1-Statistiques descriptives à une variable

1 Statistique, vocabulaire, tableaux et graphiques

1.1 Définitions, vocabulaire :

Définition

La statistique a longtemps consisté en de simples dénombrements fournissant des renseignements sur la population ou l'économie d'un pays. Si nous ouvrons un dictionnaire, nous trouvons la définition suivante : « La statistique est la science qui a pour objet l'étude numérique et graphique d'un très grand nombre de faits analogues quelle que soit leur nature ».
Cette science n'étudie pas les individus dans leur spécificité, elle permet de les rassembler dans ce qu'ils ont en commun. Les sondages sont en général anonymes et les conclusions portent sur le groupe.
L'objet de la statistique est de **rassembler, organiser, analyser, interpréter,** des observations que l'on peut mesurer ou classer.

1.1.1 Population :

Définition

Les observations que le statisticien est conduit à faire portent sur un ensemble qu'il convient de définir avec une grande précision. Cet ensemble porte le nom de **population** et chaque élément qui la constitue est un **individu** ou une **unité statistique.** Les ensembles et objets de la statistique doivent être parfaitement connus et identifiés. Cela implique une précision de temps et de définition.

Exemple

- Population des élèves de seconde année de STS biotechnologiques pour l'année 97-98 sur la France métropolitaine inscrits dans un lycée public ou en contrat avec l'état. Ces précisions permettent de cerner très exactement la population. Il n'est pas toujours simple de définir celle-ci avec précision, mais cela est nécessaire.

- Population des pièces usinées par la machine A de la chaîne1 d'un processus de fabrication pendant le mois de septembre 1998. Ici la population n'est pas vivante bien que le vocabulaire reste très humanisé. La pièce usinée est toujours l'individu que l'on étudie. Il conviendrait mieux ici de parler d'unité statistique.

1.1.2 Caractère :

Définition

On étudie certaines propriétés des unités statistiques de la population. Chacune de ces propriétés s'appelle un **caractère statistique.** On parle de caractère **qualitatif** lorsque celui-ci n'est pas mesurable (exemples : couleur des cheveux, profession, qualité...etc). Ce caractère qualitatif est dit **ordinal** lorsque l'on peut faire intervenir une notion d'ordre (exemple : les grades de l'armée), sinon le caractère qualitatif est dit **nominal**. On peut affecter un nombre à chaque attribut, cependant toute opération arithmétique doit être maniée avec précaution et exclue s'il s'agit de caractère qualitatif nominal.
On parle au contraire de caractère **quantitatif** lorsque celui-ci est mesurable (exemples : poids, taille, degré d'alcool dans le sang...etc).
Un caractère statistique est aussi appelé **variable statistique.**
Nous dirons qu'une variable statistique quantitative est **discrète** si elle ne peut prendre qu'un nombre dénombrable de valeurs numériques; en revanche, nous dirons qu'elle est **continue** si elle peut prendre toute valeur numérique appartenant à un intervalle réel.

Exemple

- « le nombre d'enfants d'une famille » est un caractère discret fini, il ne peut prendre qu'un nombre fini de valeurs

- « le poids d'un paquet de sucre » est un caractère continu car tous les réels de l'intervalle peuvent être atteints.

Remarque

Dans le cas des mesures, on effectue des observations discontinues, en raison des arrondis sur les données imposés par la manipulation alors qu'en réalité le caractère est continu.

1.2 Collecte de l'information :

Définition

une fois la population parfaitement définie et le caractère étudié choisi, on collecte les observations et on constitue ainsi une **série statistique.** Cette série est **exhaustive** si tous les éléments de la population ont été observés : on parle alors de **recensement**. Lorsque l'étude exhaustive de la population se révèle trop onéreuse ou trop longue à obtenir on observe seulement une partie de la population à l'aide d'un **échantillon**. C'est quasiment toujours le cas. La plupart du temps l'**enquête** statistique utilise un **questionnaire** qui doit être élaboré avec le plus grand soin afin de recueillir les renseignements que l'on souhaite. Il faut qu'il soit non ambigu et pas trop compliqué. On peut également recourir à des documents existants : les registres, les documents de comptabilité ...etc. Il faut ensuite **dépouiller** toutes ces données et procéder à un rangement (stockage) de toutes ces informations afin de pouvoir les exploiter.

1.3 Tableaux statistiques : trois représentations proposées.

Les observations sont le plus souvent nombreuses et se présentent sous forme désordonnée (liste de nombres, tableaux de valeurs...etc). Il faut alors les dépouiller, les ordonner, les classer pour en donner une représentation claire.

1.3.1 Le tableau exhaustif :

Exemple

On a relevé les températures des mois de décembre, janvier et février à Nancy sous abri à 3 heures et obtenu le tableau suivant :

5	**8**	**6**	**7**	**8**	**2**	**–1**	**–2**	**–7**	**–10**
2	**6**	**5**	**12**	**12**	**13**	**10**	**8**	**5**	**6**
4	**8**	**9**	**2**	**–1**	**–2**	**–1**	**–3**	**–2**	**–4**
0	**2**	**–5**	**–2**	**–1**	**–4**	**–2**	**2**	**3**	**8**
9	**5**	**8**	**3**	**5**	**4**	**3**	**2**	**–1**	**–2**
–2	**–5**	**–8**	**–12**	**–16**	**–4**	**–2**	**2**	**0**	**4**
–1	**–2**	**5**	**6**	**4**	**5**	**6**	**2**	**5**	**4**
–2	**–1**	**–5**	**–8**	**–15**	**–16**	**–13**	**–12**	**–5**	**–2**
0	**2**	**6**	**5**	**4**	**6**	**3**	**3**	**2**	**5**

Population : les 90 jours (31 en décembre, 31 en janvier et 28 en février)
Unité statistique *:* un jour (le 8 janvier par exemple)
Variable statistique *:* température en degré Celsius relevée à 3 heures et à un endroit donné.

Ce tableau est inexploitable sous cette forme. On peut juste dire qu'il ne fait pas chaud à Nancy en hiver. (mais ça, on le savait)

1.3.2 Regroupement de données :

Définition

Lorsque les données sont très nombreuses, on peut les regrouper de la manière suivante :
Désignons par X la variable statistique et par $x_1, x_2, \cdots, x_n$ les n valeurs possibles distinctes prises par la variable statistique X (en général si cela est possible, les valeurs x_i sont rangées par ordre croissant,). Nous notons n_i le nombre de fois où la valeur x_i a été observée dans la population (ou dans l'échantillon étudié). Ce nombre n_i est **l'effectif** associé à la valeur x_i de la variable statistique X. L'ensemble des couples (x_i, n_i) est appelé **série statistique**. Il peut évidemment s'agir ici d'une série statistique qualitative ou quantitative.
En désignant par N le nombre total d'observations, nous avons la relation

$$N = \sum_{i=1}^{n} n_i$$

sur l'exemple précédent on obtient

tempé. x_i	**–16**	**–15**	**–14**	**–13**	**–12**	**–11**	**–10**	**–9**	**–8**	**–7**	**–6**	**–5**	**–4**	**–3**	**–2**
effectif n_i	**2**	**1**	**0**	**1**	**2**	**0**	**1**	**0**	**2**	**1**	**0**	**4**	**3**	**1**	**11**
tempé. x_i	**–1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
effectif n_i	**7**	**3**	**0**	**10**	**5**	**6**	**10**	**7**	**1**	**6**	**2**	**1**	**0**	**2**	**1**

Remarque

Aucune information quantitative n'est perdue, seuls les jours où telle température a été relevée ne sont plus connus. Il faudra veiller à ce que cette perte ne soit pas préjudiciable à l'exploitation que l'on veut faire de cette étude. Le tableau est un peu plus lisible que le précédent. On peut par exemple noter les températures les plus souvent atteintes lors de cette période. On peut déjà avoir une idée de la moyenne.

1.3.3 Regroupement par classes :

Définition

Le nombre de valeurs est encore élevé et la lecture du tableau peu commode. On peut encore simplifier la restitution des données. Il suffit de créer des **classes** et de compter l'effectif de chaque classe. On partage alors l'**étendue** (plus grande valeur – plus petite valeur, ici $13-(-16)=29$ des valeurs en p intervalles.

Exemple

Classe	$[-16;-13[$	$[-13;-10[$	$[-10;-7[$	$[-7;-4[$	$[-4;-1[$
effectif	**3**	**3**	**3**	**5**	**15**

Classe	$[-1;2[$	$[2;5[$	$[5;8[$	$[8;11[$	$[11;14[$
effectif	**10**	**21**	**18**	**9**	**3**

Cette troisième représentation sera obligatoirement choisie s'il s'agit d'une variable continue. Les p classes sont alors disjointes et leur réunion recouvre la totalité des valeurs possibles. On dit que l'on fabrique une partition de l'ensemble. On ouvre classiquement l'intervalle à droite et on le ferme à gauche comme dans l'exemple suivant :

Classe	[0 ; 4 [	[4 ; 8 [	[8 ; 12 [	[12 ; 16 [	[16 ; 20 [
effectif	**5**	**10**	**5**	**3**	**2**

Définition

Les classes n'ont pas forcément la même **amplitude** (différence entre la borne supérieure et la borne inférieure). La perte d'information est évidemment le plus gros problème que pose cette technique de stockage. Le choix de l'amplitude permet un compromis satisfaisant au regard des conclusions que l'on veut tirer. On fait ensuite la supposition que chaque élément de la classe possède la valeur du milieu de classe appelé aussi **centre de classe**. Il est parfois difficile de préciser les classes extrêmes. On utilise souvent des classes ouvertes « Plus de » ou « moins de » qui ne possèdent pas de centre de classe. En l'absence d'informations complémentaires, on prendra alors comme centre un nombre situé à une demi amplitude de la borne de cette classe ouverte (l'amplitude choisie étant celle de la classe voisine)
exemple :

Classe	[0 ; 4 [	[4 ; 8 [	[8 ; 12 [	[12 ; 16 [	16 et plus
effectif	**5**	**10**	**5**	**3**	**2**

Pour la dernière classe, l'amplitude de la classe voisine étant 4, si on ne possède pas d'autres informations, on prendra comme centre de classe 16+2=18. On considère en fait que la dernière classe a une amplitude de 4.

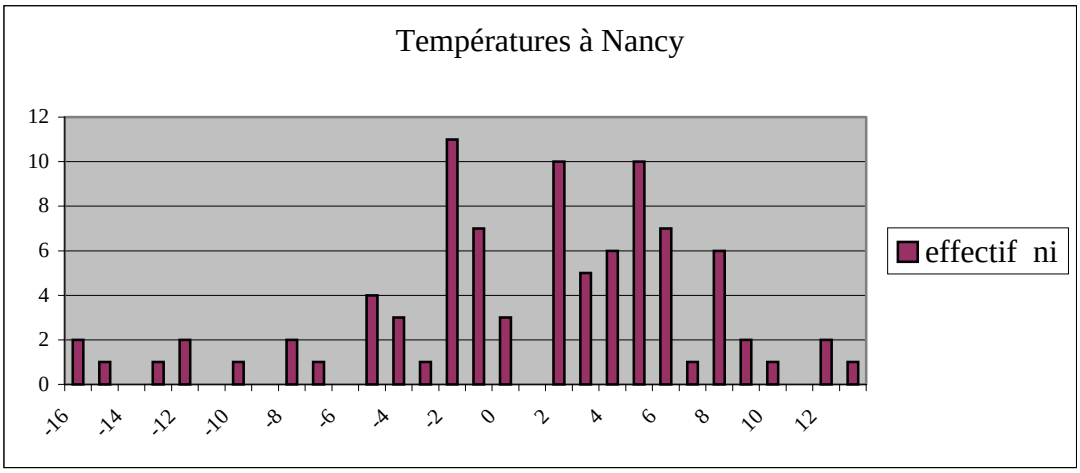
1.4 Graphiques divers :

1.4.1 Diagramme en bâtons

Définition

Lorsque les distributions sont quantitatives, et la variable discrète, le graphique est réalisé en général avec en abscisse les valeurs du paramètre observé et en ordonnée l'effectif ou la fréquence.

La représentation ainsi obtenue est appelée **diagramme en bâtons.** L'effectif ou la fréquence est illustrée par un segment de droite. (On peut également avoir cette représentation pour une variable qualitative). Reprenons les températures de l'exemple précédent. On obtient le graphique suivant :



Lorsque l'on rejoint par des segments de droite les sommets des bâtonnets, on obtient **le polygone des effectifs**.

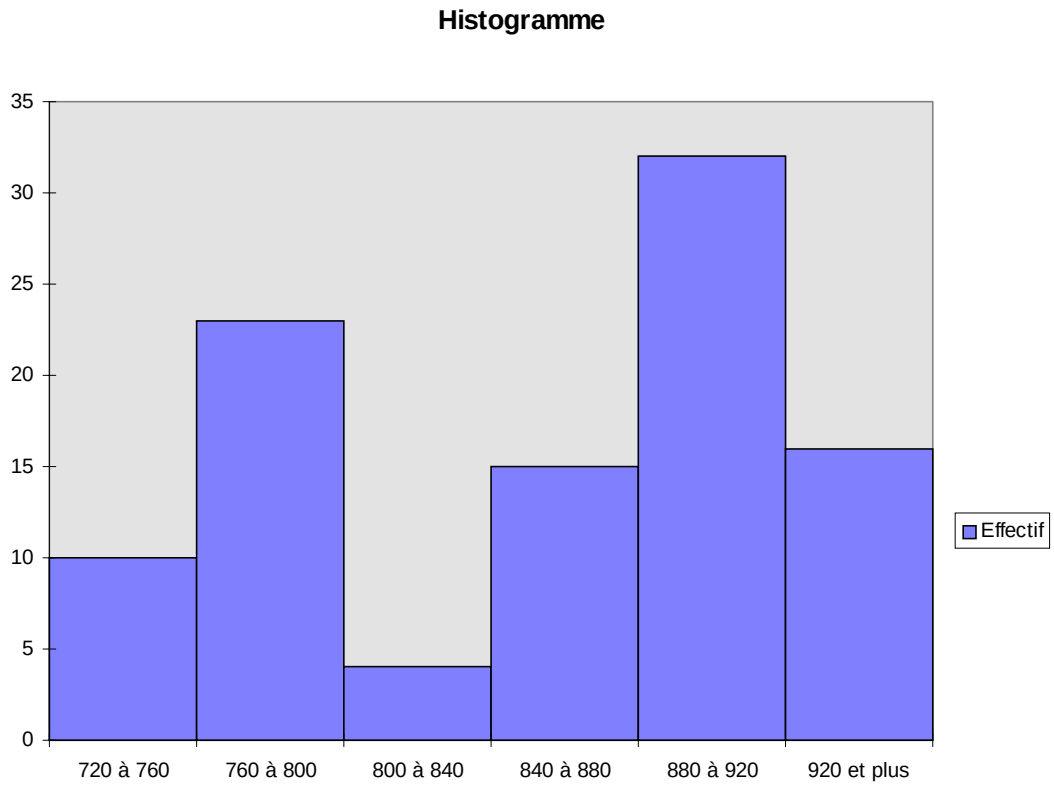
1.4.2 Histogramme

Définition

Dans le cas de la variable continue, le graphique est appelé histogramme. On suppose la répartition uniforme dans une classe et on constitue les rectangles ayant pour base l'amplitude de la classe et une hauteur telle que leur aire soit proportionnelle à l'effectif ou la fréquence de la classe.

Considérons la série statistique suivante qui décrit la charge de rupture d'un fil :

Charge en gramme	**Effectif**
[720 ; 760[	**10**
[760 ; 800[	**23**
[800 ; 840[	**4**
[840 ; 880[	**15**
[880 ; 920[	**32**
920 et plus	**16**



Si les classes ont la même amplitude, on peut retrouver le polygone des effectifs en prenant comme valeur pour chaque élément de la classe le centre.
Si les classes n'ont pas la même amplitude il faut recalculer la hauteur du rectangle.

Par exemple, pour un même effectif dans une classe d'amplitude double, la hauteur du rectangle sera deux fois plus petite.

1.4.3 Diagrammes à bandes, à secteurs, figuratifs

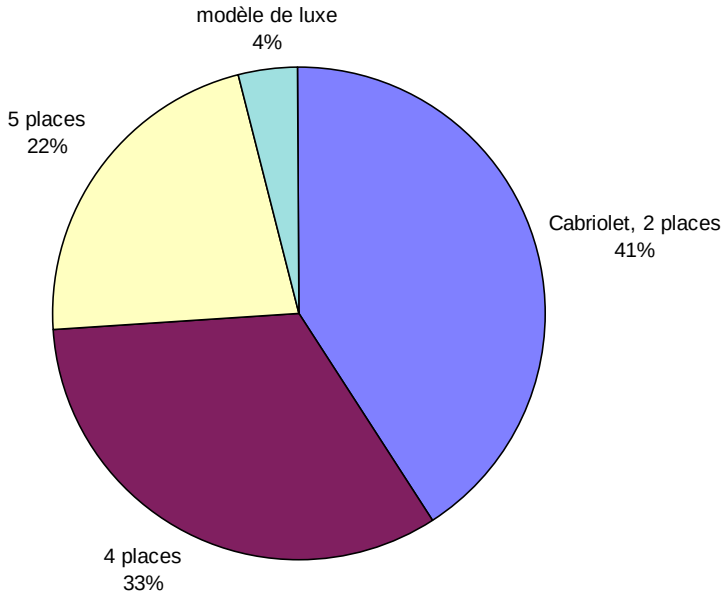
On utilisera ces différentes représentations lorsque le caractère est qualitatif.

Exemple

On considère la production d'une entreprise de fabrication d'automobiles (en milliers de véhicules)

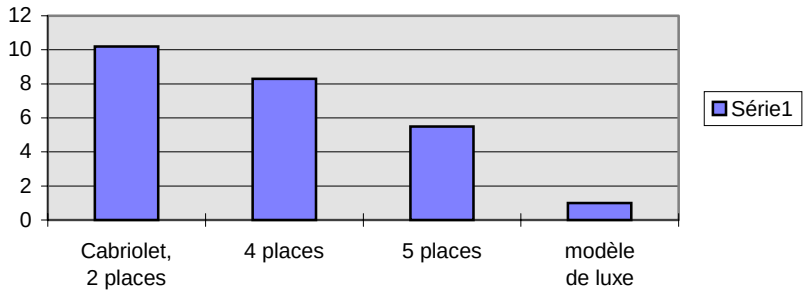
	1970		**1978**	
Véhicule	**Effectif**	**Pourcentage**	**Effectif**	**Pourcentage**
Cabriolet, 2 places	10,2	40,8	25,8	25,8
4 places	8,3	33,2	35,4	35,4
5 places	5,5	22	19,6	19,6
modèle de luxe	1,0	4	16,2	16,2
TOTAL	**25**		**100**	

Diagramme à secteurs



On fera en sorte systématiquement que le total des pourcentages soit 100. Il est parfois nécessaire de « corriger » les arrondis. On utilise pour cela la règle des moindres erreurs.

Diagramme à bandes année 1970



Bandes comparatives.