

Module : **Bioinformatique**

Enseignant de la matière : MERADSI Fouad

Semestre 6

Année universitaire : 2019 / 2020

Destiné pour les étudiants de Licence :

- 1- Biologie et physiologie végétale
- 2- Biotechnologie et génomique végétale

Remarque

En utilisant comme logiciel statistique le **SPSS**, il faut télécharger ce logiciel.

Introduction

La statistique a envahi aujourd'hui tous les champs scientifiques. Les statistiques, dans le sens populaire du terme, traitent des populations (dans des études démographiques), ce qui est très difficile dans le cas de la Bio-statistique. La statistique constitue, en science, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence.

Le traitement A est-il plus efficace que le traitement B ?

Leur objectif consiste à caractériser une population à partir d'une image plus ou moins floue constituée à l'aide d'un échantillon issu de cette population. *On peut alors chercher à extrapoler une information obtenue à partir de l'échantillon.*

On trouve des applications de la statistique dans tous les domaines : industrie, environnement, médecine, finance, marketing, sport, ... etc.

I- PARTIE THEORIQUE

Généralités et notions de base

Définition de la Bio-statistique

Ensemble de méthodes à partir desquelles on recueille, organise, résume, présente et analyse des données afin d'en tirer des conclusions et de prendre des décisions avec prudence.

Notions importantes

Parmi les notions importantes nous avons :

La variabilité

Disposition à varier, qualité de ce qui est variable.

La variabilité en biologie est la somme d'une variabilité métrologique et d'une variabilité proprement biologique.

Population : Ensemble des individus objets de l'étude, ou Ensembles des *éléments* ou *d'individus* de même nature, visés par une problématique scientifique.

Élément : Les éléments sont les unités qui composent une population.

Synonymes : Objet, individu, unité statistique, unité d'échantillonnage, sujet, événement, comportement,...etc

Echantillon : C'est un sous ensemble de la population considérée, prélevé pour juger de cet ensemble.

Echantillon représentatif : Échantillon qui reflète fidèlement la complexité et la composition de la population. *Le tirage au sort* ainsi que *l'inventaire exhaustif* (recensement), sont deux façons d'obtenir un échantillon représentatif d'une population.

Caractère statistique (ou variable statistique)

C'est ce qui est observé ou mesuré sur les individus d'une population statistique. Ce variable peut être quantitative (numérique) ou qualitative (non numérique).

Variable quantitative : C'est un paramètre expérimental qui s'exprime par un nombre.

Pouvant être classées en *variables continue* (taille, poids) ou *discontinue (discrète)* (nombre d'enfants dans une famille, nombre d'œufs pondus par un oiseau).

Variable qualitative : Une variable qualitative se caractérise par un ensemble discontinu d'états.

Pouvant être classées en *variables catégorielles* (nominales) (couleurs des plumes des oiseaux) ou *ordinales* (résistance d'une plante vis-à-vis un ravageur classée en faible, moyenne et importante).

Notion d'hypothèse

L'hypothèse est une relation hypothétique (provisoire, postulée par le chercheur).

On distingue deux formes d'hypothèses :

Hypothèse nulle (H0) et Hypothèse significative (H1) ou alternative.

- **Hypothèse nulle (H0):** $m_1 = m_2$ ou l'absence d'une différence significative entre les moyennes ;
- **Hypothèse alternative (H1):** $m_1 \neq m_2$ ou l'existence d'une différence significative entre les moyennes.

Seuil de signification

En statistique, il n'existe pas de règle rigide permettant de tirer une conclusion concernant les hypothèses ; aucun test ne nous fournit une réponse en terme de oui ou non, mais indique dans quelle mesure nous pouvons être certain de tirer des conclusions ; cette mesure se nomme niveau ou seuil de signification, ou encore probabilité d'erreur.

La loi normale

Une distribution normale correspond à la distribution de probabilités d'une variable aléatoire continue dont la courbe est parfaitement symétrique et en forme de cloche.

Lorsqu'une variable (x) se distribue de telle sorte que les fréquences de ses différentes éventualités suivent la loi normale, alors elle est dite variable normale.

Types de test

On parle de *tests paramétriques* lorsque l'on stipule que les données sont issues d'une distribution paramétrée. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide de paramètres estimés sur l'échantillon (moyenne, mode et médiane). La distribution des données suit la loi normale.

Les *tests non paramétriques* ne font aucune hypothèse sur la distribution sous-jacente des données (la distribution des données ne suit pas la loi normale). On les qualifie souvent de *tests distribution free*.

II- PARTIE PRATIQUE

Ce module est basé essentiellement sur la pratique

Statistiques inférentielles

Les statistiques inférentielles, consistant en des tests permettant de confirmer ou infirmer une hypothèse.

Pour effectuer une analyse statistique (partie pratique), il faut suivre les six étapes suivantes :

1. Identification des variables ;
2. Introduction de données ;
3. Analyse proprement dite ;
4. Lecture, interprétation des résultats ;
5. Conclusion ;
6. Représentation graphique des résultats.

I. Variables quantitatives

1. Test Student

1.1 . Test Student pour deux échantillons indépendants

C'est un test permet de comparer deux moyennes de deux échantillons indépendants.

Exercice 01 : Le domaine vital de l'ours noir (km^2) a été mesuré pour des individus mâles et femelles.

Mâle ($n_1 = 6$)	Femelle ($n_2 = 9$)
94	37
504	72
173	60
274	49
560	102
168	18
	50
	49
	20

Existe-t-il une différence significative dans la taille du domaine vital chez le mâle et la femelle de l'ours noir ?

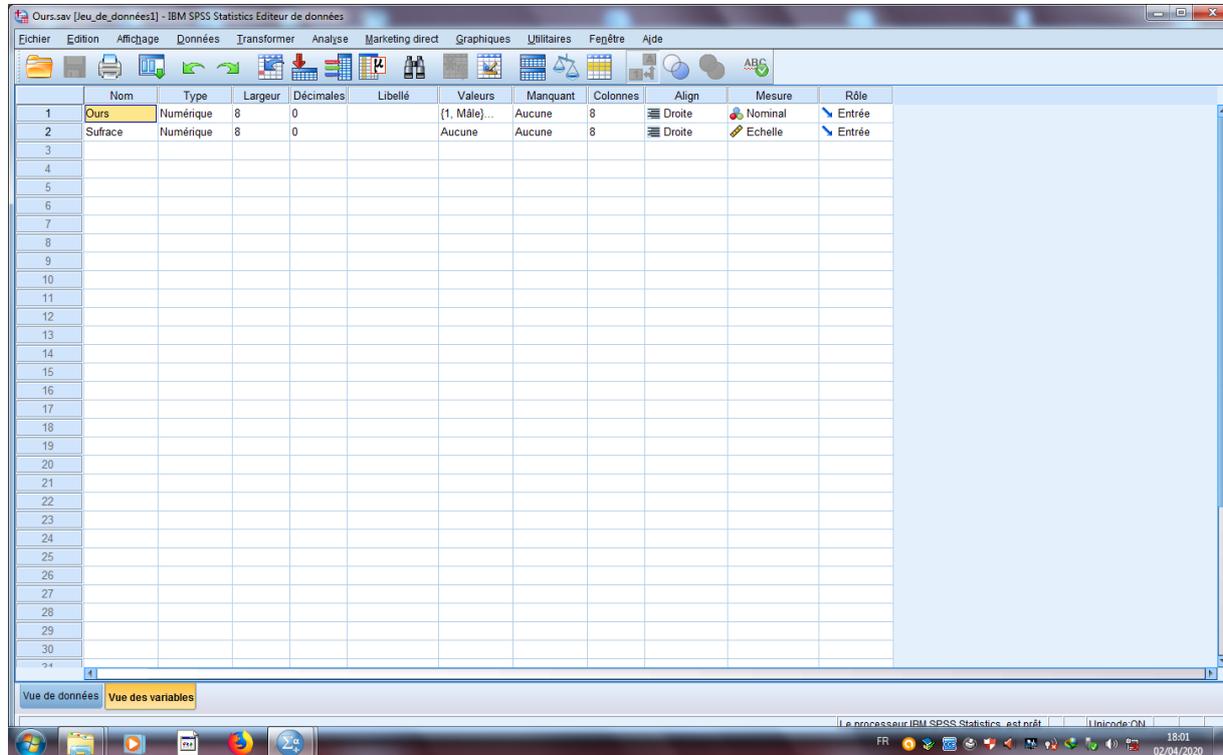
1. Identification des variables

Premièrement il faut commencer par l'identification des variables

Le facteur est nommé par **Ours** et la seule variable est la **Surface** occupée par l'animal

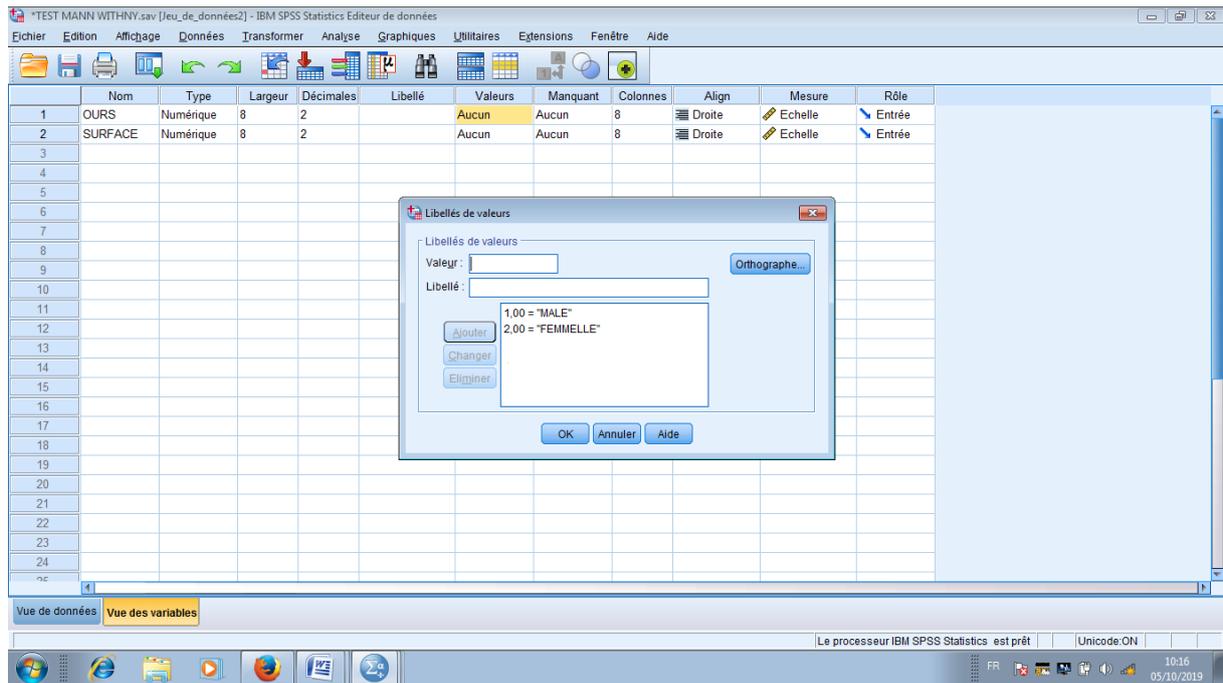
Cliquez premièrement sur l'option **Vue des variables** en bas de la page

Commencer toujours par le facteur (Ours), puis les variables (Surface).



Codez les deux modalités du facteurs

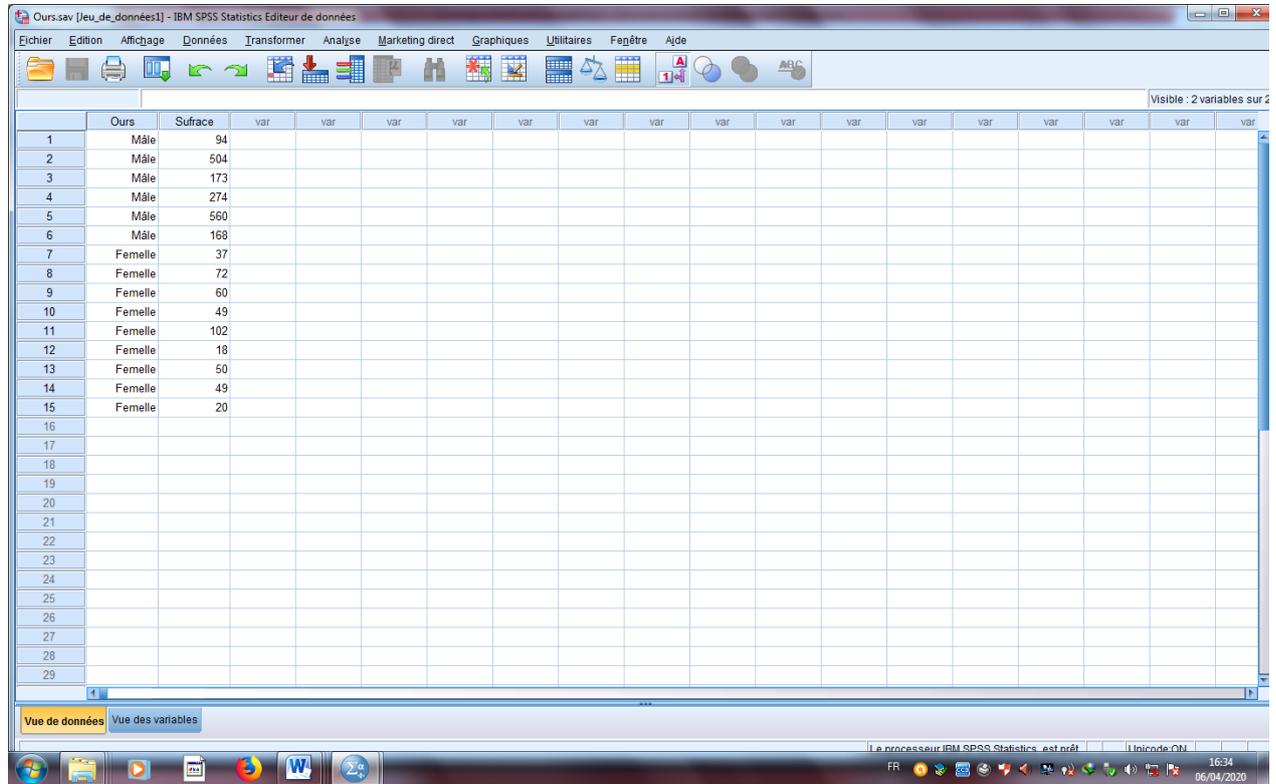
Le mâle par le code **1** et la femelle par le code **2**



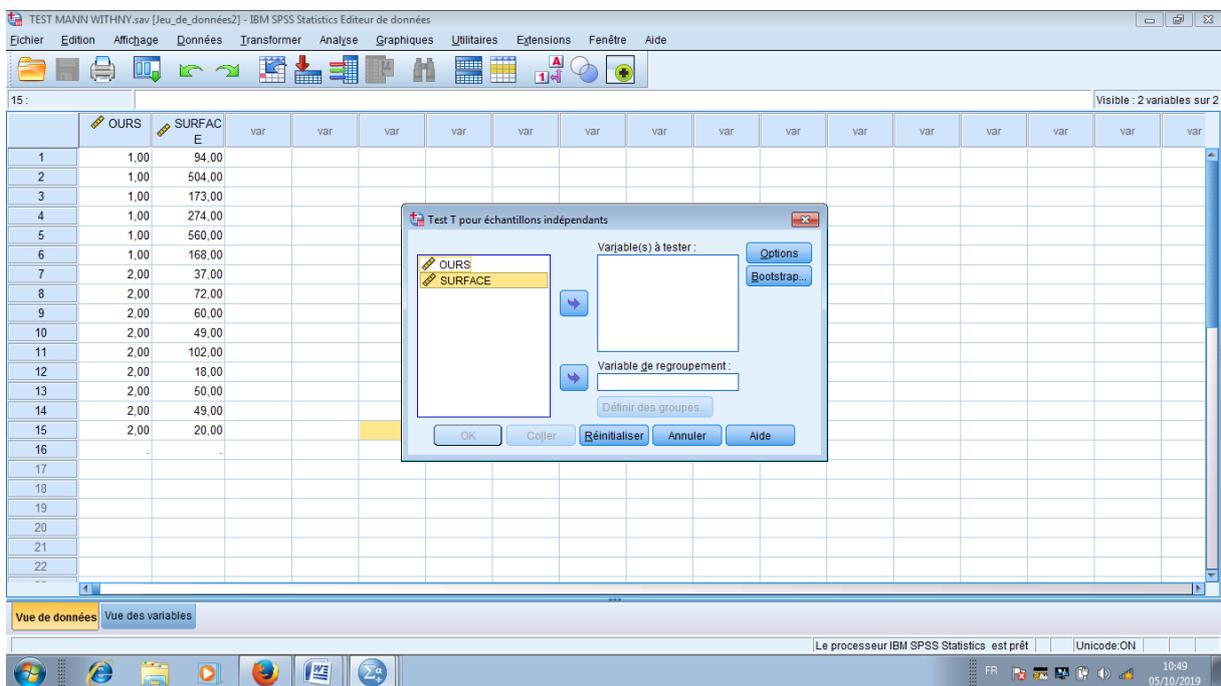
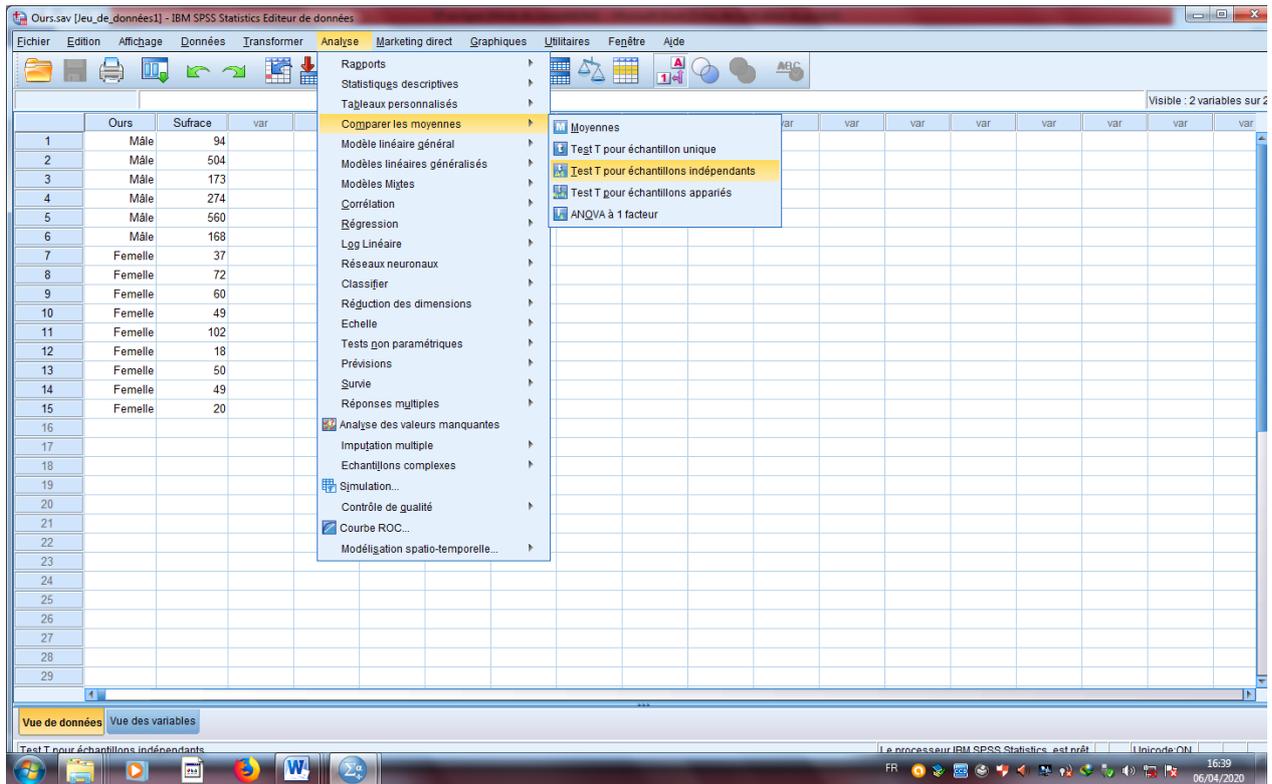
2. Introduction de données

Cliquez sur l'option **Vue de données** en bas de la page ;

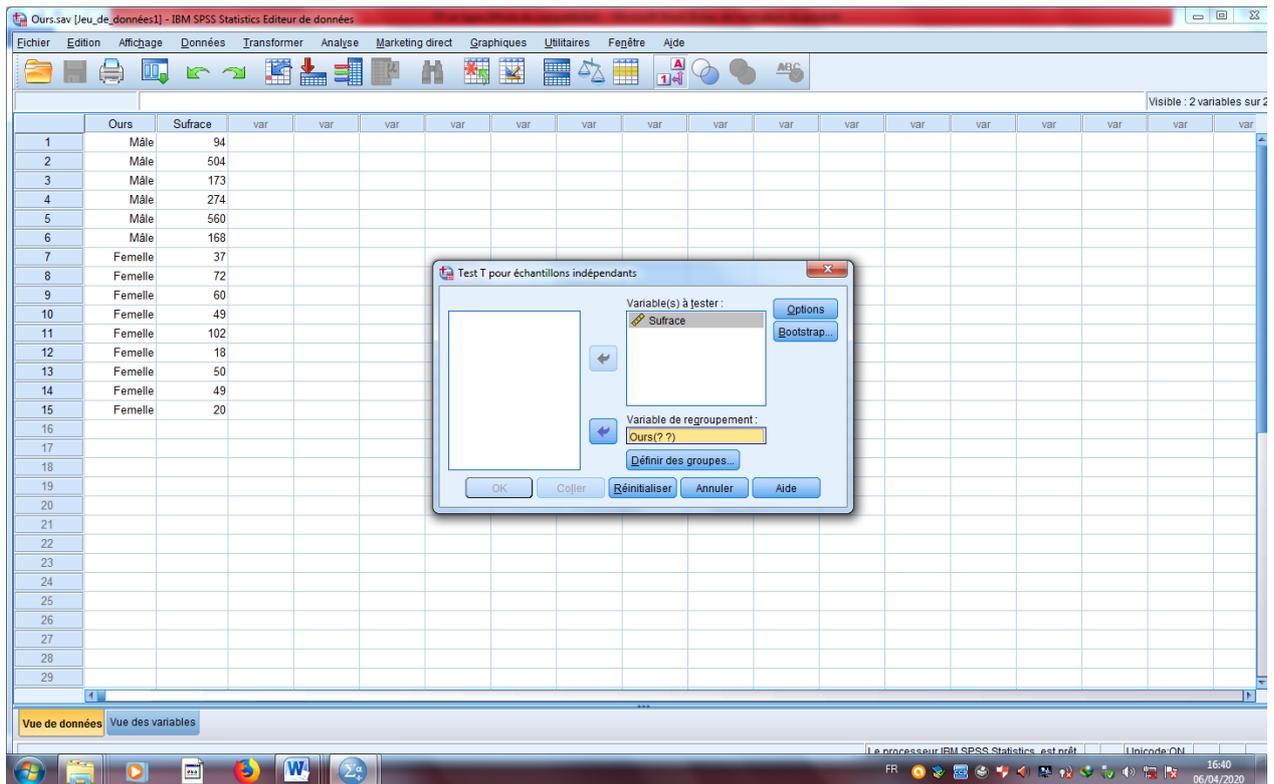
Il faut introduire les codes 1 et 2, puis les données pour chaque individu



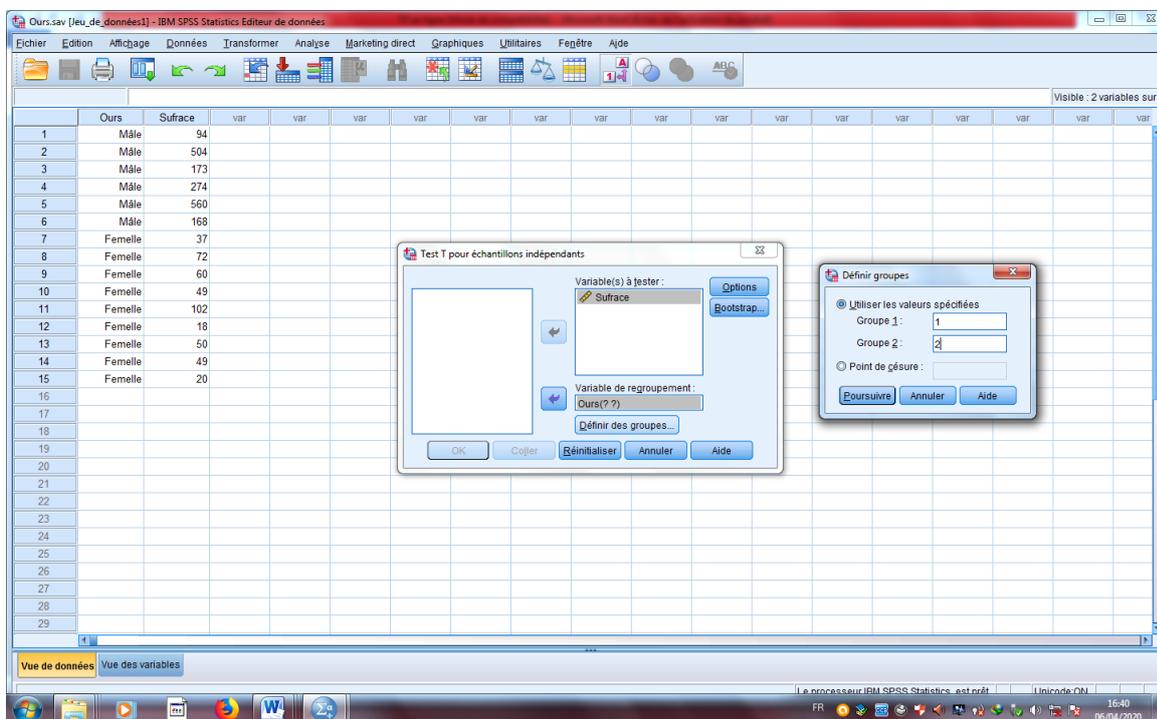
3. Analyse proprement dite



Déplacez l'Ours vers **variable de regroupement** et Surface vers **Variable à tester**



Il faut définir les groupes 1 et 2



4. Lecture, interprétation des résultats

Cliquez sur **poursuivre**, puis sur **OK** pour obtenir les résultats

	Ours	N	Moyenn e	Ecart type
Sufrace	Mâle	6	295,50	192,748
	Femelle	9	50,78	25,927

Il faut vérifiez les résultats de test d'égalité des variances puis choisir la bonne ligne : le test de Levene a montré que les variances sont inégales car le $P = 0,000$ c à d $P < 0,05$ (5%).

Il faut prendre les résultats de la deuxième ligne

$P = 0,026$; il existe une différence significatives entre les deux surfaces.

		Test de Levene sur l'égalité des variances				
		F	Sig.	t	ddl	Sig. (bilatéral)
Sufrace	Hypothèse de variances égales	23,346	,000	3,829	13	0,002
	Hypothèse de variances inégales			3,091		0,026

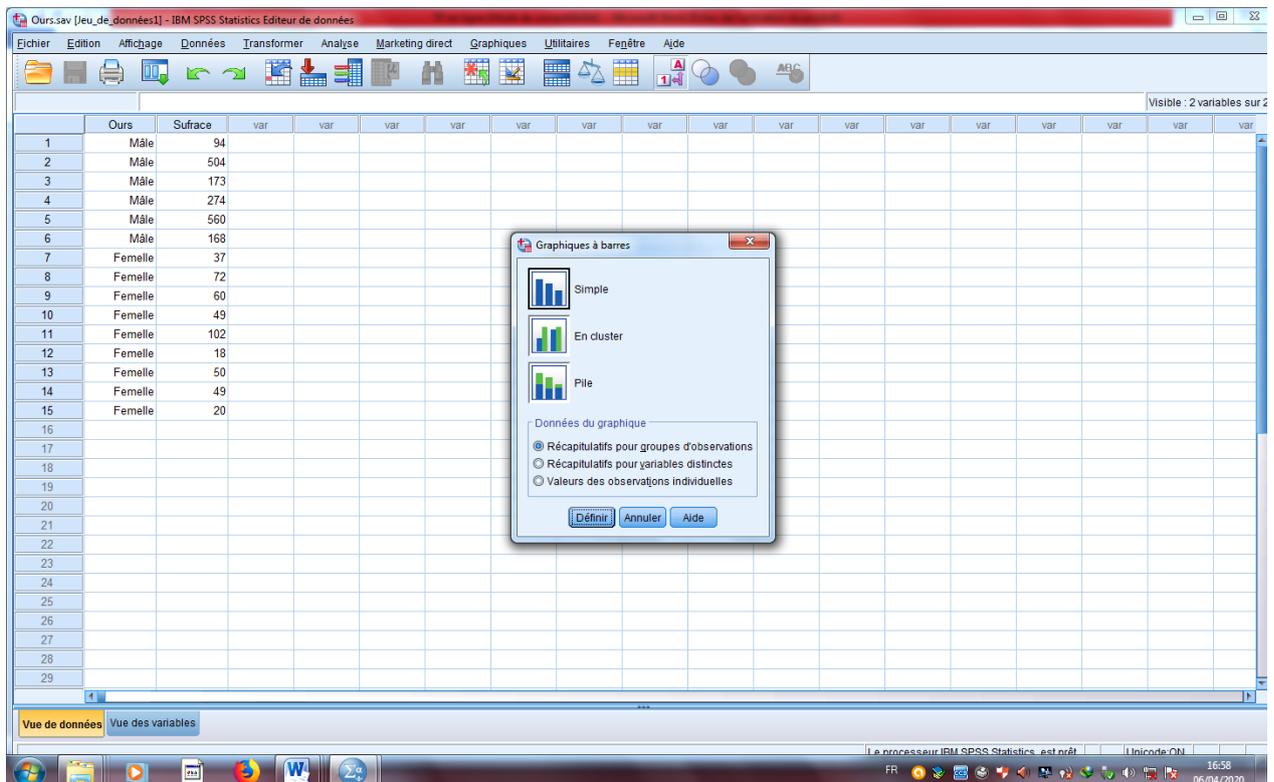
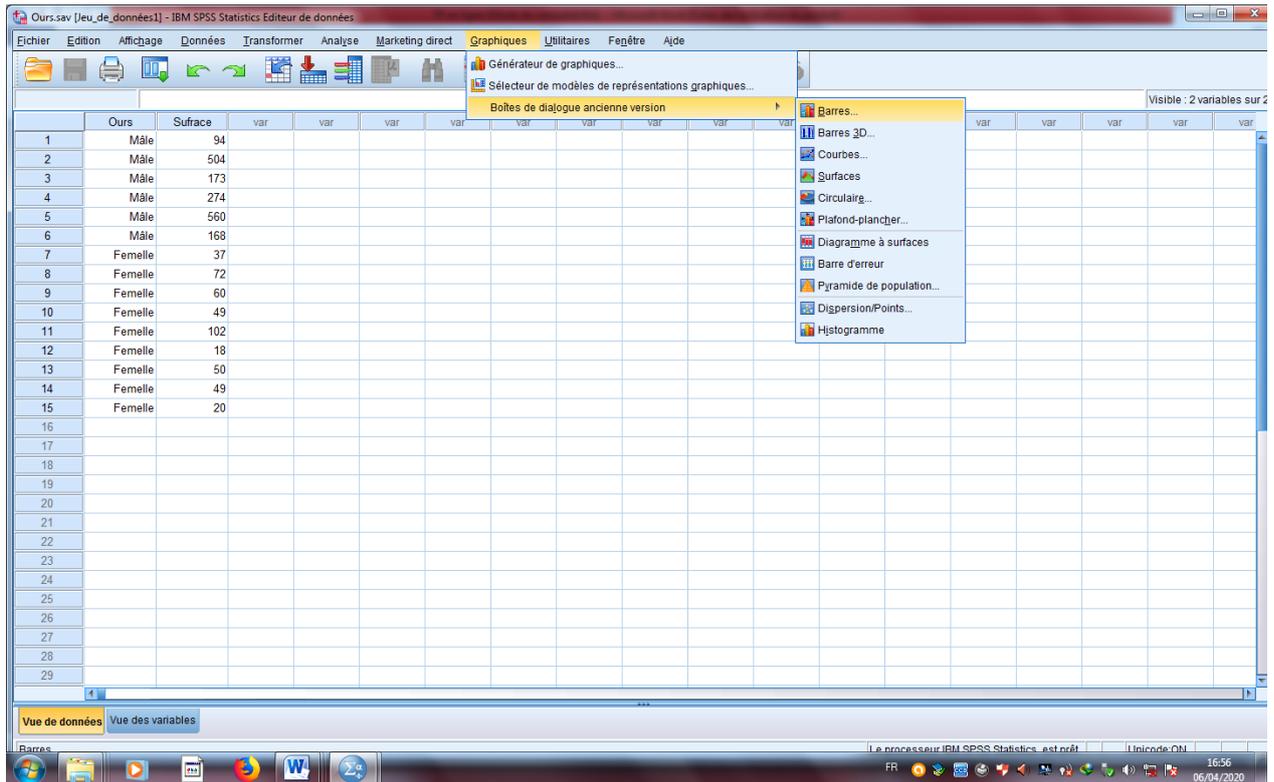
5. Conclusion

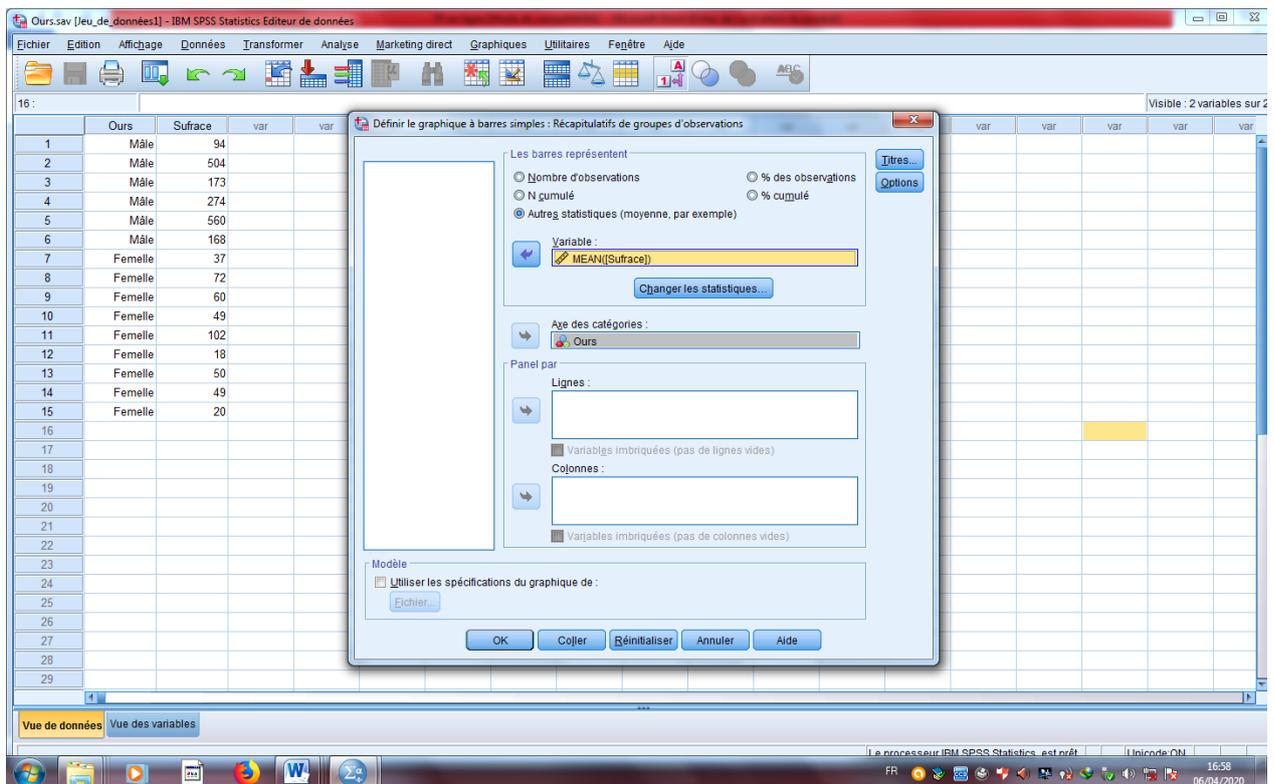
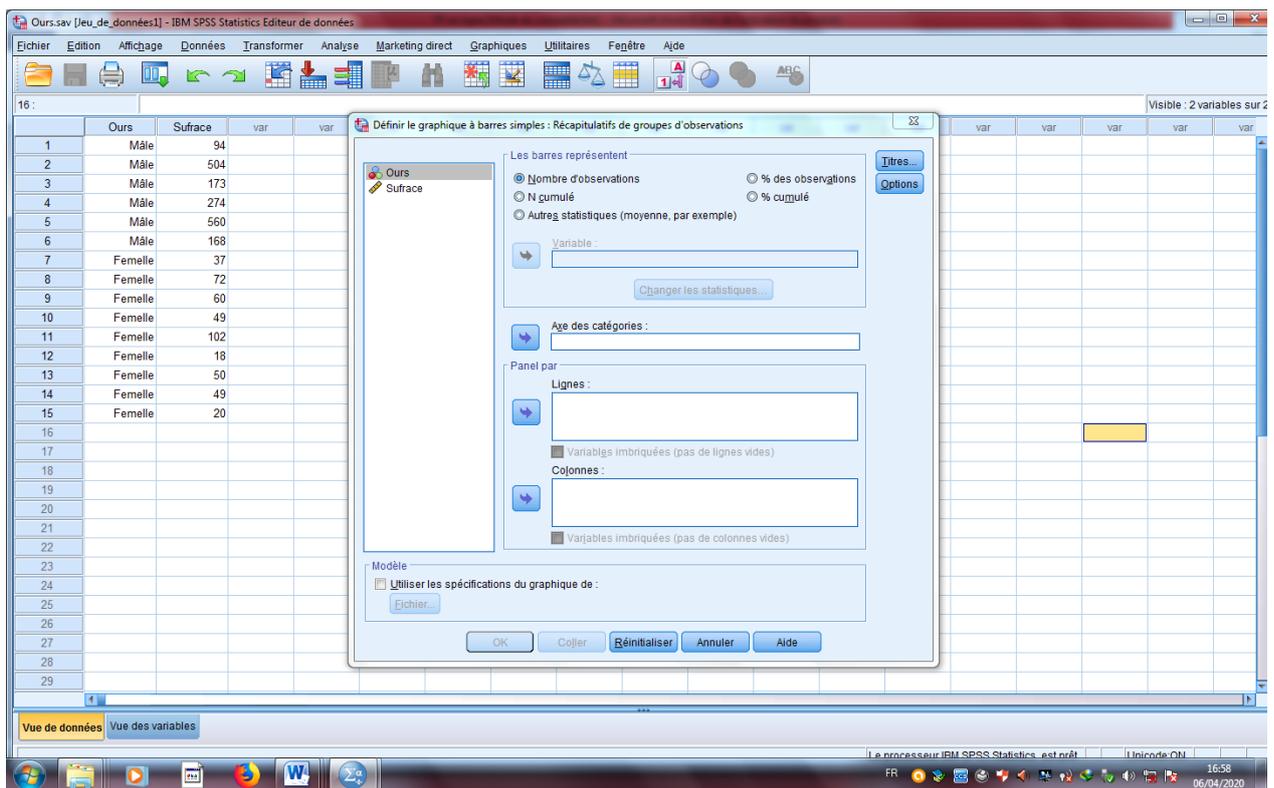
Le mâle occupe une surface six fois plus larges que la femelle (295,5 Km² par rapport 50,78 Km²).

6. Représentation graphique des résultats

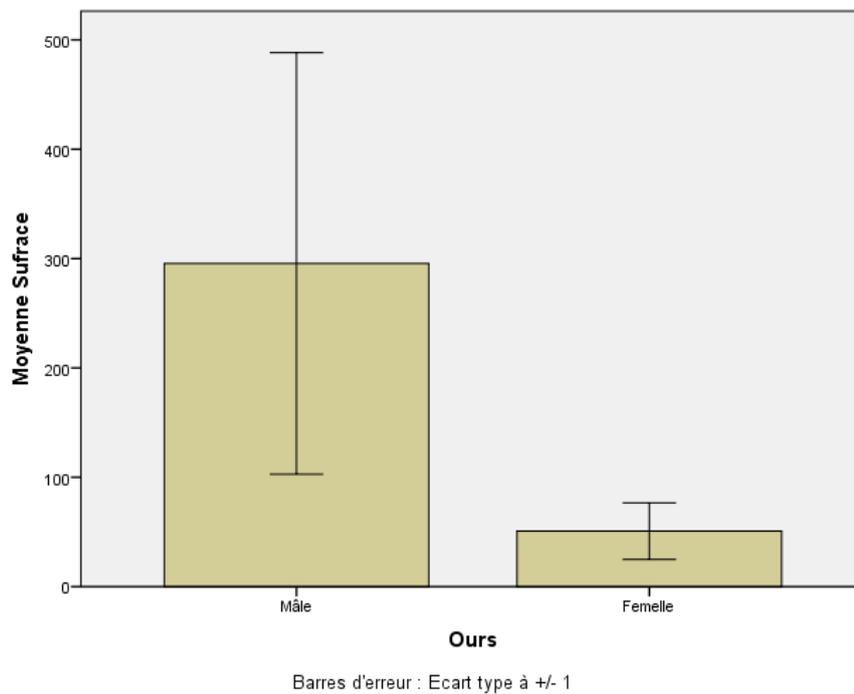
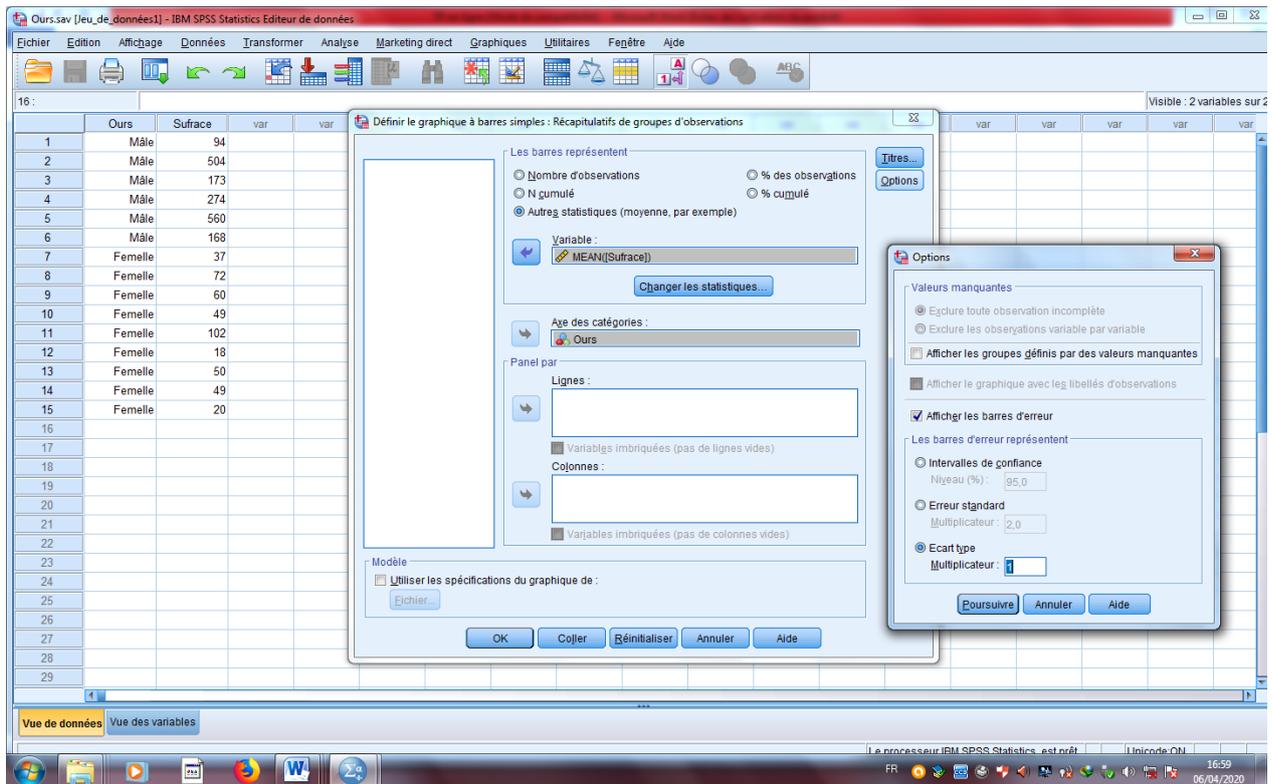
En réalité, on peut exploiter les résultats sous plusieurs formes, mais la meilleure est l'histogramme avec barres.

Allez vers Graphiques et suivre les étapes comme représentent les captures suivantes :





Mettre le multiplicateur **1** au lieu de **2** pour obtenir la taille réelle de la barre écart-type



Titre de la figure : Surfaces occupées (Km²) (Moyenne ± Ecart-type) (Mean ± SD) par le mâle et la femelle de l'ours noir.