Université Mohamed Khider de Biskra
Faculté des Sciences Exactes et de la Vie
Département : Biologie
Module: Biostatistique
Année Univ: 2023-2024
Master 1 LMD Science Biologique (Tous les spécialités)

Prof Chala Adel

Protocole de TP 01 :

Validation du test de comparaison entre plusieurs échantillons indépendants

Test de normalité (Shapiro-Wilk ou de Kolmogorov-Smirnov)

Test d'homogénéité pour la variance (Test de Levene)

Introduction aux théories des probabilités et statistique

Lorsqu'on veut étudier les données relatives aux caractéristiques d'un ensemble d'individus ou d'objets il est difficile d'observer toutes les données lorsque leurs nombres sont élevés. Au lieu d'examiner l'ensemble qu'on appelle population on examine un nombre restreint qu'on appelle échantillon, pour être représentatif l'échantillon doit être pris au hasard (une population peut être finie ou in…nie).

Population : C'est l'ensemble sur lequel porte l'étude statistique.

Individus : Les éléments de cet ensemble.

Echantillon : Est un sous-ensemble de la population.

Caractère : C'est le trait (ou propriété) choisi pour l'étude statistique.

Modalités : Les différentes positions que peut prendre un caractère. Usage on numérote les modalités de 1 à k la modalité numéro i est notée C_i

Effectifs : Lorsque la population est répartie sur les différentes modalités nous obtenons pour chacune d'elles un nombre c'est le nombre des individus ayant cette modalité. On note habituellement n_i l'effectif correspondant à la modalité C_i : les fréquences absolu.

Fréquence relative : Par définition c'est le rapport entre n_i et N, où N est la somme totale des individus.

Nous allons ainsi adopter les définitions suivantes :

*Un caractère est dit quantitatif quand ses différentes modalités sont mesurables par des nombres qui en indiquent l'intensité.

*Un caractère est dit qualitatif quand ses différentes modalités ne peuvent être désignées que par leurs qualités.

*Une variable statistique est dite discrète lorsque ses modalités ne peuvent être que des nombres isolés.

*Une variable statistique est dite continue quand elle peut prendre n'importe quelle valeur dans un intervalle donné.

Mode : c'est la valeur la plus fréquente.

Médiane : C'est la valeur de la variable statistique qui partage la population en deux populations d'effectifs égaux.

Quartiles : Comme on a défini la médiane on peut définir des paramètres qui la répartissent en quarts.

Moyenne arithmétique : est égale par définition :

b/ Caractéristiques de dispersion :

Etendue : C'est la longueur de l'intervalle sur lequel se disperse la variable.

Ecart-interquartiles : C'est la différence entre les deux quartiles Q_1 et Q_3:

Variance : C'est la caractéristique qui est réellement utilisée pour mesurer la dispersion :

Validation du test de comparaison entre plusieurs échantillons indépendants

Pour appliquer le test de comparaison que nous avons déjà vu dans l'année passée *L3*, qui peut être une comparaison entre deux échantillons (test de Student), soit une comparaison entre plus que trois échantillons (Test ANOVA), il faut que les conditions suivantes soient valides :

a) Une seule variable quantitative mesurable, et une variable qualitative avec deux modalités (pour test de Student), ou avec k modalités pour test d'ANOVA.
b) La distribution de la variable quantitative soit gaussienne.
c) L'échantillon est homogène pour la variable quantitative. (toutes les valeurs de la variable sont proches de la moyenne), on ne peut pas trouver des valeurs plus loin que la moyenne.
d) Echantillon doit être pris au hasard.

Pour cela notre objectif dans cette TP est de savoir comment vérifier la condition de normalité (c'est la condition de validité d'un test de comparaison d'anova, et aussi de vérifier l'homogénéité de la variation pour échantillon.

En utilisant pour cette raison les techniques du test de normalité soit de Shapiro-Wilk (le cas des petits échantillons) ou bien test de normalité de Kolmogorov-Smirnov (le cas des grands échantillons), par utilisation du logiciel SPSS.

I) Test de normalité

On utilise ce type du test lorsqu'on a une seule variable quantitative.

Pour cela on doit prendre un exemple d'explication :

Exemple 1

Dans des études d'anesthésie, voulant tester la normalité de la distribution ainsi l'homogénéité (la durée de somnifères), on a noté les durées de sommeil qui ont suivi les injections d'une dose bien définie. Les durées étant exprimées en minutes :

Somnifère 01	170	175	187	180	190	165	175	174	173	181		
Somnifère 02	133	143	152	122	132	134	126	129	120	131	135	
Somnifère 03	155	160	164	150	160	159	154	156	160	167	153	158

1) Déterminer l'objectif pour cette expérience.
2) Déterminer la variable qualitative qui détermine les trois échantillons, et la variable quantitative mesurable.
3) Déterminer l'hypothèse nulle et alternative pour la normalité et pour l'homogénéité de la variation.
4) Tracer le tableau de de la statistique descriptive.
5) Avec un risque de signification 6%, que peut-on dire sur la normalité de la variable quantitative ? et sur l'homogénéité sur les modalités de la variable quantitative ?

Remarque 01:

Pour test de la normalité, on utilise test de Kolmogorov-Smirnov lorsque la taille des échantillons est très élevée (n>30).

On utilise test de Shapiro-Wilk lorsque la taille des échantillons est assez petit (n<30).

Remarque 02:

Pour la décision, on utilise souvent la règle suivante :

Si Signification (bilatérale) inférieure à $\alpha\%$, alors on rejette H0.

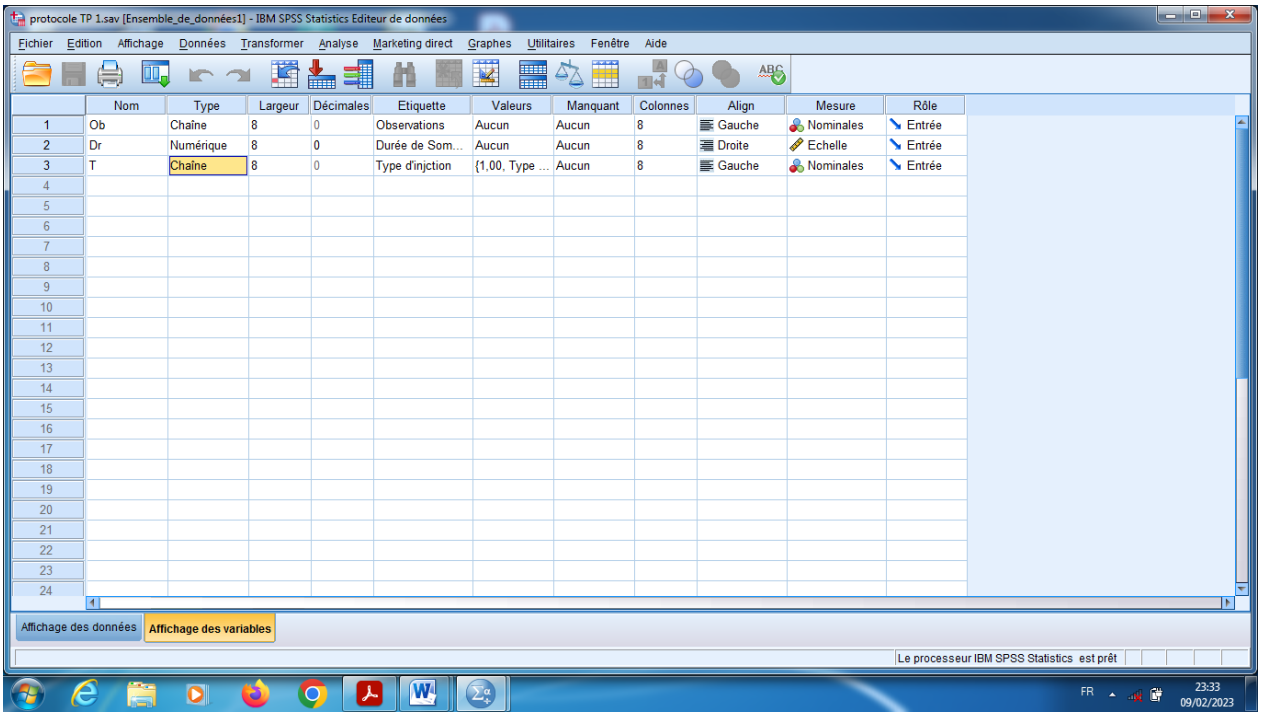
Si Signification (bilatérale) supérieure à $\alpha\%$ alors on accepte H0.

Solution d'exemple

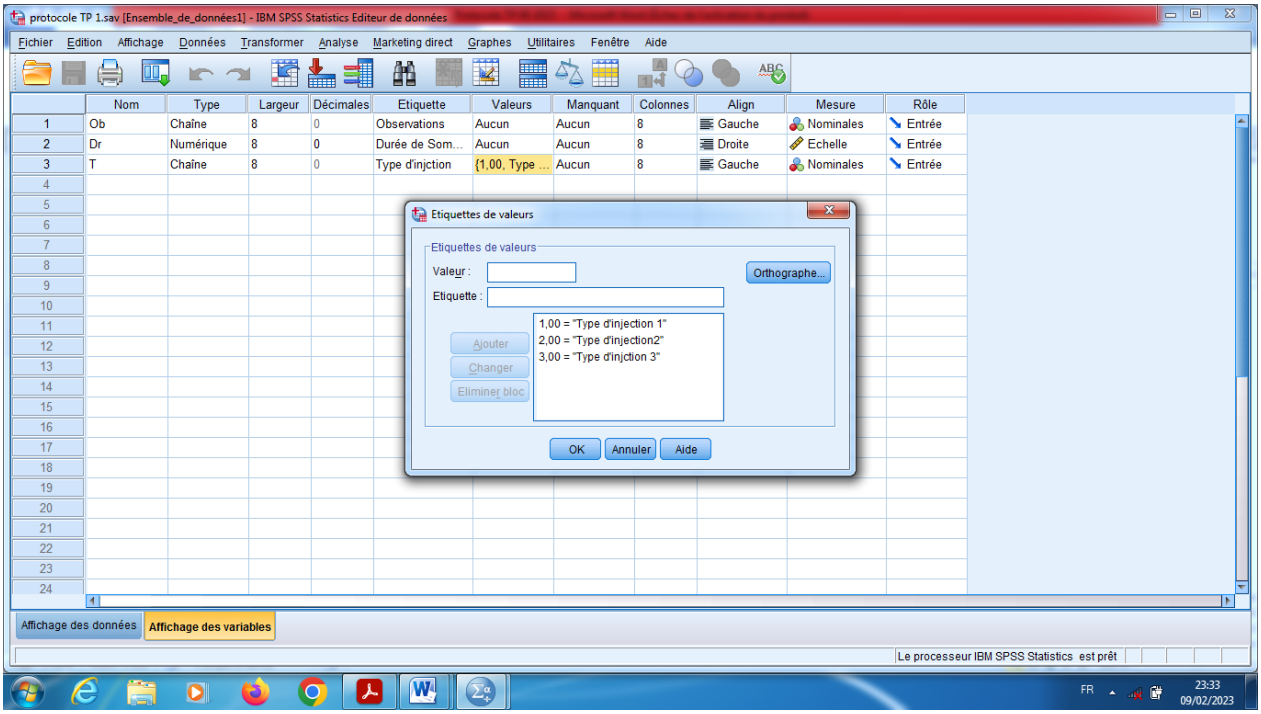
Tout d'abord il faut saisir ces données dans SPSS.

En suivant les étapes suivantes :

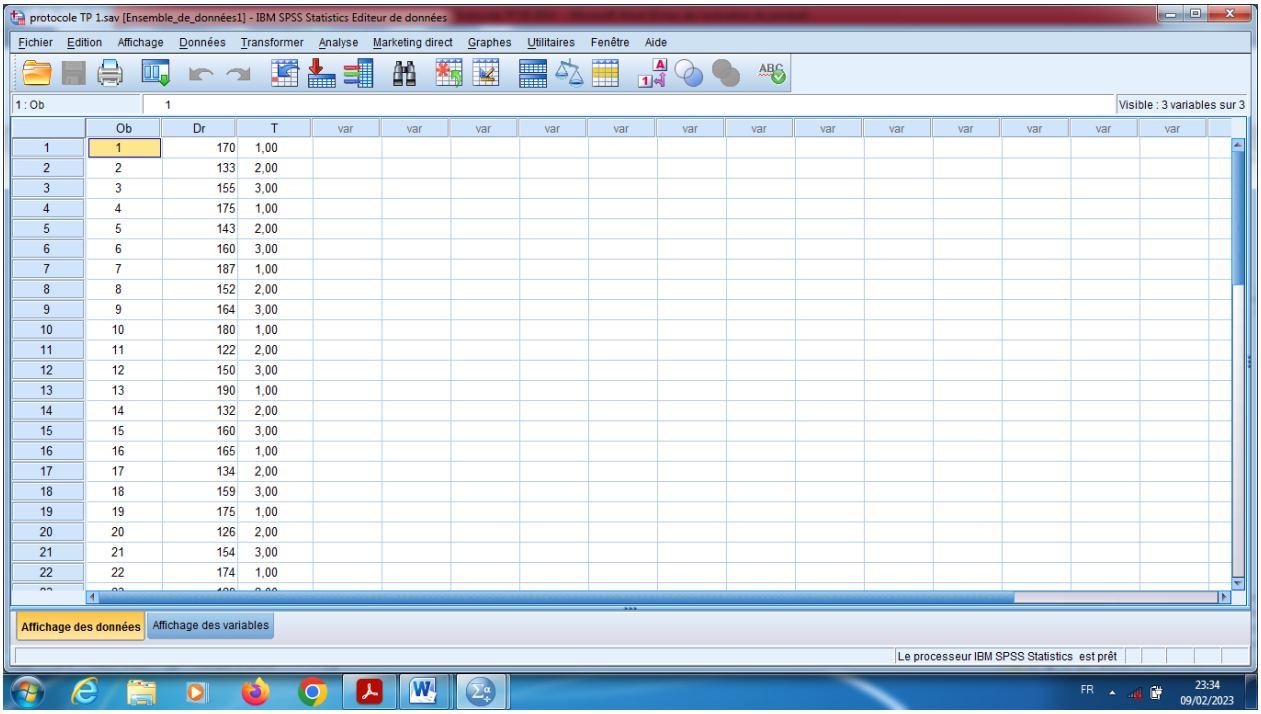
a) Il faut définir dans la barre en bas « Affichage des variables » : les variables qualitative et quantitative suivantes : observations, la durée de sommeil, et les type de somnifères.



b) On choisit les modalités pour la variable qualitative qui représente les trois échantillons de somnifères dans icône « valeurs », on peut prendre comme un exemple (la valeur 1 pour le type de somnifère 1 et la valeur 2 pour le type de somnifère 2 et 3 pour le type de somnifère3).

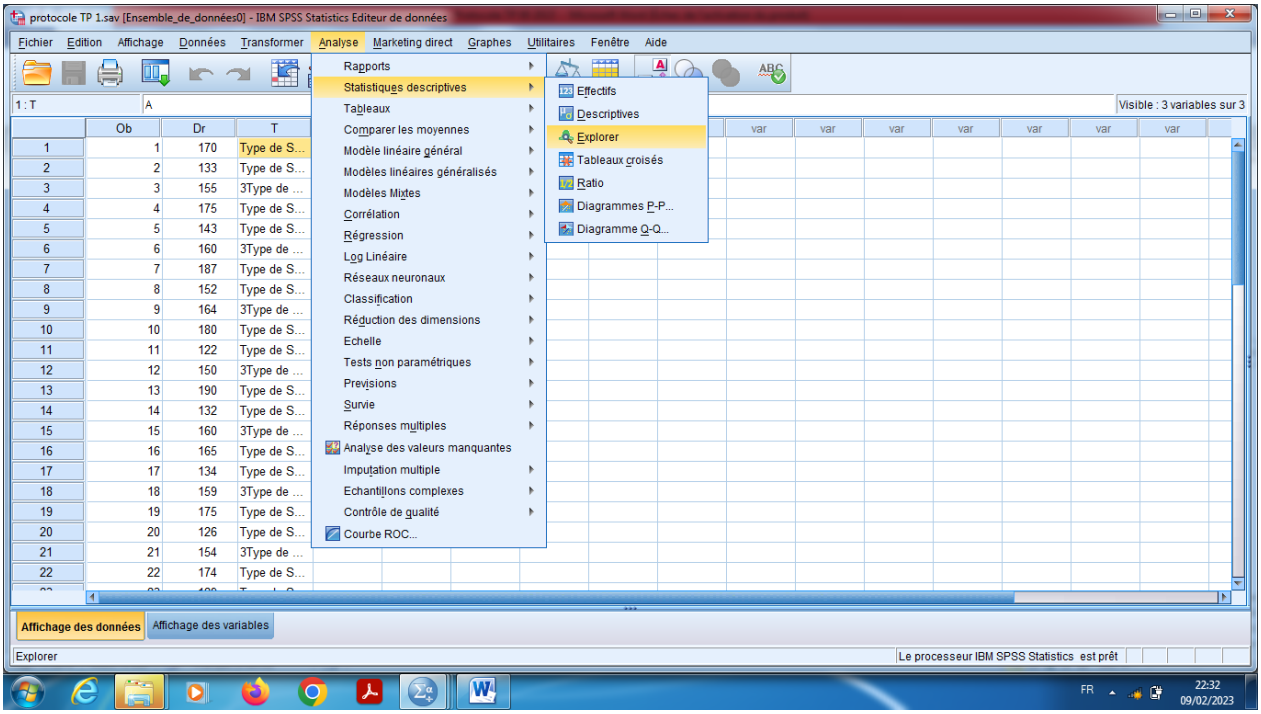


c) On introduit les données dans la barre « Affichage des données ».



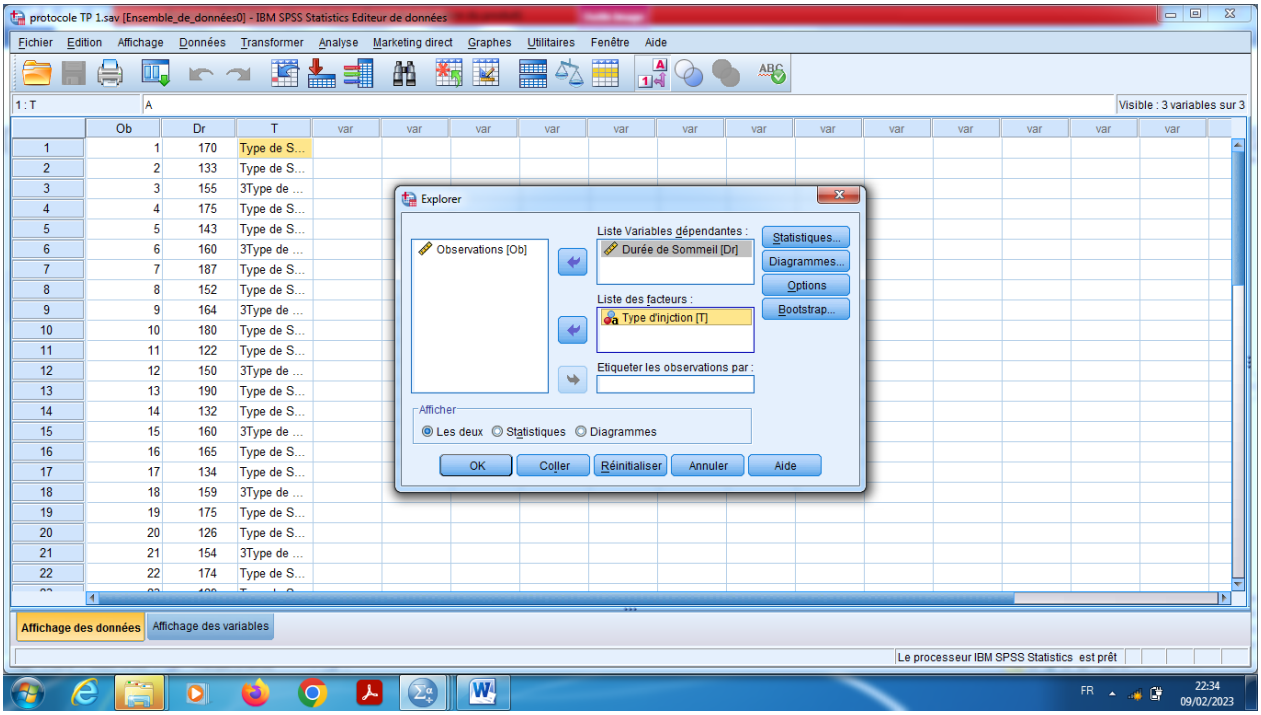
d) En cliquant sur le bouton « Etiquettes des valeurs » pour visualiser le codage des échantillons.

e) En suit, pour obtenir le test de normalité (c'est-à-dire que la variable quantitative à mesuré suit la loi de Gauss), en cliquant sur le bouton « Analyse » qui se trouve dans la barre des outils, et choisir « Statistiques descriptives», et puis « Explorer».



f) On choisit dans « liste variables dépendantes» la variable quantitative mesurable « Durée de Sommeil».

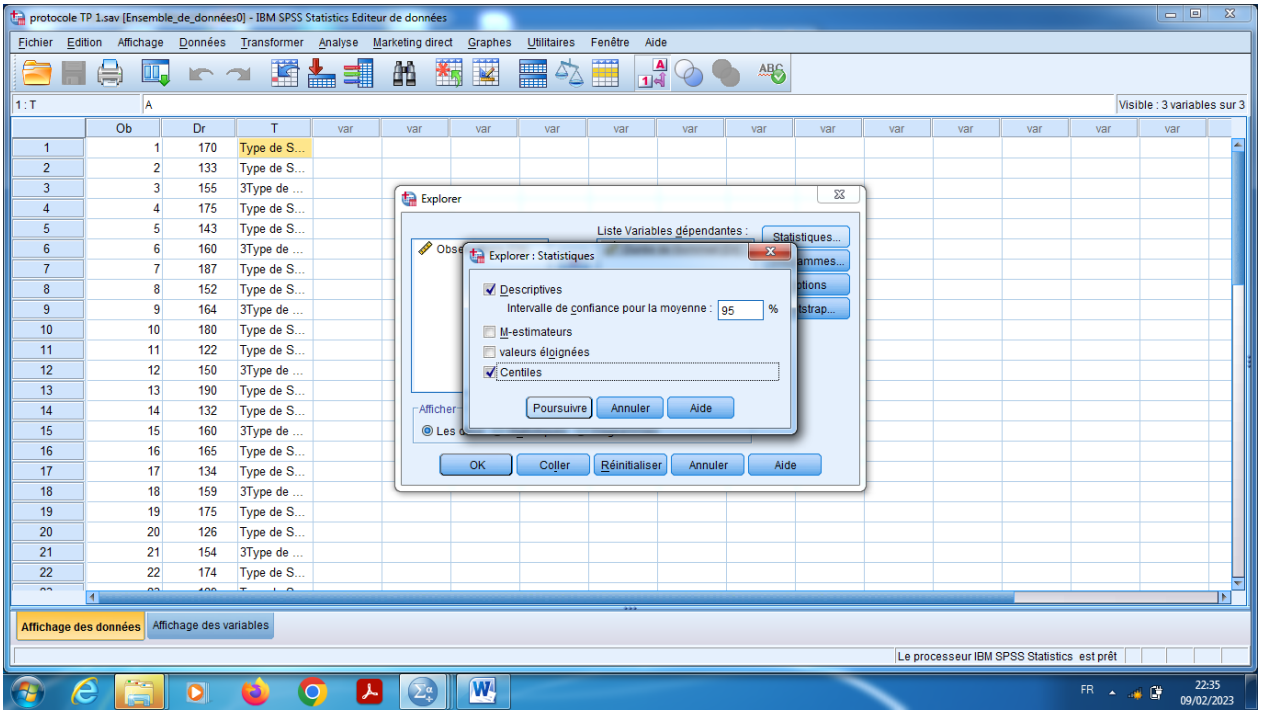
g) On choisit dans « liste des facteurs» la variable qualitative (Echantillons=type de somnifères)



h) Dans le Bouton « Statistiques » on peut changer le pourcentage pour l'intervalle de confiance le seuil de signification 95%.

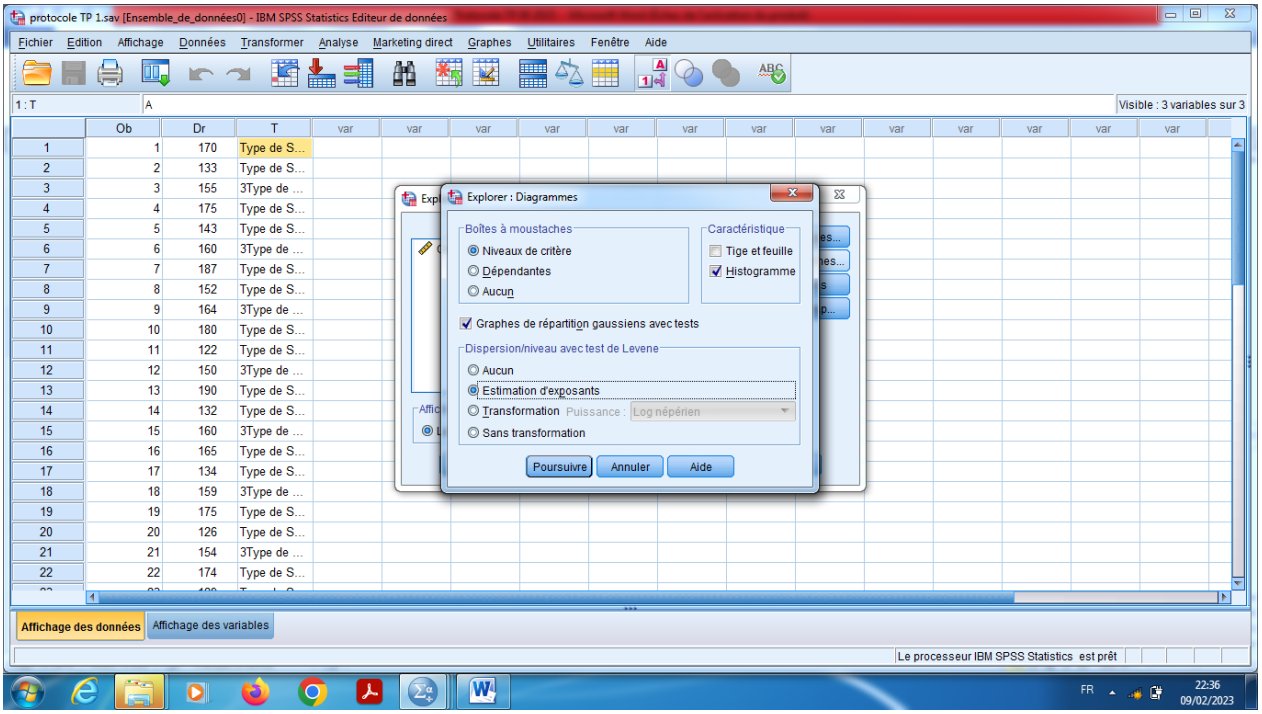
i) On choisit « Centiles » pour visualiser les quartiles (Q1, Q2, Q3).

j) En cliquant sur « poursuivre ».



k) Sur le bouton « Diagrammes », on garde dans « Boites à moustaches » le niveau de critère, et dans « Caractéristiques » en cliquant sur le choix Histogramme, et aussi on coche sur le choix « Graphes de répartitions gaussiens avec tests » pour validé test de normalité, en fin sur «Dispersion /Niveau avec test de Levene » on prend le choix « Estimation d'exposants » pour validé test d'homogénéité.

l) En cliquant sur « poursuivre ».



En fin en cliquant sur « Ok ».

Interprétation des résultats

1) Notre objectif est de valider les conditions de comparaison entre les trois types des somnifères (Test d'ANOVA) qui s'appelle test de normalité c'est à dire que la variable quantitative (Duré de sommeil) est Gaussienne, ainsi que le test d'homogénéité de la variation sur échantillon.
3) Proposition d'hypothèses :

a) Pour la normalité :

Hypothèse nulle, H0 : « La distribution de la variable Duré de sommeil est gaussienne».

Hypothèse alternative, H1 : « La distribution de la variable Duré de sommeil n'est pas gaussienne».

b) Pour l'homogénéité

Hypothèse nulle, H0 : « La variation de la variable Duré de sommeil est Homogène».

Hypothèse alternative, H1 : « La variation de la variable Duré de sommeil n'est pas homogène».

4) Dans la table 1 «Statistique descriptive» on remarque que 'il y a 10 observations dans les somnifères 1, et 11 observations pour la somnifère 2, et 12 observation pour la somnifère 3, et non des valeurs manquantes.

Récapitulatif du traitement des observations

	Type d'injction	Observations					
		Valide		Manquante		Total	
		N	Pourcent	N	Pourcent	N	Pourcent
Durée de Sommeil	Type d'injection 1	10	100,0%	0	0,0%	10	100,0%
	Type d'injection2	11	100,0%	0	0,0%	11	100,0%
	Type d'injction 3	12	100,0%	0	0,0%	12	100,0%

(Table 01)

Dans la table 2 : Les moyennes 177 pour le somnifère 1 et 132,45 pour la somnifère 2 et 158 pour la somnifère 3, de plus pour l'écart-type c'est 7,601 et 9,070 et 4,74821 respectivement pour la somnifère 1 et 2 et 3.

Pour les médianes 175 et 132,45 et 158,5 respectivement pour les somnifères 1 et 2 et 3.

Descriptives

	Type d'injction			Statistique	Erreur standard
Durée de Sommeil	Type d'injection 1	Moyenne		177,00	2,404
		Intervalle de confiance à 95% pour la moyenne	Borne inférieure	171,56	
			Borne supérieure	182,44	
		Variance		57,778	
		Ecart-type		7,601	
		Minimum		165	
		Maximum		190	
		Intervalle		25	
		Intervalle interquartile		10	

	Type d'injection2	Moyenne		132,45	2,735
		Intervalle de confiance à 95% pour la moyenne	Borne inférieure	126,36	
			Borne supérieure	138,55	
		Médiane		132,00	
		Variance		82,273	
		Ecart-type		9,070	
		Minimum		120	
		Maximum		152	
		Intervalle		32	
		Intervalle interquartile		9	
	Type d'injction 3	Moyenne		158,00	1,371
		Intervalle de confiance à 95% pour la moyenne	Borne inférieure	154,98	
			Borne supérieure	161,02	
		Médiane		158,50	
		Variance		22,545	
		Ecart-type		4,748	
		Minimum		150	
		Maximum		167	
		Intervalle		17	
		Intervalle interquartile		6	

(**Table 02**)

5) **<u>Pour le test de normalité,</u>** il faut choisir le test de Shapiro-Wilk (car n_1=10 et n_2=11 et n_3=12 sont tous inferieurs à 30).

On remarque que dans somnifère 1 (Sig=0,828>0,05) on accepte H0, alors la distribution de la variable durée pour échantillon 1 est gaussienne.

De plus dans somnifère 2 (Sig=0,492>0,05) on accepte H0, alors la distribution de la variable durée pour échantillon 2 est gaussienne.

D'autre part dans somnifère 3 (Sig=0,969>0,05) on accepte H0, alors la distribution de la variable durée pour échantillon 3 est gaussienne.

Tests de normalité

	Type d'injction	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistique	ddl	Signification	Statistique	ddl	Signification
Durée de Sommeil	Type d'injection 1	,204	10	,200*	,964	10	,828
	Type d'injection2	,208	11	,200*	,937	11	,492
	Type d'injction 3	,170	12	,200*	,977	12	,969

<u>Test d'homogénéité de la variance</u>

Mais pour l'homogénéité de la variance pour la variable quantitative « Durée de Sommeil », on peut l'établir à partir de la deuxième table : La durée de sommeil qui basé sur la moyenne que la distribution est homogène, c'est-à-dire qu'il n ya pas des valeurs critiques pour cette variable, (échantillon est homogène par rapport à la variable durée).

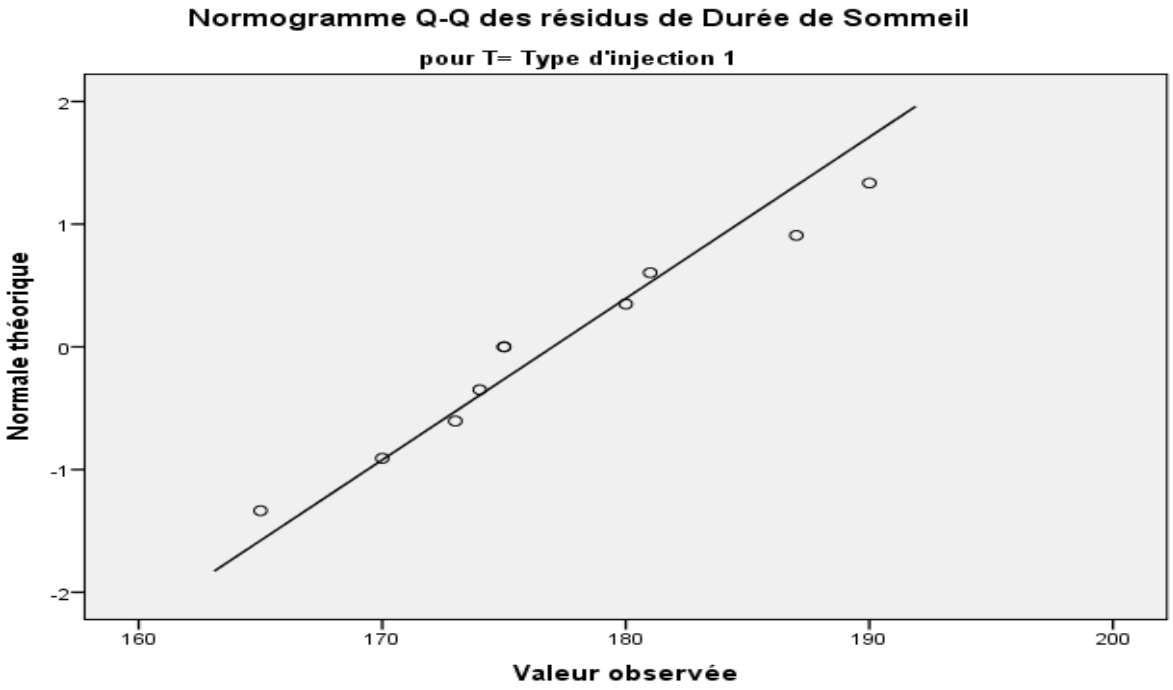
Test d'homogénéité de la variance					
		Statistique de Levene	ddl1	ddl2	Signification
Durée de Sommeil	Basé sur la moyenne	1,145	2	30	,332

Test de normalité graphiquement

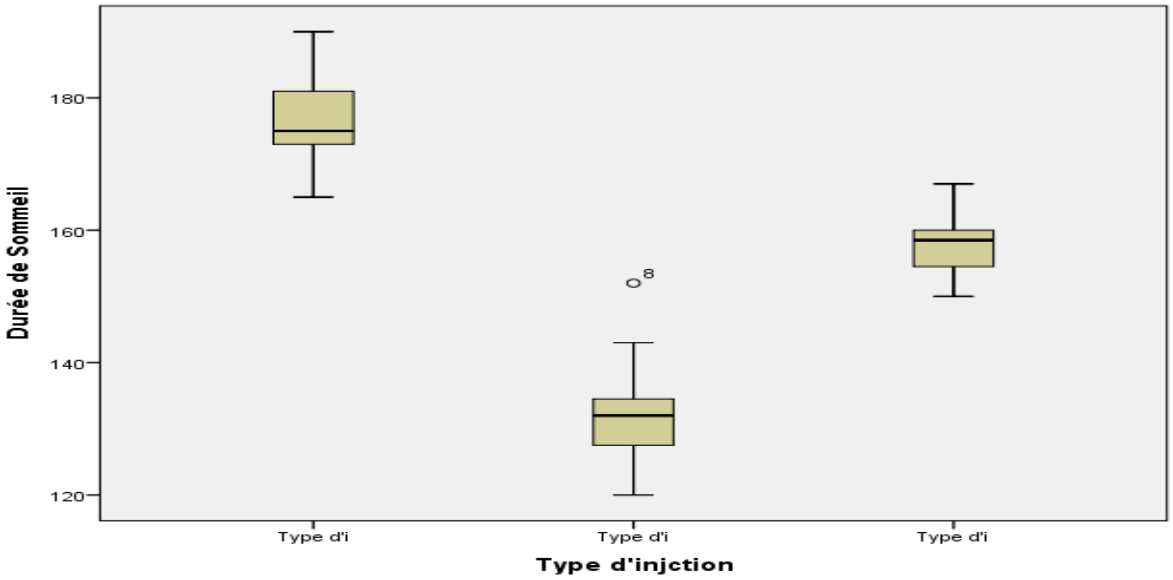
D'autre part on peut verifier la validité du (test de comparaison entre les moyennes) en utilisant quelques type des graphes.

Diagramme Q-Q gaussien

Pour Somnifère 1 :



Pour Somnifère 1 : On compare la droite de normalité avec nuage des points, on remaque facilement que les pluparts des points se trouve au voisinage de la droite.



On remarque ici que dans la deuxième échantillon que l'observartion N° :8 n'est pas bien définit, c'est-à-dire il faut que refaire encore fois cette experiance dans la 8^{ime} observation.

Conclusion :

D'après de tout ce qui précède, on conclut que la distribition du paramètre quantitative « Durée de sommeil » est gaussien au taux de confiance 95%, de plus l'échantillon est homogène pour la variation, alors on peut appliquer le test de comparaison des moyennes.