**Introduction à la statistique**

**Statistiques descriptives**

Les statistiques descriptives nous permettent de donner un sens à l’information recueillie. Elles rendent plus intelligible une série d’observations en permettant de dégager les caractéristiques essentielles qui se dissimulent dans une masse de données. Une **statistique** est un **nombre** résultant de la manipulation de données tirées d’un échantillon. Ce nombre varie d’un échantillon à l’autre et représente un **paramètre**.

1. **Terminologie :**

**La population**est un ensemble d’individus bien défini (**qui**, **où**, **quand ?** Exemple : la population du pourtour méditerranéen en 2015). Souvent, l’étude ne peut pas être faite sur la population totale, nous utilisons donc un **échantillon** ; mais pour faire de l’inférence statistique il faut **un échantillon représentatif**, donc sélectionné au hasard (Exemple : Un échantillon de 10000 habitants choisi au hasard dans cette région pendant cette période).

Nous définissons également **l’unité statistique**(souvent l’individu) et la **distribution statistique**qui correspond à l’ensemble des couples (Xi, Ni) où Xi est une modalité de la variable X et Ni lenombre de fois où cette modalité est observée (c’est-à-dire l’**effectif** ou la **fréquence absolue**).

Exemple : Modalité (Xi) : groupe sanguin (X1 = groupe O Rhésus négatif; X2 = autre ; ici i varie de 1 à 2) ; effectif : N1 = nombre d’individu appartenant au groupe sanguin O Rh-) ; N2= nombre d’individus appartenant à d’autres groupes sanguin). N = N1 + N2 = Ni. N représente le nombre d’individu de l’échantillon étudié.

Cettedistribution peut être également l’ensemble des couples (Xi, Fi) où Fi est la **fréquence relative** de lamodalité Xi. Dans ce cas Fi = Ni/N.

Exemple : parmi 10000 individus choisi au hasard au sein de la population méditerranéenne, 600 possèdent un groupe sanguin O Rhésus négatif.

N = 10000 individus ou **effectif total** sur lequel a porté l’observation ou l’étude (échantillon)

N1 = 600 individus ou effectif portant O Rh – (fréquence absolue) ; F1 (fréquence relative)= (N1/N) x 100 = (600/10000) x 100= 6%

N2 = 10000-600 = 9400 individus possédant des groupes sanguins autre que O Rh– (fréquence absolue) ; F2 (fréquence relative) = 100-6 = 94%, ou bien F2 = (9400/10000) x 100 = 94%

**Les variables :**

1. **Variables qualitatives nominales :**

Ce sont des variables binaires, sans relation d’ordre.

Exemples : Sexe

Répartition de la fréquence des hommes et des femmes dans un groupe donné.

1. **Variables qualitatives ordinales :**

Nous ne pouvons pas faire de moyenne.

Exemples. – Intensité d’une douleur, stade d’une maladie.

1. **Variables quantitatives :**

Elles peuvent êtres discrètes (exemple : nombre d’enfants dans une famille) ou continues (exemples : taille, masse). Pour les variables continues, on peut faire de la discrétisation c’est-à-dire l’arrondi (exemples : taille en cm par exemple ou âge en année).

1. **Série :**

**Exemple de séries statistique** : taille de 5 personnes pris au hasard au sein d’un groupe :

1,80 m ; 1,75 m ; 1,78 m ; 1,82 m et 1,70 m

Cette série représente les données des mesures recueillies à partir d’observations : ce sont les valeurs du **caractère étudié** qu’on appelle aussi **modalité.**

**Remarque :** On calcule une statistique à partir d’un échantillon pour faire une estimation du paramètre de la population.

**B) Procédure de dépouillement des observations et distribution des effectifs :**

A partir d’un ensemble d’observations associé à un caractère, on veut être en mesure d’effectuer une présentation succincte et intelligible. Il faut donc :

* Les dépouiller suivant une distribution d’effectifs ;
* Visualiser ce dépouillement à l’aide d’un histogramme ou d’un diagramme en bâtons.

**Dépouillement des observations :**

Le dépouillement consiste à grouper les observations en classe (dans le cas où le caractère est continu), en modalités dans le cas où il est discontinu ou en qualité dans le cas où il est qualitatif, en indiquant par un trait vertical chaque observation appartenant à sa classe (ou égale à sa modalité ou identique à sa qualité) respective. On dépouille les observations par bloc de 5 en marquant d’un trait horizontal un ensemble de quatre traits verticaux déjà notés.

**Distribution d’effectifs :**

La somme du nombre de traits appartenant à chaque classe (modalité ou qualité) donne l’effectif de cette classe (modalité ou qualité). La répartition des effectifs dans les classes (modalité ou qualité) s’appelle la distribution d’effectifs.

Exemple : Soit la série statistique suivante :

# 83,5 -102 - 104,5 - 56 – 66 – 68 – 93 - 69,5 – 71 – 73 – 106 -74 -79,5 – 80 84 - 91- 92 - 84,5 - 86 – 87 -76 -77 - 89 - 90,5 - 92,5 -94  - 108 -57,5 - 63 78 -100 - 78,5 - 65 -111 - 83 .

|  |  |  |
| --- | --- | --- |
| Classes | Dépouillement | Effectifs |
| 55 < x < 65 |  |  |
| 65 < x <75 |  |  |
| 75 < x < 85 |  |  |
| 85 < x < 95 |  |  |
| 95 < x < 105 |  |  |
| 105< x< 115 |  |  |
| Total |  |  |

Limites des classes :

Ce sont les nombres entre lesquels sont classées les observations

Exemple : 55 est la limite inférieur de la première classe ;

65 est la limite supérieure de la première classe

Amplitude des classes :

C’est l’écart qui existe entre les limites de la classe.

Exemple : l’amplitude de la première classe est 65-55 = 10

Centre des classes :

C’est la somme des limites de chaque classe divisée par 2

Exemple : le centre de la première classe (point milieu) est (65+55)/2 = 60

Etendu de la série : 111-56 = 55

Nombre de classe selon la règle de Sturges : 6

**Etapes à suivre :**

1. **Ranger les observations par valeurs croissantes**
2. **Déterminer le nombre de classes :**

Le nombre de classe ne doit pas être inférieur à 5 ni supérieur à 20. On peut utiliser la règle de **Sturges**: nombre de classes = 1+3,3 x Log(n) ou la règle de **Yule**:

2,5 x racine quatrième de (n) ou utiliser le tableau suivant (déduit de la règle de Sturges)

Nombre d’observations à dépouiller (n) nombre de classes (k)

|  |  |
| --- | --- |
| 10 | 4 |
| 10 < n < 22 | 5 |
| 22 < n < 44 | 6 |
| 44 < n < 90 | 7 |
| 90 < n < 180 | 8 |
| 180 < n < 360 | 9 |
| 360 < n < 720 | 10 |
| 720 < n < 1000 | 11 |

1. **Déterminer l’amplitude de chaque classe :**

On calcule l’étendue E de la série de données (différence entre la plus grande valeur et la plus petite valeur) et on divise l’étendue par le nombre de classes souhaitées :

Amplitude = E/k

1. **Fixer les limites des classes :**

La limite inférieure de la première classe peut être la valeur minimale de la série ou une plus petite mais voisine de la valeur minimale.

1. **Dépouiller les observations :**

Cette opération concerne les données en vrac (Non ordonnées)

Lorsque le caractère étudié ne prend que des valeurs entières, le dépouillement ne s’effectue pas par classe.

1. **Distribution des effectifs avec classes ouvertes :**

A cause de l’étalement des observations, il se peut qu’on ait recours à une distribution d’effectif avec classes ouvertes.

Exemple :

|  |  |  |  |
| --- | --- | --- | --- |
| Revenus en DH | Effectifs | Fréquences relatives | Fréquences cumulées |
| Moins de 15000 | 2 | (2/40) x 100 = 5% | 5 |
| 15000 à moins de 20000 | 5 | (5/40) x 100 = 12,5% | 5+12,5 = 17,5 |
| 20000 à moins de 25000 | 9 | (9/40) x 100 = 22,5% | 17,5 + 22,5 = 40 |
| 25000 à moins de 30000 | 14 | (14/40) x 100 = 35% | 40 + 35 = 75 |
| 30000 à moins de 35000 | 7 | (7/40) x 100 = 17,5% | 75 + 17,5 = 92,5 |
| 35000 et plus | 3 | (3/40) x 100 = 7,5% | 92,5 + 7,5 = 100 |
| Total | 40 | 100 % |  |

**C) Présentations graphiques d’une distribution d’effectifs :**

**1) Diagramme en bâtons :**

Le diagramme en bâtons est utilisé lorsque le caractère étudié est quantitatif discontinu ou lorsqu’il s’agit d’un caractère qualitatif. Il est constitué en portant en abscisse les valeurs (ou les qualités) des caractères en traçant parallèlement à l’axe des ordonnées un trait de longueur proportionnelle à l’effectif de chaque modalité du caractère.

Exemple d’un tracé :

|  |  |
| --- | --- |
| Note sur 20 | Nombre d’étudiants ayant la note |
| 14 | 6 |
| 15 | 14 |
| 15,5 | 8 |
| 17 | 2 |
| TOTAL | 30 |

**2) Histogramme et polygone d’effectifs :**

Ces deux types de graphes sont utilisés lorsque le caractère quantitatif est continu. Dans ce cas les valeurs observées sont généralement dénombrées suivant une distribution en classes.

* **Histogramme :**

L’histogramme est constitué de rectangles juxtaposés dont chacune des bases est égales à l’intervalle de classe et dont la hauteur est telle que la surface soit proportionnelle à la fréquence (ou l’effectif) de la classe correspondante. Il faut laisser un espace entre le premier rectangle de l’histogramme et l’axe des ordonnées.

* **Polygone des effectifs :**

Le polygone des effectifs permet de représenter l’allure générale de la distribution. Il est obtenu en joignant les points ayant pour coordonnées les centres des classes et les effectifs correspondants. Il faut ajouter de part et d’autre de la courbe une classe d’effectif nul.

Exemple (voir cours)

**Remarque :** Les valeurs regroupées dans la même classe se verront attribuer la même valeur soit celle du point milieu de la classe. Si on veut comparer plusieurs histogrammes constitués à partir d’échantillons de tailles différentes, alors on utilise en ordonnée les fréquences relatives.

* **Rectification d’effectifs :**

Exemple (voir le cours)

**3) Courbe cumulative :**

La courbe cumulative d’une série continue s’obtient à partir des effectifs cumulés ou des fréquences relatives cumulées. Pour se faire, on détermine une succession de points dont les abscisses correspondent aux limites supérieures des classes et dont les ordonnées sont égales aux effectifs cumulés correspondant aux classes.

Exemple (voir cours)