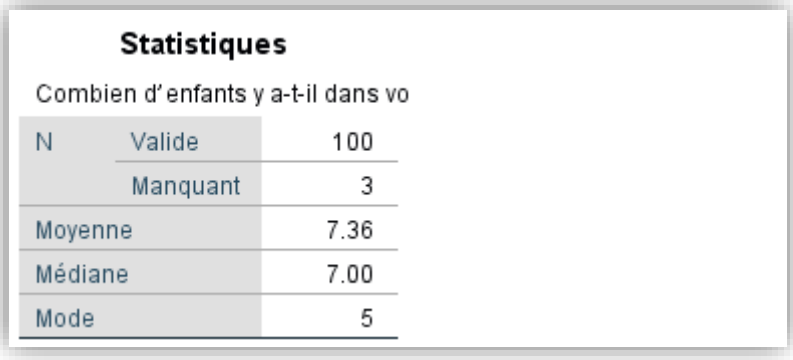
médiane et moyenne. Poursuivez !

On peut visualiser la distribution de la variable Nbenf à l'aide du diagramme en bâtons linéaires. Ce type de diagramme est approprié pour rendre compte du caractère discret d'une variable. Cliquons sur Graphiques !



On obtient la fenêtre contextuelle ci-dessous :

Poursuivez et validez le tout pour voir apparaître les résultats!



Statistiques

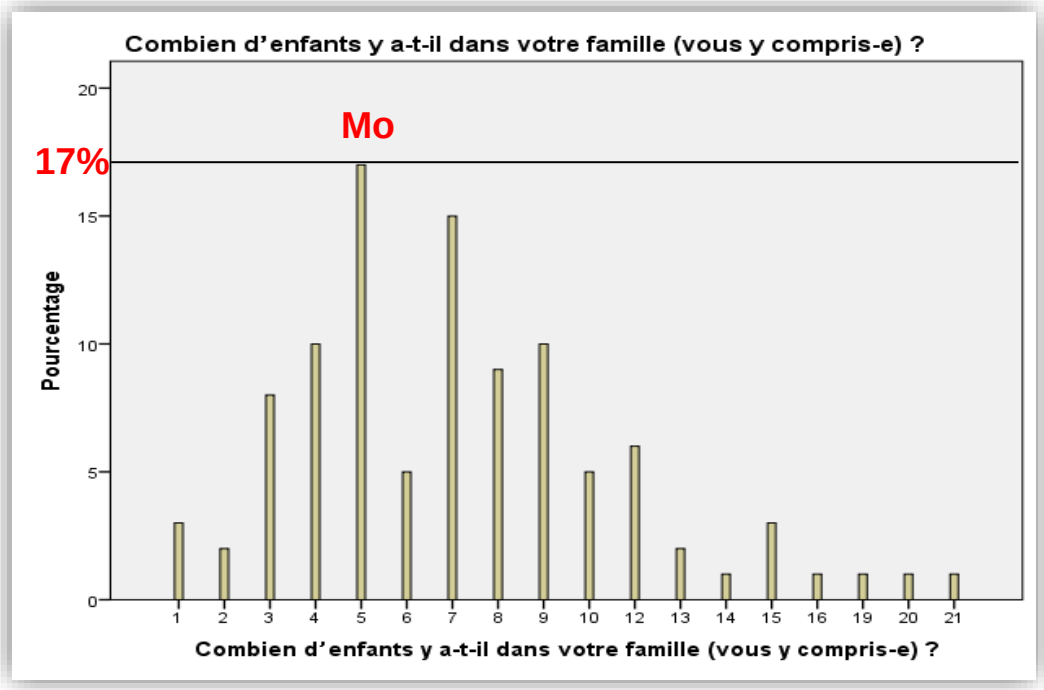
Combien d'enfants y a-t-il dans vo

N	Valide	100
	Manquant	3
Moyenne		7.36
Médiane		7.00
Mode		5

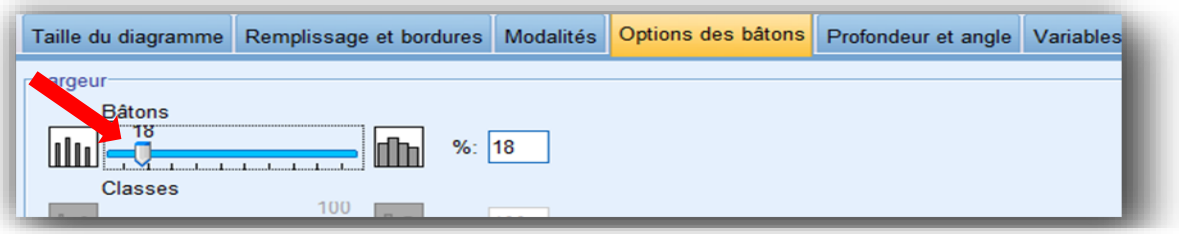
Moyenne:	Mesure de tendance centrale la plus utilisée, soit la somme des scores de tous les cas, divisée par le nombre de scores ou cas.
Médiane:	Valeur qui occupe la place du milieu dans le rangement ascendant ou descendant des valeurs d'une variable. Autrement dit, c'est le score de la variable qui divise une distribution de telle sorte que 50% des scores se trouvent au-dessus et 50% des scores se trouvent en dessous.
Mode:	Valeur la plus fréquente dans une série de scores.
Somme:	Somme de tous les scores d'une distribution.

1.1.3. Reconfiguration ou visualisation des résultats

Lorsqu'une sortie SPSS comporte seulement trois statistiques ou moins (ici mode, médiane, moyenne), il est inutile de reconfigurer et de présenter le tableau. Les statistiques sont directement présentées dans le texte. Cependant, on peut visualiser la distribution de la variable Nbenf à l'aide du diagramme en bâtons linéaires.



Par contre, les barres du diagramme ne sont pas vraiment des bâtons, et n'épousent pas une forme linéaire. Il faut donc diminuer les intervalles entre les barres. Pour réduire les barres du diagramme et leur conférer un caractère discret, double-cliquez sur le graphique, ensuite double-cliquez sur une barre pour obtenir la fenêtre ci-dessous. Sur « Options des bâtons », réduisez en tirant le bouton vers la gauche…



1.1.4. Interprétation statistique et théorique des résultats

<u>Interprétation statistique :</u> (Que disent les chiffres ? Que suggèrent-ils?)

Le mode est 5, suggérant que la plupart des 100 étudiants ayant répondu (17%) ont 5 enfants dans leur famille. Le diagramme en bâtons en offre une illustration visuelle.

La médiane est 7. *Autrement dit, **au moins 50%** des 100 étudiants ayant répondu ont 7 enfants **ou moins** dans leur famille.*

La moyenne est égale à 7,36. Ce qui signifie que les 100 étudiants ayant répondu ont *en moyenne* 7,36 (ou 7) enfants dans leur famille.

Somme toute, au regard de la taille de la famille, la moyenne semble plus élevée que la médiane ou le mode. Ce qui atteste la présence de familles dont le nombre d'enfants est anormalement élevé par rapport à la tendance centrale.

<u>Interprétation théorique/sociologique :</u> (Comment expliquer la conclusion ?) Globalement, les mesures de tendance centrale indiquent la présence d'un nombre élevé d'enfants dans les familles des étudiants. Cette importance de la fécondité peut s'expliquer par la forte prégnance de la polygamie dans la société sénégalaise.

2. Mesures de variation

Jusqu'à quel point les inégalités sociales, en termes de revenu, scolarité, de santé, sont-elles importantes dans une communauté ? Certes, les mesures de tendance centrale sont appropriées lorsqu'il s'agit d'obtenir un résumé quantitatif précis typique d'une distribution. Toutefois, elles ne renseignent pas sur la façon dont les scores d'une distribution sont éparpillés ou dispersés les uns des autres. Les **mesures de variation** sont adéquates à cet effet. Elles permettent de connaître, surtout, la dispersion ou l'éparpillement des scores par rapport à la tendance centrale. Il en est ainsi de l'étendue et de l'intervalle interquartile, de la variance et de l'écart-type, et du coefficient de variation. Ces mesures sont particulièrement pertinentes pour décrire une variable quantitative en vue de rendre compte de la **variabilité d'un phénomène**. Une large dispersion des scores signifie que la distribution est **hétérogène** tandis qu'une petite dispersion révèle une distribution **homogène**. Il est possible d'obtenir une idée plus nette et visuelle de la dispersion des données, en se basant sur la **forme des distributions**. Celle-ci comporte deux caractéristiques majeures : l'asymétrie (ou symétrie) et la kurtose (aplatissement).

À l'aide de SPSS, nous apprenons ici et maintenant à :

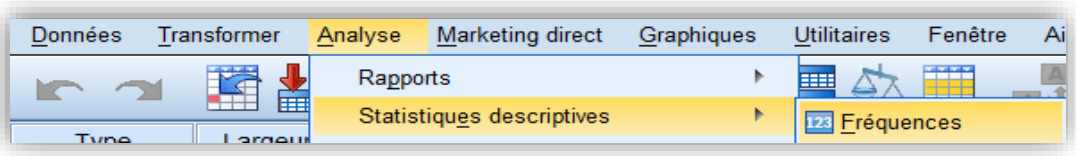
- ✓ Calculer les mesures de variation et de la forme d'une distribution ;
- ✓ Représenter la forme d'une distribution à l'aide de la boîte à moustaches, notamment en vue de détecter des cas déviants ;
- ✓ Représenter la forme d'une distribution d'une variable continue à l'aide de l'histogramme en regroupant automatiquement les valeurs en classes ;

2.1. Mesures de variation et de la forme d'une distribution

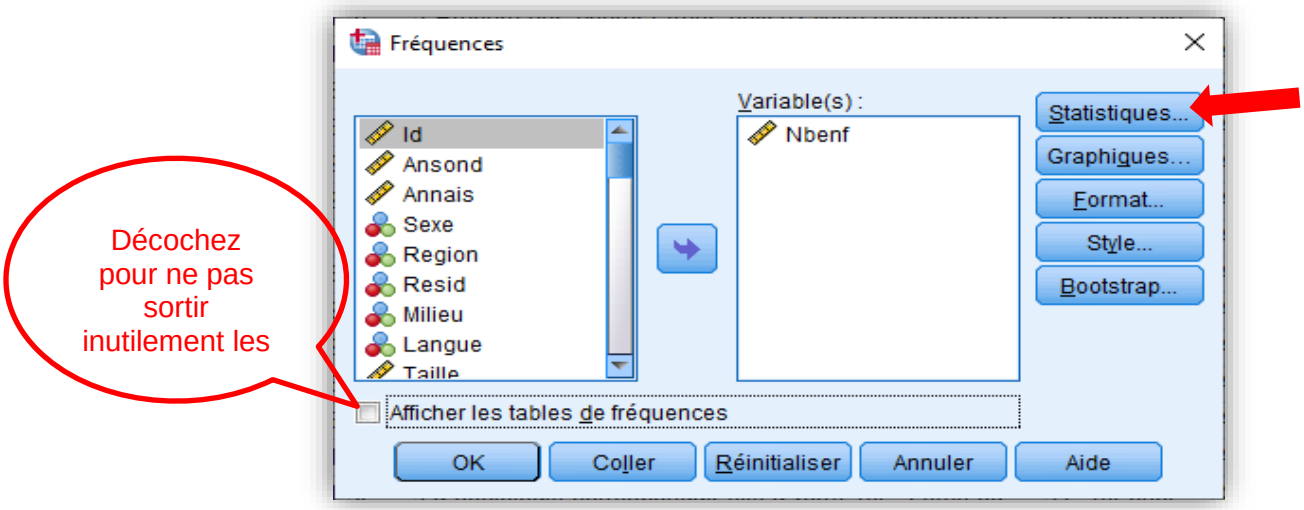
Nous nous intéressons à la variabilité du nombre d'enfants dans les familles des étudiants : « **nbenf** ». Pour faire calculer les mesures de variation et de forme à l'aide de SPSS, il y a au moins deux procédures.

2.1.1. La procédure FREQUENCIES

Analyse
Statistiques descriptives
Fréquences
Statistiques.



Cliquez sur **Fréquences** pour obtenir la fenêtre ci-dessous :



Cliquez sur la variable « Nbenf » pour la faire passer dans l'autre rectangle, puis sur Statistiques pour obtenir l'écran de dialogue ci-dessous.

Sélectionnez les mesures suivantes : Moyenne, Quartiles, Ecart type, Variance, Etendue, Minimum, Maximum, Skewness (asymétrie) et kurtosis (aplatissement).

Les mesures de variation	
Écart-type	mesure la dispersion des scores autour de la moyenne. Un écart-type qui est grand par rapport à la moyenne indique la présence de données dispersées autour de la moyenne, donc hétérogènes, alors qu'un écart-type petit témoigne de la présence de données concentrées autour de la moyenne, donc relativement homogènes. Plus des 2/3 des données (68%) sont situées à plus ou moins l'écart-type de la moyenne.
Variance :	Écart-type élevé au carré. S'interprète en termes d'unités carrées.
Plage:	Étendue, c'est-à-dire différence entre la plus grande valeur et la plus petite valeur d'une série de scores.
Minimum / Maximum:	Plus petite et plus grande valeurs dans une distribution.
Écart-moyen (E.S):	Distance moyenne (en valeur absolue) entre les scores et la moyenne. Il se calcule ainsi : $\sum\|X_i - \bar{X}\|/n$.
Les mesures de la forme d'une distribution	
Aplatissement (kurtosis)	indique si la distribution est haute et étroite (recentrée autour de sa moyenne) ou aplatie et large (écartée). Le coefficient est nul pour une distribution normale ou moyenne (mésokurtique), négatif pour une distribution aplatie (platykurtique), positif pour une distribution haute (leptokurtique). L'aplatissement se lit verticalement.
Asymétrie (skewness)	indique le degré de symétrie d'une distribution : un coefficient négatif indique la présence d'une distribution négativement asymétrique ou étendue vers la gauche, un coefficient positif indique la présence d'une distribution positivement asymétrique ou étendue vers la droite, zéro indiquant une distribution symétrique. Contrairement à l'aplatissement, l'asymétrie se lit horizontalement.
Les fractiles ou quantiles:	
Quartiles	Le premier quartile est la valeur de la variable qui divise la distribution de telle sorte que 25% des valeurs se trouvent en dessous d'elles (25^e centile). Le troisième quartile est la valeur de la variable qui divise la distribution de telle sorte que 75% des valeurs se trouvent en dessous d'elles (75^e centile). Le deuxième quartile renvoie à la médiane, et donc partage une distribution en deux parties égales. Les quartiles 1 et 3 permettent de calculer l'intervalle interquartile.

Poursuivez et validez pour afficher la page des résultats.

Statistiques

Combien d'enfants y a-t-il dans votre famille

N	Valide	100
	Manquant	3
Moyenne		7.36
Ecart type		3.999
Variance		15.990
Asymétrie		1.152
Erreur standard d'asymétrie		.241
Kurtosis		1.634
Erreur standard de Kurtosis		.478
Plage		20
Minimum		1
Maximum		21
Percentiles	25	5.00
	50	7.00
	75	9.00

<u>Interprétation statistique :</u> (Que disent les chiffres ? Que suggèrent-ils ?)

L'écart-type est de 3.999. Ce qui signifie que*, environ 2/3 des 100 étudiants répondants ont dans leur famille 7 enfants (7,36) plus ou moins 4 (3,999)*. Autrement dit, quelque 68% des étudiants appartiennent à des familles ou le nombre moyen d'enfants se situe entre 3 et 11 (7,36-4 et 7,36+4). Encore, si l'on prend au hasard un étudiant, il y a 68% de chances que sa famille compte 3 à 11 enfants.

Pour obtenir une idée plus nette de la dispersion, nous pouvons calculer le coefficient de variation relative en relativisant l'écart-type par la moyenne et en multipliant le dividende par 100. Il est égal à 54,3% ((3,999/7,36) * 100). Le coefficient étant de loin supérieur à 15%, nous pouvons dire que la taille de la famille est fortement disparate chez les répondants étudiants. ***La distribution est très hétérogène***.

La variance est égale à 15.99. Autrement dit, ***les scores relatifs au nombre d'enfants dans les familles des étudiants divergent de la moyenne de 15.99***[1]**.**

L'étendue est donnée par la plage, qui est de 20. **La distance qui sépare le nombre d'enfants maximal du nombre d'enfants minimal est de 20 (max 21 – min 1)**[2].

Le quartile 1 ou centile 25 (5.00) signifie qu'**au moins 25% des répondants étudiants**

[1] [Toutefois, la variance n'est pas appropriée pour décrire la dispersion des données, sa valeur n'apparaissant pas sur une même échelle que les scores de la distribution. Elle sert surtout à des raisonnements statistiques plus avancés, inférentiels notamment.]

[2] [Cependant, cette mesure est moins pertinente que l'écart-type pour mesurer la dispersion des données, puisqu'elle demeure trop sensible aux valeurs extrêmes et ne dépend que de deux valeurs.]

ont des familles comprenant 5 enfants ou moins. Le quartile 3 ou centile 75 (9.00), quant à lui, s'interprète ainsi : **au moins 75% des répondants étudiants ont des familles comprenant 9 enfants ou moins**.

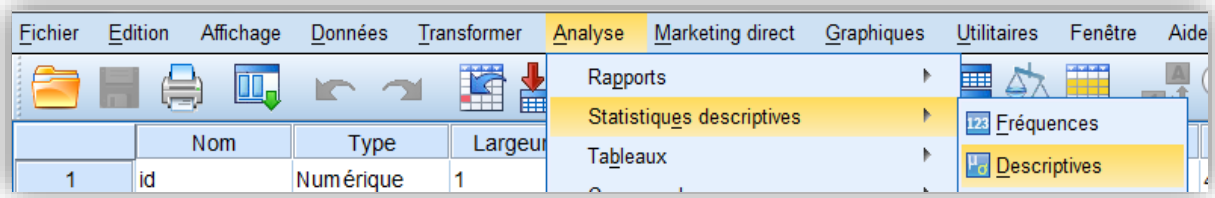
L'intervalle interquartile peut être calculé en soustrayant le quartile 1 ou centile 25 (5.00) du quartile 3 ou centile 75 (9.00). Précisément, 9.00-5.00 = 4.00. **Au moins 50% des étudiants ont dans leur famille entre 5 et 9 enfants**, l'intervalle interquartile étant de 4 enfants autour de la médiane ou centile 50 (7.00).

Concernant la forme de la distribution, **le coefficient d'asymétrie étant de 1,152, nous sommes en présence d'une asymétrie positive**. De l'autre côté, **le coefficient de kurtosis étant de 1.634, la distribution est plutôt leptokurtique**. Nous verrons ci-après comment visualiser ces différentes formes de distribution.

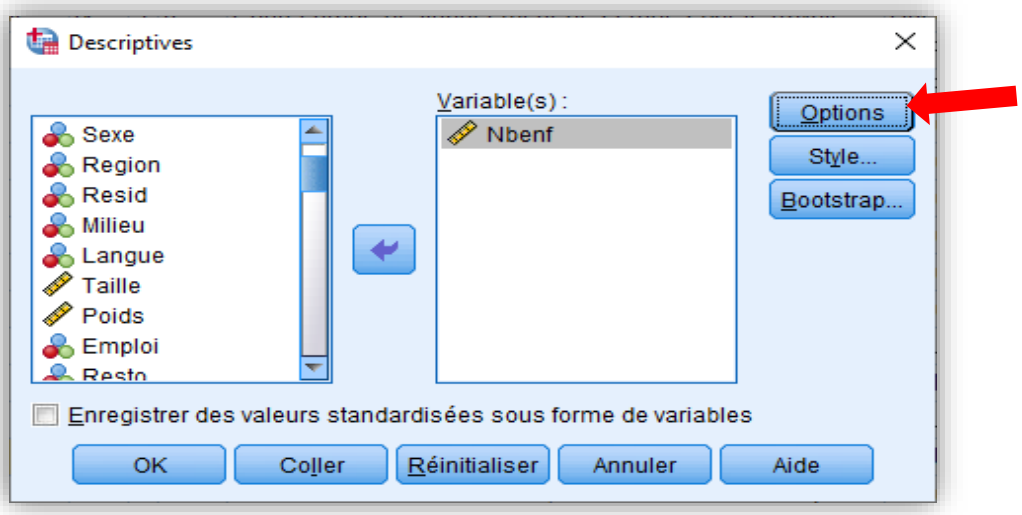
Retenez qu'« **Erreur standard d'asymétrie** » et « **Erreur standard de kurtosis** » constituent ce qu'on appelle des erreurs-types (une forme particulière d'écarts-types) que nous verrons dans la leçon consacrée à l'inférence statistique. Ces erreurs-types permettent de déterminer jusqu'à quel point une distribution s'éloigne suffisamment de la forme typique d'une courbe normale. Il suffit de diviser les coefficients d'asymétrie ou de kurtose par leur erreur standard respective. Les ratios > 2 ou -2 révèlent un problème majeur d'asymétrie ou de kurtose.

2.1.2. La procédure DESCRIPTIVES

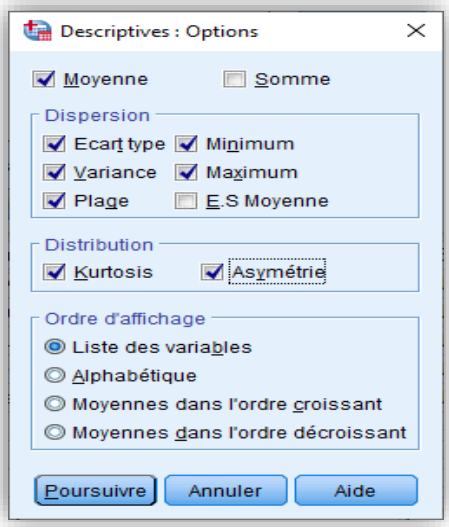
Une autre procédure pour obtenir les mesures de variation et de forme consiste à aller sur « Analyse », puis « Statistiques descriptives », puis « Descriptives ».



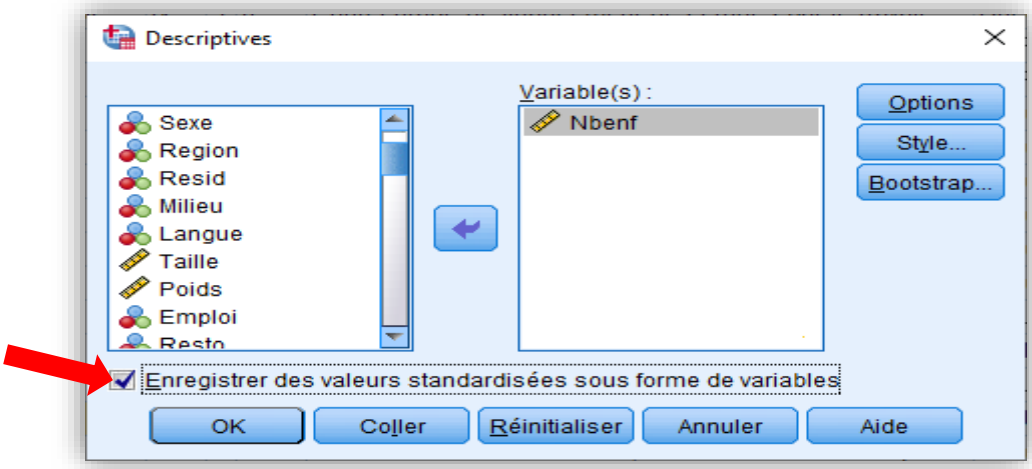
Les sorties sont les mêmes que celles de la procédure FREQUENCIES. Par contre, comme on le voit sur la fenêtre ci-dessous, les possibilités sont un peu plus limitées. Par exemple, on n'y voit plus les percentiles. Dans la fenêtre qui s'ouvre, cliquez sur ***Options***.



Et cochez les statistiques d'intérêt, tel qu'illustré ci-dessous.



En revanche, ce qui est intéressant avec la procédure DESCRIPTIVES, c'est qu'elle permet de sortir les scores-Z ou standardisés. Pour ce faire, cochez la case : « Enregistrer des valeurs standardisées sous forme de variables »

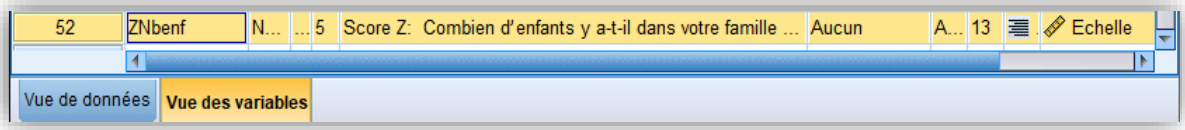


Validez pour voir s'afficher la page des résultats !

Statistiques descriptives

	N	Plage	Minimum	Maximum	Moyenne	Ecart type	Variance	Skewness		Kurtosis	
	Statistiques	Statistiques	Statistiques	Statistiques	Statistiques	Statistiques	Statistiques	Statistiques	Erreur std.	Statistiques	Erreur std.
Combien d'enfants y a-t-il dans votre famille (vous y compris-e) ?	100	20	1	21	7.36	3.999	15.990	1.152	.241	1.634	.478
N valide (liste)	100										

Vérifiez que la variable standardisée **ZNbenf** est créée au bas de la liste des variables!



Nous avions appris que la moyenne des scores-Z d'une distribution est toujours égale à 0 et leur écart-type 1. Confirmons ces équivalences en sortant les statistiques descriptives de la variable standardisée. Allez sur « Analyse », puis « Fréquences », et mettez la variable **ZNbenf** dans la cage d'instruction. Cliquez sur Statistiques et sortez seulement la moyenne, l'écart-type, le score minimum, le score maximum. N'oubliez pas de décocher les tables de fréquences. Validez le tout !

Statistiques

Score Z: Combien d' enfants y a-t-il d

N	Valide	100
	Manquant	3
Moyenne		.0000000
Ecart type		1.00000000
Minimum		-1.59048
Maximum		3.41103

Les scores-Z de la variable **ZNbenf** s'étendent de -1.59 à 3.41. Leur moyenne est égale à 0 et leur écart-type 1. Ainsi, on peut ramener les scores-X de plusieurs variables mesurées différemment en scores-Z de telle sorte que ces variables puissent être comparables, et les approximer à l'aide de la loi normale.

2.2. Représentation graphique de la forme d'une distribution

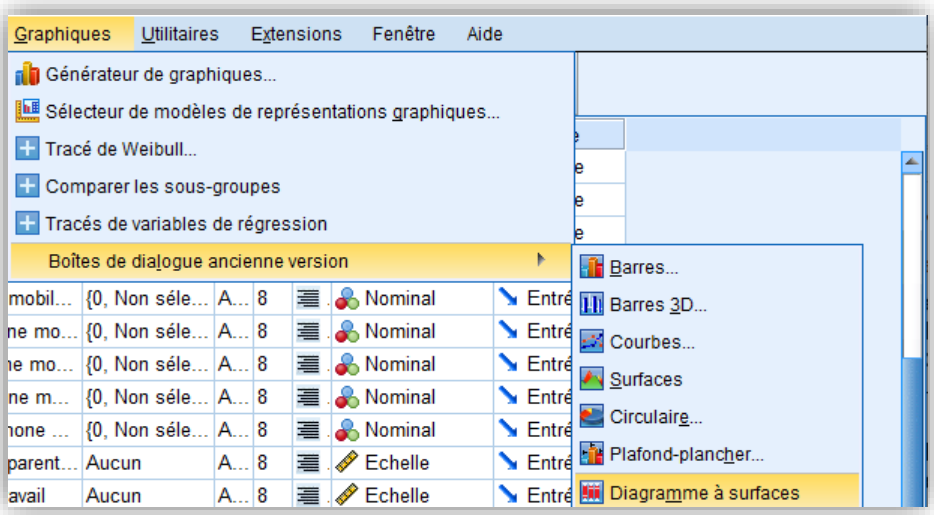
2.2.1. Le diagramme en boîte à moustaches

Le diagramme en **boite à moustaches** permet de visualiser le degré de dispersion des données d'une distribution quantitative, notamment en mettant en évidence la présence ou non de cas déviants, l'étendue et l'intervalle interquartile. Pour ce faire, intéressons-nous toujours à la variable « Nbenf ». La procédure est la suivante :

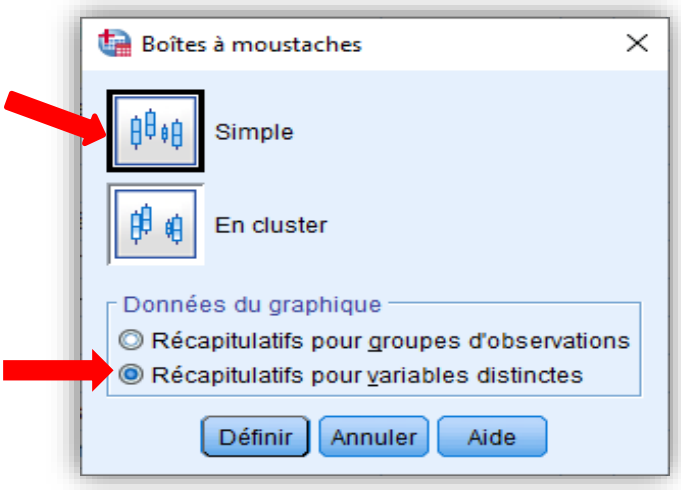
Graphiques

Boîtes de dialogue ancienne version

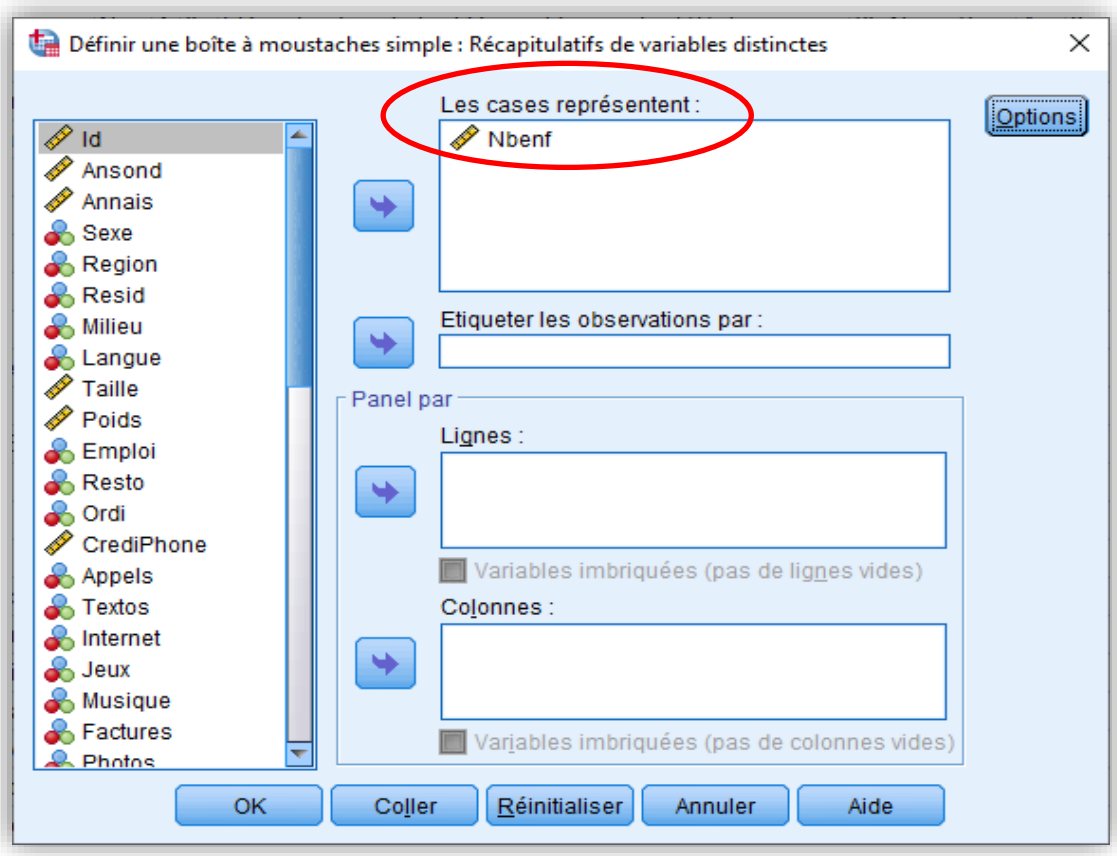
Diagramme à surface (Boîte à moustaches : Autre version SPSS).



Un écran de dialogue apparaît :

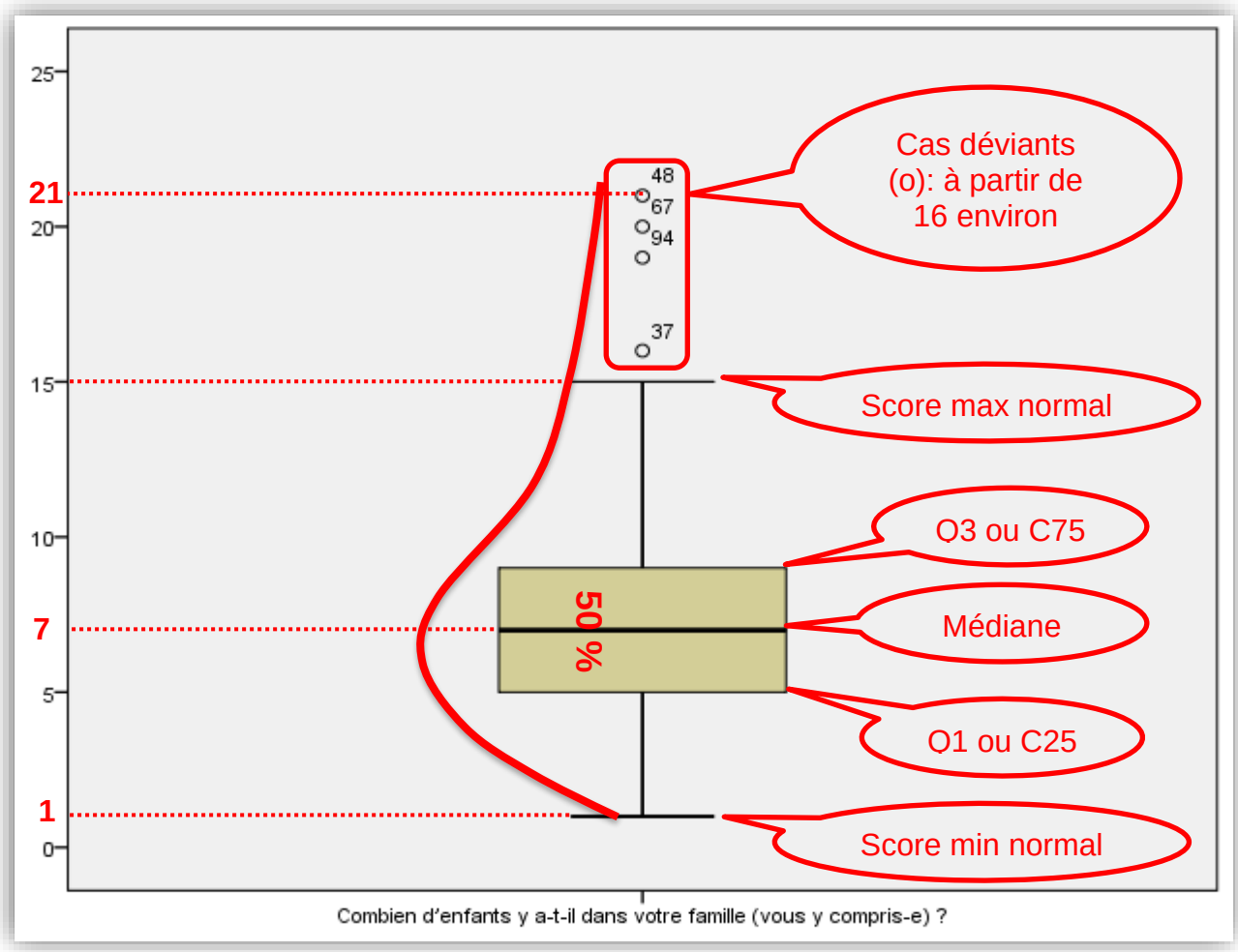


Cliquez sur « Simple », puis sur « Récapitulatifs pour variables distinctes », puisque nous avons affaire à une variable quantitative, càd distincte.



Faites passer la variable en question dans le rectangle « Les zones représentent ». On aurait pu aussi étiqueter les observations par Id. Mais, cela n'est pas nécessaire dans cette base de données puisque l'identification Id (analyste) est la même que l'identification automatique de SPSS.

Si vous validez le tout, la page des résultats s'affiche ! Contrairement aux autres graphiques, le diagramme en boîtes à moustaches est plus compact et se lit verticalement.



Analyse/interprétation statistique :

La lecture du diagramme montre clairement que la médiane (ligne horizontale foncée à l'intérieur de la boîte) se situe se situe autour de 7 enfants.

Situé entre la bordure inférieure de la boîte (25e centile ou 1er quartile) et la bordure supérieure (75e centile ou 3e quartile), l'intervalle interquartile est visible sur le diagramme puisqu'il encadre la boîte dans le sens vertical. On peut lire que 50% des étudiants ont dans leur famille entre 5 et 9 enfants approximativement.

Par ailleurs, le diagramme en boîtes permet de visualiser la présence de quatre cas déviants (nº37, nº 94, nº 67, nº 48)[3]. Ce sont des étudiants qui appartiennent à des familles ayant un très grand nombre d'enfants : 16 enfants ou plus. Ce qui témoigne d'une distribution asymétrique positive. Attestée aussi par la longue moustache supérieure (ligne verticale s'étendant de la bordure supérieure de la boîte et à la limite du diagramme), la courbe étant tirée par les scores élevés. On aurait pu éliminer ces

[3] La boîte à moustache permet de détecter deux types de cas déviants :
1. Les cas déviants synbolisées par le code **o** : ce sont les cas dont les scores sont moyennement éloignés car se situant au-delà des bordures inférieure ou supérieure du diagramme.
2. Les cas extrêmemenet déviants symbolisées par le code ***** : ce sont les cas dont les scores sont extrêmes car se situant aux limites du diagramme.

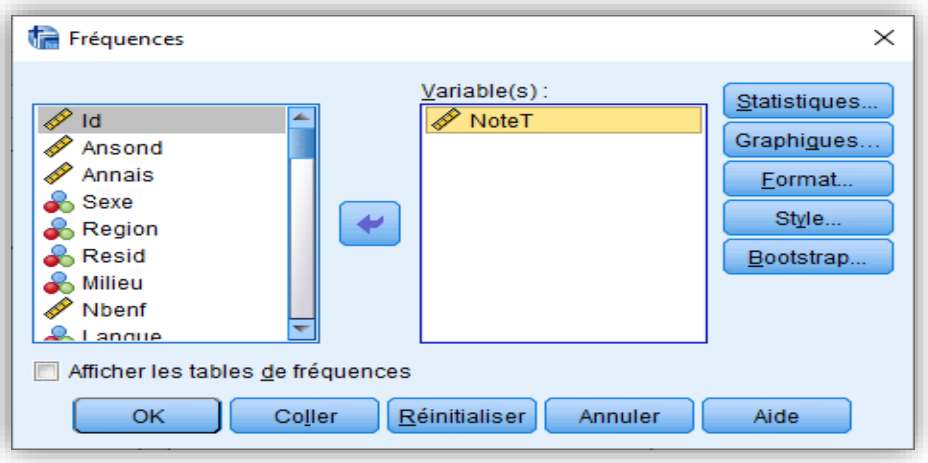
quatre cas déviants pour obtenir des statistiques davantage valides.

À juste titre, SPSS se base sur l'intervalle interquartile pour détecter les cas déviants. La logique repose sur la règle de Tukey. Si le score X s'écarte d'au moins 1,5*intervalle interquartile au-dessus de Q3 ou en dessous de Q1, il est considéré comme relevant d'un cas déviant (codé o). Si le score X s'écarte d'au moins 3*intervalle interquartile au-dessus de Q3 ou en dessous de Q1, il est considéré comme relevant d'un cas extrêmement déviant (codé *). Dans le diagramme en boîte à moustaches, il n'y a pas de cas extrêmement déviants.

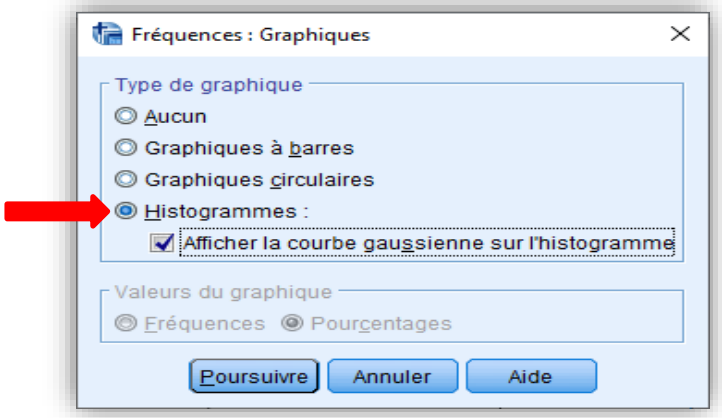
2.2.2. L'histogramme avec la courbe normale

Nous nous intéressons à la variation de la moyenne générale obtenue en Terminale par les étudiants (**NoteT**). La variable étant quantitative continue, l'histogramme est le graphique indiqué pour en représenter la distribution. Au fait, SPSS peut regrouper automatiquement les scores d'une variable continue pour en créer des classes. Pour ce faire, il construit l'histogramme en utilisant par défaut des amplitudes de classes qui s'adaptent à la distribution. Reprenons la procédure pour obtenir les fréquences.

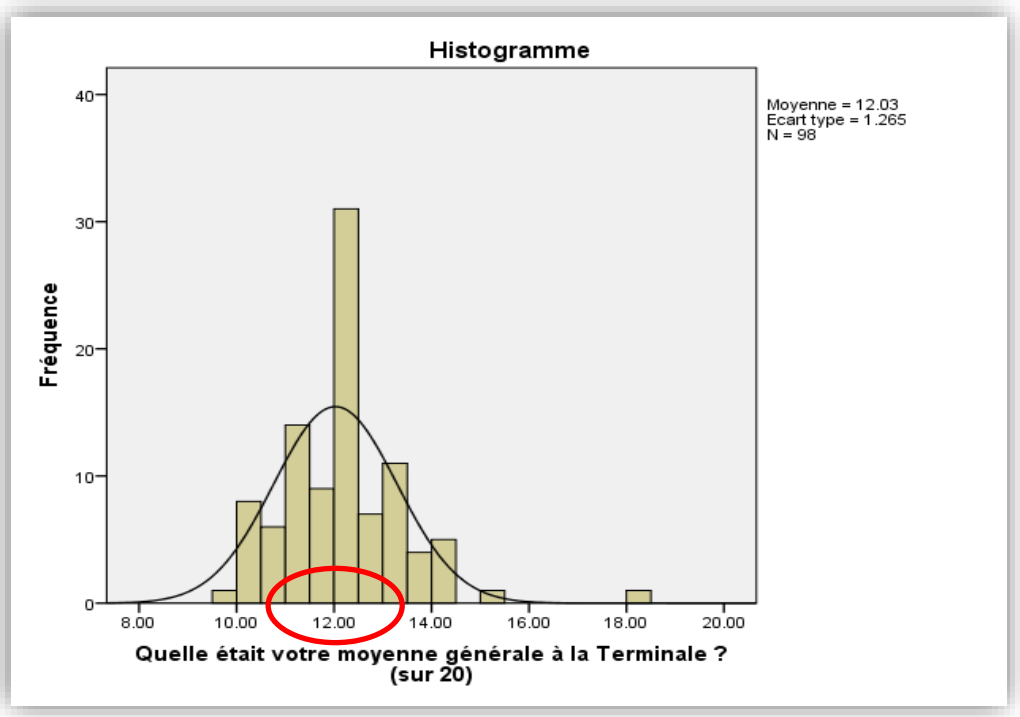
Analyse
Statistiques descriptives
Fréquences
Graphiques



Cliquez sur la variable « NoteT » pour la faire passer dans l'autre rectangle, puis sur « Graphiques » pour obtenir l'écran de dialogue ci-dessous.

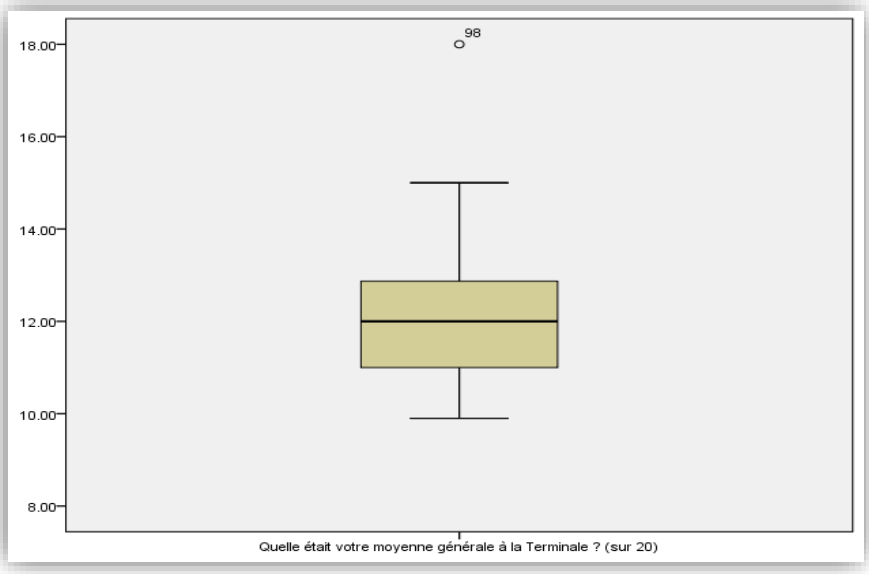


Sélectionnez « Histogrammes », puis « Afficher la courbe gaussienne sur l'histogramme ». La courbe gaussienne est la « courbe normale », en l'honneur de Gauss, l'inventeur de la loi normale. Poursuivez! Et validez le tout pour obtenir le formidable graphique ci-dessous : l'intervalle de classes est d'environ 0,5.



Analyse/interprétation statistique :

D'après l'histogramme, la moyenne générale « typique » en Terminale se situe probablement entre 11 et 13 chez les répondants étudiants, la majorité des cas se situant dans cette plage. La lecture de la courbe gaussienne suggère une distribution légèrement asymétrique vers la droite et assez haute (leptokurtique). La présence d'un cas déviant, dont la note est de 18, tire la courbe vers la droite. On aurait pu éliminer ce cas numéro 96, tel qu'illustré sur le diagramme en boîte à moustaches.



2.3. Exercice pratique

Vous vous intéressez au temps passé sur Internet en moyenne par jour (**HeureNet**). Le problème de recherche est double : *Quel est le nombre d'heures typique passé passées sur Internet chez les étudiants de L2 de la Section de sociologie de l'UGB ? Comment varie ce temps ?* Pour l'élucider, répondez aux questions ci-dessous :

a) Quelle est la nature de la variable ? (Qualitative nominale, qualitative ordinale, quantitative discrète ou quantitative continue). Justifiez ?

b) Calculez la moyenne et la médiane de cette distribution et interprétez-les.

c) Y a-t-il une différence entre la moyenne et la médiane ? Si oui, que signifie la différence ? Si non, que signifie l'égalité ?

d) Faire calculer les mesures de variation et de forme de la distribution.

e) Interprétez statistiquement chacune de ces mesures.

f) Demander à SPSS de créer automatiquement des classes en regroupant les valeurs, puis de représenter les classes à l'aide d'un histogramme avec la courbe gaussienne. Qu'observe-t-on ? Analysez !

g) Représentez la variable à l'aide d'un diagramme en boîte à moustaches. Qu'observe-t-on ? Analyser !

3. Analyses descriptives : tableau synthétique

Les études quantitatives portent très souvent sur un ensemble de variables dont on cherche à mettre en évidence les relations fonctionnelles ou structurelles. Une toute première analyse, cependant, consiste à résumer minutieusement les données de chaque variable de façon à produire une radiographie sociologique des profils des répondants. Dans un article scientifique, un mémoire ou une thèse, une bonne pratique consiste à présenter les résultats de l'analyse univariée dans un tableau synthétique avant d'en faire l'interprétation. On peut construire un tableau synthétique pour les variables qualitatives et un autre pour les variables quantitatives. Tout comme, il est possible d'opter pour un seul tableau compact afin de résumer quantitativement les données des variables analysées. Voici un exemple d'une façon de faire :

Tableau 1. *Caractéristiques des étudiants de L2 inscrits en sociologie*

	Statistiques	n
Sexe		103
Femmes	56 (54)	
Hommes	47 (46)	
Région de provenance		102
Saint-Louis	27 (26)	
Dakar	18 (18)	
Autre	57 (56)	
Nombre d'enfants	7,36 ± 4.00	100
Nombre d'heures sur Internet/jour	7.14 ± 5.72	88

Notes. Les entrées correspondent à des fréquences (avec les % entre parenthèses) pour les variables qualitatives. Pour les variables quantitatives, les entrées sont des moyennes en plus ou moins des écarts-types.
Source. Sondage_EtudiantsSocioL2_2021.

Interprétation statistique (analyse succincte) :

Les étudiants sondés sont constitués pour la plupart de femmes (54%) et de résidents des régions de Saint-Louis (26%) et de Dakar (18%). Quelque 68% des étudiants passent en moyenne 7.14 heures ± 5.72 sur Internet, ou proviennent d'une famille constituée en moyenne de 7,36 enfants ± 4.00. Par conséquent, le profil des étudiants de sociologie de L2 est particulier : femmes pour la plupart, résidents de la région saint-louisienne, grands internautes, ils sont issus d'une famille nombreuse.

4. Estimation par intervalle de confiance

L'inférence statistique comporte deux volets : l'**estimation** et le **test d'hypothèse**. L'estimation, en particulier, consiste à estimer la valeur d'un paramètre de la population à partir d'une statistique d'un échantillon. Elle peut être obtenue ponctuellement : dans ce cas, on avance simplement que la statistique observée, par exemple la moyenne, correspond approximativement à la vraie valeur, à condition toutefois que l'échantillon soit **représentatif** de la population (n ≥ 30 et aléatoire). Au demeurant, pour plus de précision, on peut se demander ***quel est l'intervalle à l'intérieur duquel est susceptible de se trouver la moyenne d'une population, alors que l'on ne connaît que la moyenne de l'échantillon.*** SPSS peut aider à estimer, par intervalle, la moyenne d'une population à un niveau de confiance donné, à partir d'une moyenne calculée sur des données d'échantillon.

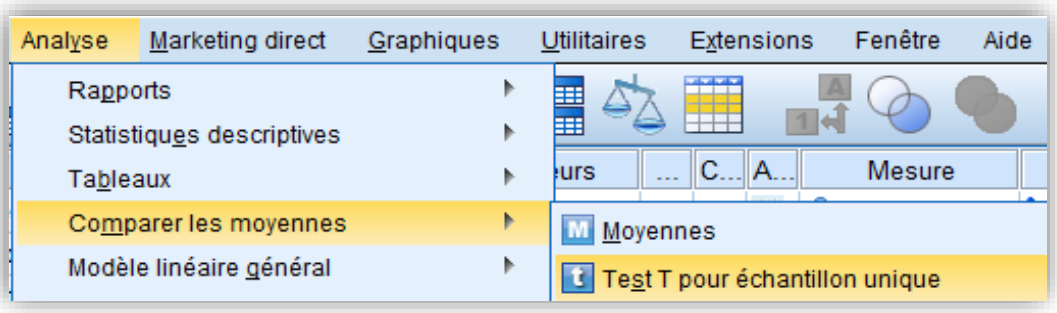
Selon le **théorème central limite**, la moyenne de la distribution d'échantillonnage des moyennes est égale à la moyenne de la population, soit $\mu_{\bar{x}} = \mu$. De même, l'écart-type de la distribution d'échantillonnage (appelé encore erreur-type) est égal à l'écart-type de la population divisé par la racine carrée de ***n***, soit $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Puisque la distribution d'échantillonnage suit la loi normale, si ***n*** aléatoire ≥ 30 à tout le moins, 95% des moyennes des échantillons possibles sont comprises à plus ou moins **1,96 erreur-type** de la moyenne de la population. À partir de ces constats, on peut calculer l'intervalle de confiance à 95% en soustrayant 1,96 erreur-type de la moyenne et en ajoutant 1,96 erreur type à la moyenne. Donc, il y a une probabilité de 95% que l'intervalle ainsi calculé contienne la moyenne de la population. Le problème est facilement réglé par SPSS! Dans ce labo, nous apprendrons à *calculer l'intervalle de confiance autour d'une moyenne à 95% ou 99%* et le représenter graphiquement à l'aide de la barre d'erreur.

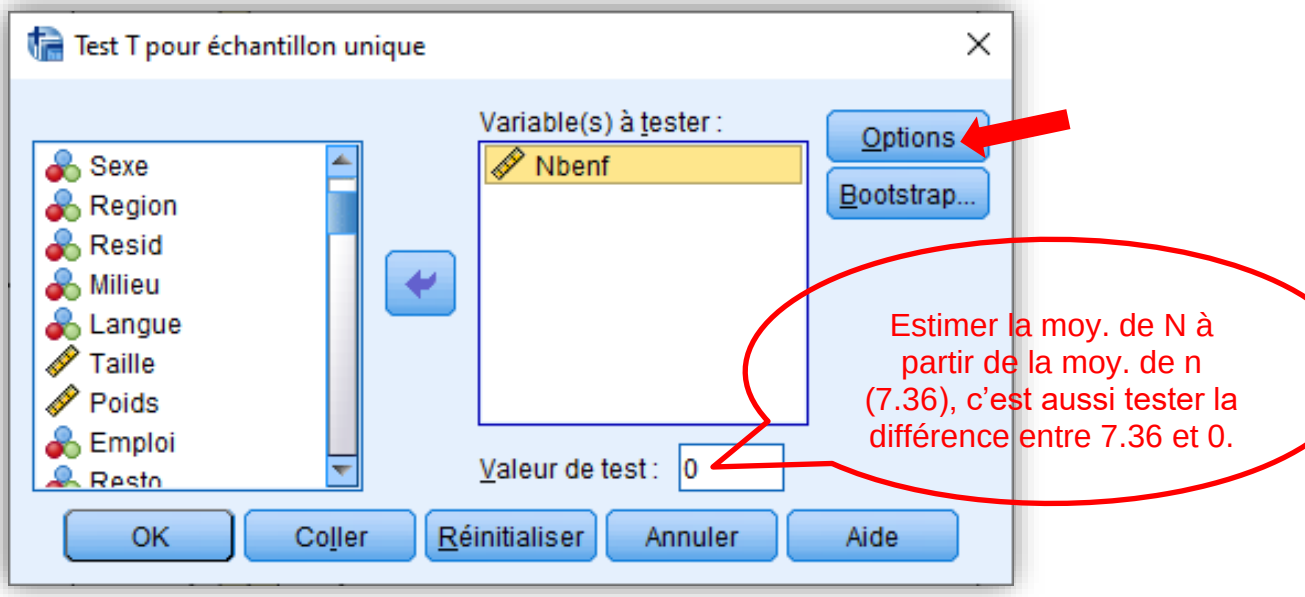
4.1. Calculer l'intervalle de confiance autour d'une moyenne

Prenons l'exemple du « nombre d'enfant dans les familles des étudiants » (**Nbenf**). Si l'on considère l'enquête comme un sondage aléatoire de taille n=103, il est possible de connaître à 95% l'intervalle à l'intérieur duquel se trouve la moyenne de la population étudiée. Pour ce faire, allons sur :

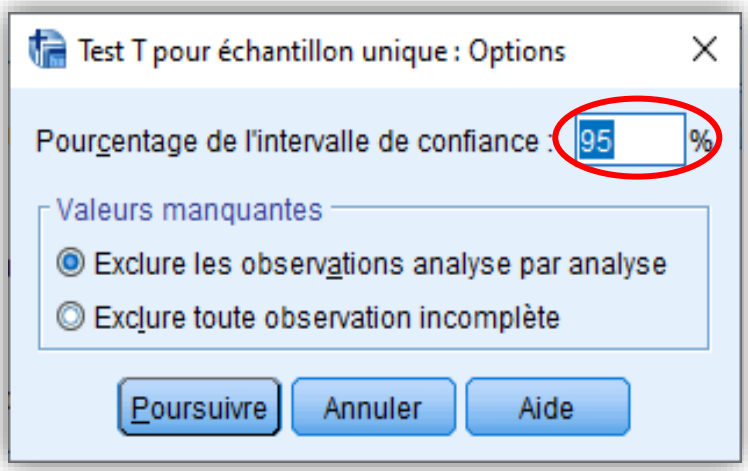
Analyse
Comparer les moyennes
Test T pour échantillon unique



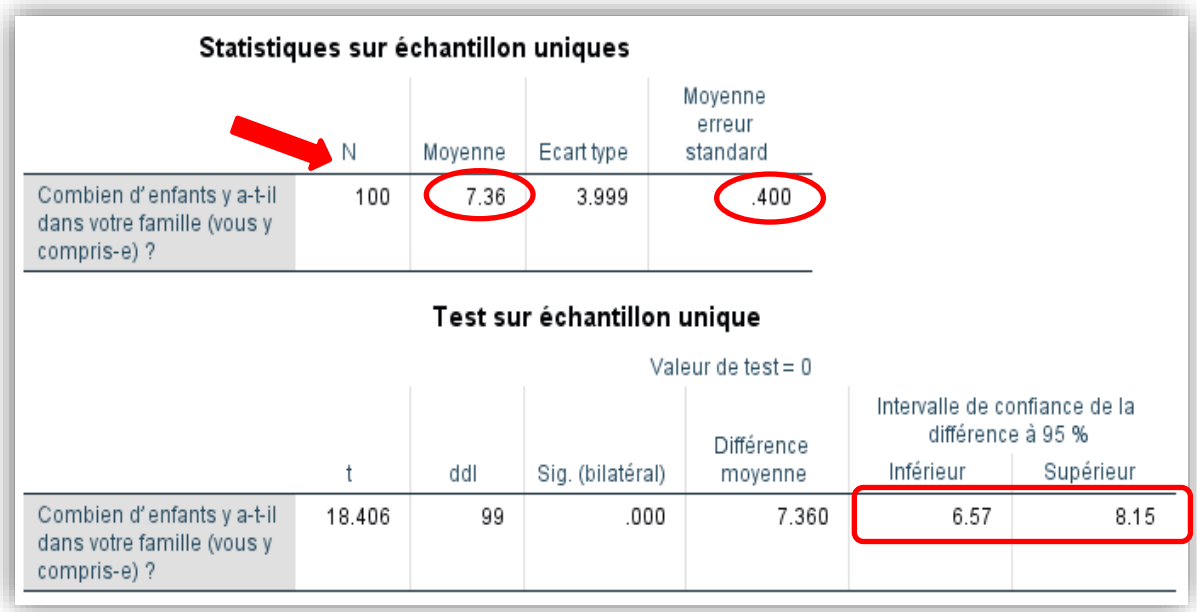
Une fenêtre apparaît! Cliquons sur la variable « Nbenf » puis sur la flèche pour la faire passer dans l'autre rectangle.



Ensuite, cliquons sur « Options » pour fixer le niveau de confiance à 95%. **C'est d'ailleurs le niveau de confiance par défaut de SPSS**. Il est possible que la commande « **Options** » ne soit pas disponible. Dans ce cas, continuez !



Poursuivons notre chemin ! Validons le tout pour voir s'afficher les résultats !



Statistiques sur échantillon uniques

	N	Moyenne	Ecart type	Moyenne erreur standard
Combien d'enfants y a-t-il dans votre famille (vous y compris-e) ?	100	7.36	3.999	.400

Test sur échantillon unique

Valeur de test = 0

	t	ddl	Sig. (bilatéral)	Différence moyenne	Intervalle de confiance de la différence à 95 % Inférieur	Supérieur
Combien d'enfants y a-t-il dans votre famille (vous y compris-e) ?	18.406	99	.000	7.360	6.57	8.15

Le premier tableau révèle que la taille moyenne de la famille est de 7,36 enfants dans l'échantillon, pour un écart-type de 3.999 et un nombre de cas valides ***n*** de 100. L'*erreur-type* (Erreur standard moyenne) *de la moyenne* est de 0,40. Autrement dit, si l'on répétait l'expérience de sélection de l'échantillon 100 fois, environ 68% des échantillons auraient une taille moyenne de la famille de 7.36 enfants plus ou moins 0,40. On obtient le même résultat si l'on applique la formule de l'erreur-type :

$$\sigma\overline{x} = \frac{\sigma}{\sqrt{n}} = \frac{3.999}{\sqrt{100}} = 0{,}40$$

La moyenne (7,36) est celle obtenue sur la base des 100 cas constituant l'échantillon valide. ***Est-elle proche ou éloignée du paramètre, c'est-à-dire de la vraie moyenne au niveau de la population étudiée ?*** Une estimation par intervalle de confiance peut aider à répondre à cette question. Pour y parvenir, examinons le deuxième tableau. Nous savons d'ores et déjà que dans une distribution normale, 95% des échantillons possibles se trouvent à l'intérieur de 1,96 écart-type de la moyenne.

Le deuxième tableau montre que l'intervalle susceptible de contenir la moyenne de la population étudiée avec une probabilité égale à 95% s'étend de 6.57 à 8.15, soit 7.36 - (1,96*0,40) et 7.36 + (1,96*0,40), 1,96 représentant le *score-z* à 95%, 0,40 l'*erreur-type de la moyenne*, leur produit la *marge d'erreur* (0,784).

$$IC\grave{a}95\% = \overline{X} \pm 1.96\sigma\overline{x}$$

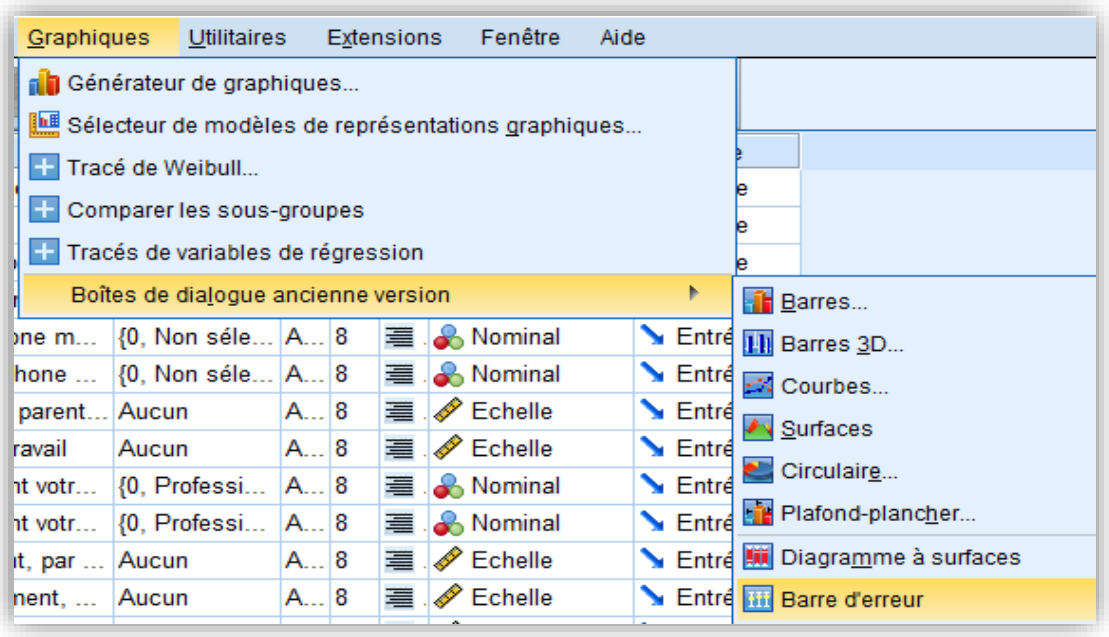
$$= 7.36 \pm 1.96(0{,}40) = 6.57\ \grave{a}\ 8.15$$

Interprétation statistique (ce que disent et suggèrent les chiffres): L'intervalle de confiance à 95% s'étend de **7,36 à 8,15**. Ce qui suggère qu'on est sûr au moins à 95% que la **taille moyenne de la famille** se situe entre 7,36 et 8.15 enfants dans l'ensemble de la population étudiante étudiée. Ou bien encore, la taille moyenne de la famille au sein de la population étudiée est de 7,36 enfants, avec une marge d'erreur de ±0,78, 95 fois sur 100 (19 fois sur 20, si l'on répète l'expérience).

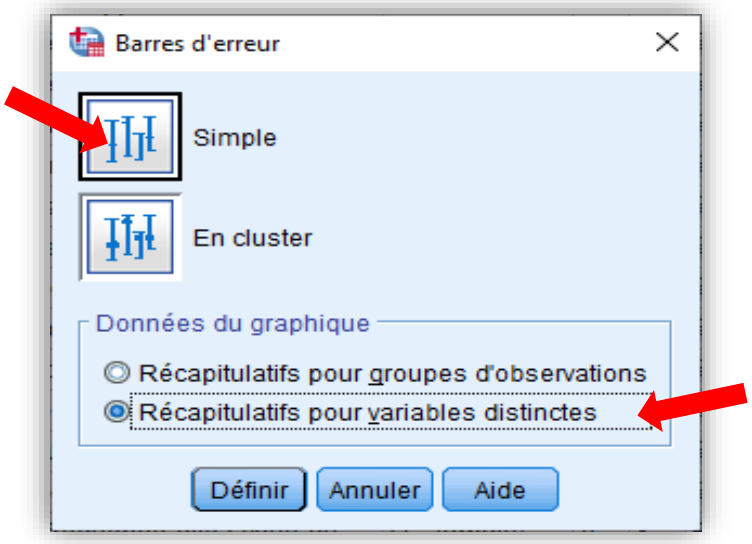
La statistique **t** est égale à: $\frac{\bar{x}}{\sigma x} = \frac{7{,}36}{0{,}40} = 18{,}41$. Nous la verrons plus tard.

4.2. Représenter graphiquement l'intervalle de confiance d'une moyenne

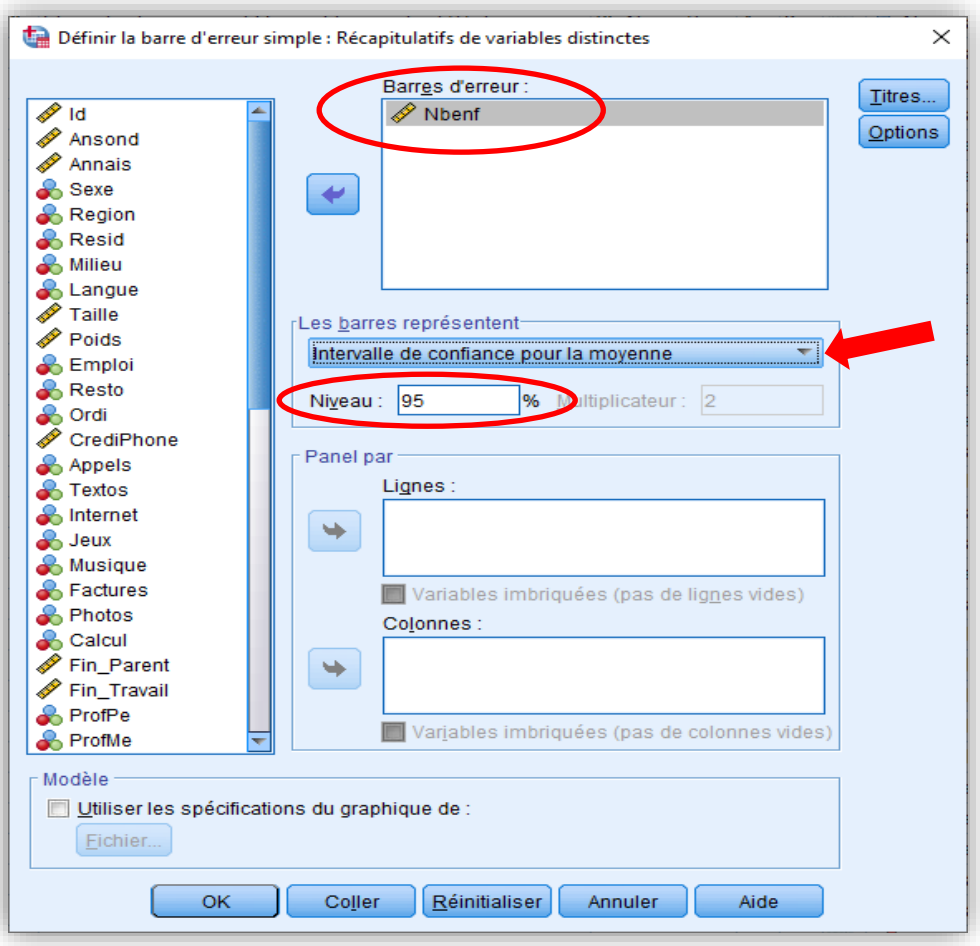
Pour représenter l'IC autour d'une moyenne sous forme graphique à l'aide de SPSS, on utilise la **barre d'erreur** (**<Graphiques <Boîtes de dialogue**), tel qu'illustré :



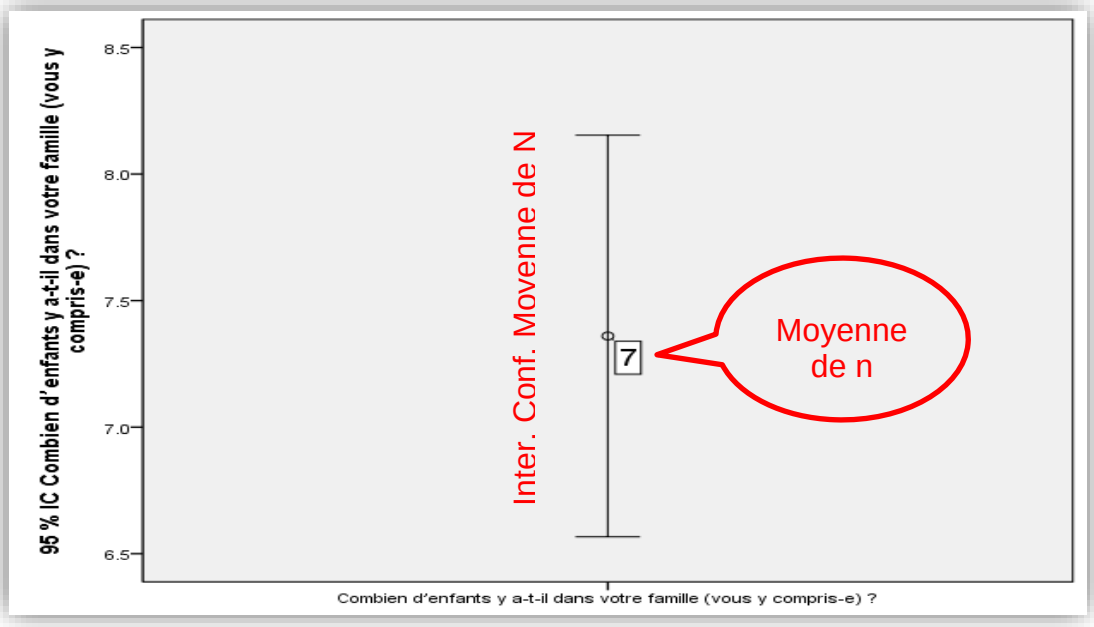
Cliquez sur Barre d'erreur ou Barre de variation, selon la version de SPSSS. On obtient la boîte de dialogue ci-dessous!



Cliquez sur « Simple », puis sélectionnez « Récapitulatifs pour variables distinctes » (la variable étant quantitative) et définissez pour obtenir la fenêtre ci-dessous!

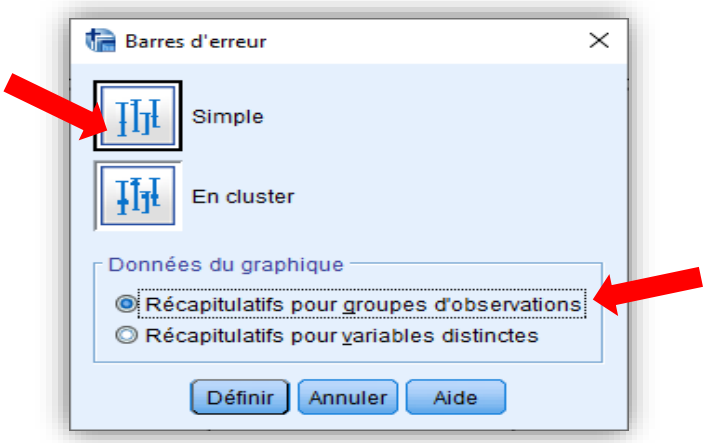


Sélectionnez la variable « **Nbenf** » pour la faire passer dans le rectangle « Barres d'erreur » (ou de variation), choisissez « intervalle de confiance pour la moyenne », précisez le niveau de confiance à 95% et validez pour obtenir la barre d'erreur. Cliquez deux fois sur le graphique pour « afficher les étiquettes de données ».

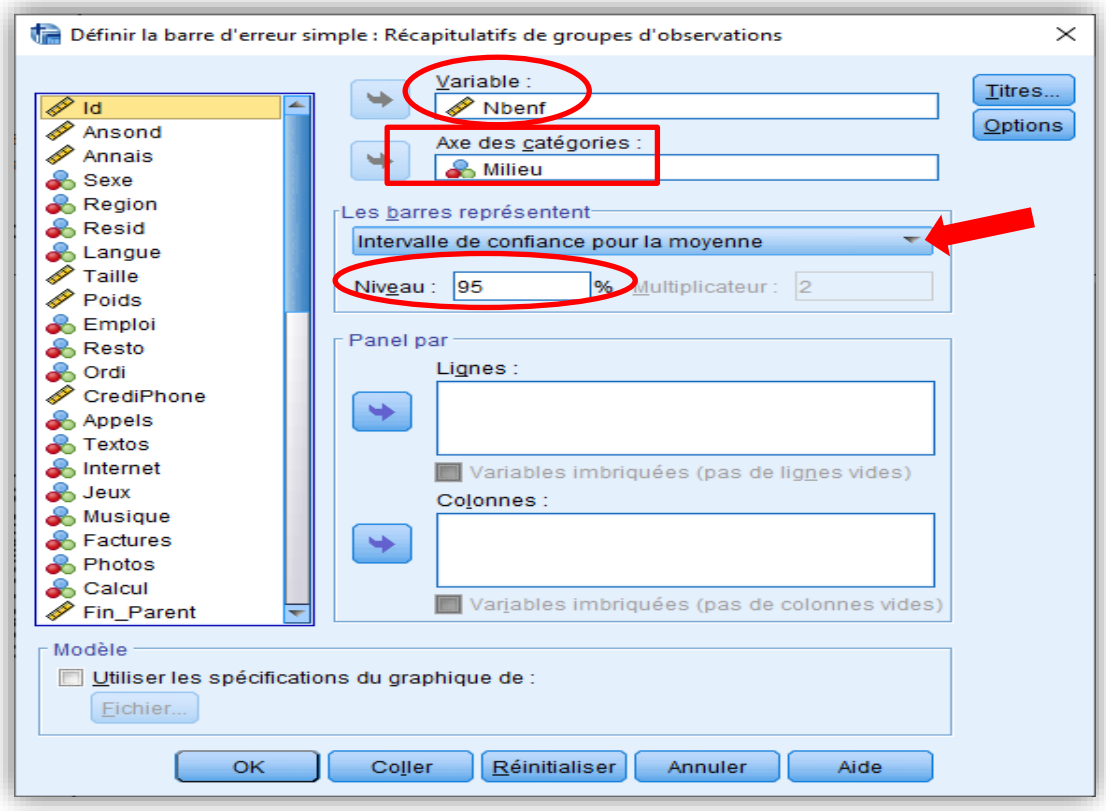


4.3. Comparer deux groupes à l'aide de la barre d'erreur

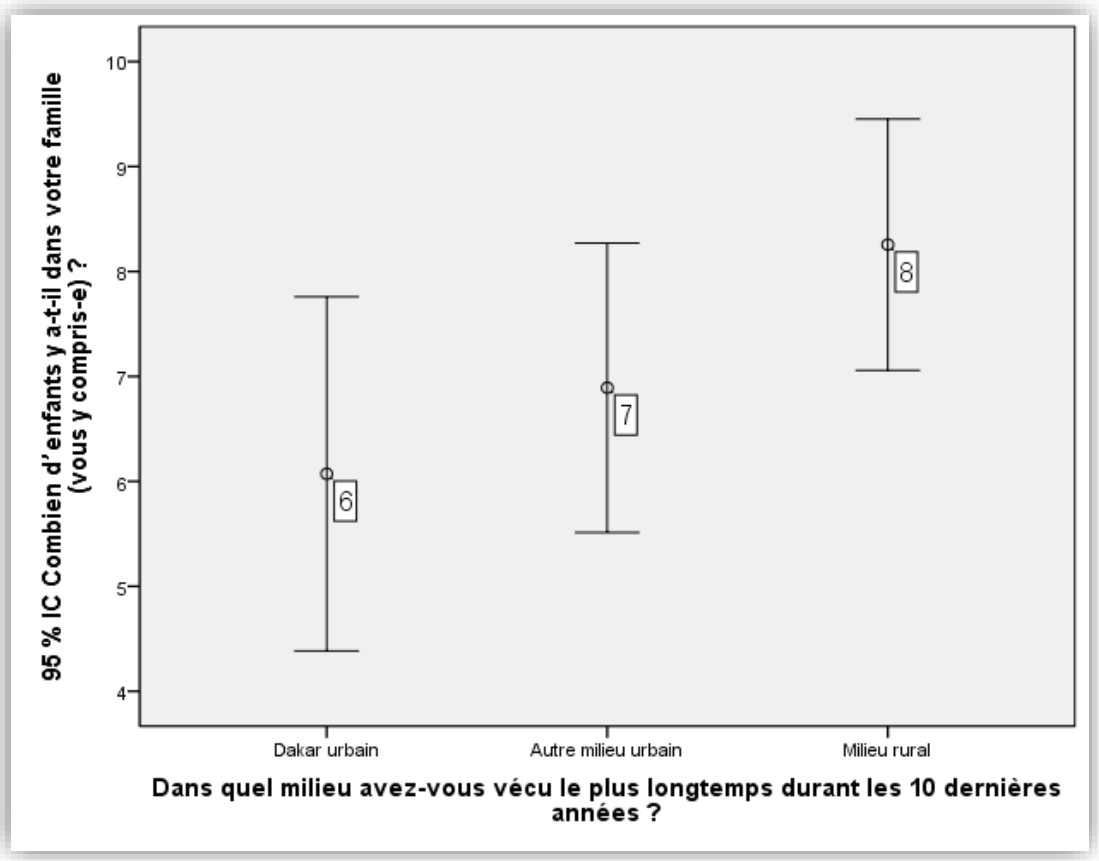
La barre d'erreur ci-dessus montre l'intervalle à l'intérieur duquel est susceptible de se retrouver la taille de famille moyenne dans la population étudiée : entre 6,57 et 8,15 enfants. Elle est particulièrement utile pour comparer des groupes eu égard à un phénomène quantitatif. Par exemple, on peut comparer la taille de famille moyenne (**Nbenf**) entre les résidents de Dakar urbain, les résidents d'un autre milieu et les résidents du milieu rural (**milieu**) pour voir s'il y a une différence statistiquement significative dans la population étudiée. On peut supposer que les étudiants du monde rural appartiennent à des familles plus nombreuses. Suivons les étapes suivantes :



Cliquez sur « Simple », puis sélectionnez « Récapitulatifs pour groupes d'observations » (la variable comparative étant catégorielle) et définissez pour obtenir la fenêtre ci-dessous!



Mettez la variable d'intérêt (**Nbenf**) sous *Variable*, et la variable de comparaison (**Milieu**) sous *Axe des catégories*. Validez pour afficher les barres d'erreur!



Dans le diagramme de variation ci-dessus, la barre d'erreur est plus haute chez les résidents du milieu rural, qui ainsi ont un nombre d'enfants plus élevé dans leur famille (8 enfants en moyenne), comparés aux résidents de la région dakaroise urbaine (6 enfants en moyenne). Mais, les barres n'étant pas **disjointes** (elles se chevauchent), les différences ne sont pas **significatives** au niveau de la population étudiée. Les différences observées dans l'échantillon semblent, en effet, être dues au hasard de l'échantillonnage. Évidemment, il aurait suffi d'augmenter la taille de l'échantillon pour que les différences se révèlent probablement significatives.

4.4. Exercice pratique

Vous vous intéressez à la variation du temps passé sur Internet en moyenne par jour (**HeureNet**) selon différents sous-groupes dans l'ensemble de la population étudiée.

a) Faire calculer d'abord l'intervalle de confiance autour de cette moyenne à 95%.

b) Ensuite, interprétez statistiquement chacun des résultats.

c) À l'aide de la barre d'erreur, représentez graphiquement l'intervalle de confiance autour du temps sur Internet (**HeureNet**) pour les hommes et pour les femmes (**Sexe**). La différence est-t-elle significative statistiquement?

d) À l'aide de la barre d'erreur, représentez graphiquement l'intervalle de confiance autour du temps sur Internet (**HeureNet**) pour les trois milieux de provenance (**Milieu**). Y a-t-il des différences significatives statistiquement?