Cours du 22/01/14

Statistiques

Les statistiques c'est objectiver un fonctionnement en quantifiant ses observations

Chapitre I : Séries Statistiques - Généralités

I. Définitions

Population : Ensemble d'éléments faisant l'objet d'une étude statistique

Ex : Ensemble des élèves de l'EFREI (étude sur l'âge, poids, notes …), ensemble des pièces produites par une machine (étude sur la taille, la qualité …)

Echantillon : Partie de la population

Ex : Groupe A des L1, pièces fabriquées un jour donné

Unité statistique : Un élément de la population ou de l'échantillon dont on veut étudier 1 ou plusieurs caractères

Caractère : Qualitatif : Non mesurable

Quantitatif : Mesurable avec une variable statistique

Variable statistique

- Discrète : Valeurs isolées (nombres d'individus etc …)
- Continue : Elle peut prendre toutes les valeurs d'un intervalle

Série statistique : L'ensemble des valeurs prises par une variable statistique sur l'ensemble de la population ou sur un échantillon

II. Séries statistiques d'un caractère quantitatif discret

On considère un échantillon de n éléments. Soit x_1, x_2, …, x_p les valeurs possibles du caractère x mesurée ici.

Série statistique : Les n valeurs prisent par les n éléments

Effectif total : n

Fréquence absolue ou Effectif potentiel de x : C'est le nombre n_i de fois qu'apparaît x_i

$n_1 + n_2 + … + n_p = n$

Fréquence relative de x_i : $f_i = \frac{ni}{n}$

Etendue de la série : Ecart entre le plus grand x_i et le plus petit

Ex :

On considère 100 familles de 4 enfants. On va étudier le nombre de garçons.

Série statistique : 3, 1, 0, 2, 4, …, 2

Effectif total : 100

Etendue : 4

Valeurs de x : $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$

On va compter les valeurs et les regrouper

x_i	0	1	2	3	4
n_i	7	20	43	25	5

Fréquence relative des familles avec 2 garçons : $f_2 = \frac{n2}{n} = \frac{43}{100} = 43\%$

III. Série statistique à caractère quantitatif continue

Le nombre p de valeurs possibles de x_i est infini.

On constitue des classes (intervalles) en divisant l'étendue de la série en un certain nombre de classes.

Par convention, pour chaque classe : (x_{min} inclus) [[(x_{max} exclu)

Centre des classes : $x_1, x_2, \ldots, x_p$

Effectifs des classes : $n_1, n_2, \ldots, n_p$

Ex :

Pesées à 10g près, donnent un poids entre 2,240 et 4,490 kg => classes d'amplitude 0,3 kg

	2,2	2,5	2,8	3,1	3,4	3,7	4,0	4,3	4,6
Centre	2,35	2,65	2,95	3,25	3,55	3,85	4,15	4,45	
Effectif	5	11	24	40	42	20	13	6	
Fréquence %	3,1	6,8	14,9	24,8	26,1	12,4	8,1	3,7	

IV. Séries statistiques d'un caractère qualitatif

On va regrouper les résultats en autant de classes qu'il existe de modalités du caractère

Ex : étude de la couleur de fleurs, avec 3 couleurs possibles : Rouge, Bleu, Vert.

On constitue 3 classes, une par couleur, avec un effectif par classe.

Variables ordinales : Si on peut ordonner les valeurs

Variables nominales : Si les valeurs sont non-ordonnables

V. Représentation graphique des séries statistiques

1) Cas discret

On va travailler avec des diagrammes en bâtons où on va mettre :

- n_i en fonction de x_i
- f_i en fonction de x_i

Ex :

x_i	0	1	2	3	4
n_i	7	20	43	25	5

Graphique

- Polygône des effectifs et des fréquences relatives

On trace une courbe qui va rejoindre l'extrémité des bâtons

- Diagramme cumulatif
  - Effectif cumulé jusqu'à la i-ème valeur x_i du caractère
    $n_1 + n_2 + \dots + n_i$
  - Fréquence cumulée
    $f_1 + f_2 + \dots + f_i$

⇨ Diagramme cumulé des effectifs :

x_i	0	1	2	3	4
n_i	7	20	43	25	5
$\sum n_i$	7	27	70	95	100

Graphique

Cours du 24/01/14

2) Caractère continu

Diagramme en bâton impossible car on a trop de x_i. On réalise donc un histogramme.

- Avec des classes égales (de même amplitude) :

Graphique

- Avec des classes inégales :

Classes	5 - 6	6 - 7	7 - 8	8 – 10
Effectifs	12	13	16	6

3 premières classes : amplitude de 1
Dernière classe : amplitude de 2

On va corriger l'effectif de la dernière classe : on va la diviser par 2 : $n_i = \frac{6}{2} = 3$

On a donc deux classes d'effectif 3 : 8 – 9 et 9 – 10

D'où l'histogramme suivant :

Graphique

- Polygône des effectifs et des fréquences :

Il s'agit d'un ligne brisée (segment de droites) joignant les milieux des sommets des rectangles de l'histogramme :

Graphique

- Polygône des effectifs (ou fréquence) cumulé(e) :

Attention, ne pas confondre avec le cas discret, ici, c'est une droite.

Exemple des nourissons :

Graphique

Chapitre II : Paramètres de position et de dispersion

Objectif : On veut condenser l'information contenue dans une série statistique

Paramètres de position : Donner l'ordre de grandeur des mesures et l'existence de valeurs centrales autout desquelles se groupent les mesures (moyenne …)

Paramètres de dispersion : Donne une estimation de la dispersion des mesures autour d'un paramètre de position (écart-type, variance …)

I. Paramètres de position

On a une série statistique avec n valeurs : $x_1, x_2, x_3, \ldots, x_n$

1) Moyenne arithmétique

$$\bar{x} = \frac{x1 + x2 + x3 + \cdots + xn}{n} = \frac{1}{n}\sum_{i=1}^{n} xi$$

Série à caractère discret : On va avoir p valeurs de x_i distinctes : $x_1, x_2, \ldots, x_p$ et n_i effectifs associés à chaque x_i ainsi que f_i, la fréquence relative associée à chaque x_i

Ainsi, $\bar{x} = \frac{n1x1+n2x2+\cdots+npxp}{n} = \frac{1}{n}\sum_{i=1}^{p} nixi = \sum_{i=1}^{p} fixi$

Ex : Famille de 4 enfants, x_i = nombres de garçons

x_i	0	1	2	3	4
f_i	7%	20%	43%	25%	5%

$\bar{x} = \sum_{i=0}^{4} fixi = \frac{7}{100} * 0 + \frac{20}{100} * 1 + \frac{43}{100} * 2 + \frac{25}{100} * 3 + \frac{5}{100} * 4 = \frac{201}{100} = 2{,}01$ garçons (pour cette série)

Remarque : On parle de moyenne arithmétique pondérée par les fréquences relatives f_i

Série à caractère continu : On a p classes de centre x_i et d'effectif n_i

$$\bar{x} = \frac{n1x1 + n2x2 + \cdots + npxp}{n} = \frac{1}{n}\sum_{i=1}^{p} nixi = \sum_{i=1}^{p} fixi$$

Ex : Les nourissons

Classes (z_i)	1	2	3	4	5	6	7	8
Centre (x_i)	2,35	2,65	2,95	3,25	3,55	3,85	4,15	4,45
Effectif (n_i)	5	11	24	40	42	20	13	6
f_i (%)	3,1	6,8	14,9	24,8	26,1	12,4	8,1	3,7

$$\bar{x} = \sum_{i=1}^{p} fixi = \frac{3,1}{100} * 2,35 + \cdots + \frac{3,7}{100} * 4,45 = 3,406 \text{ kg (pour cette série)}$$

Remarque : Astuce de calcul

On peut remarquer que $x_i = M + a * z_i$

Avec M = valeur initiale (ici, 2,05)

a = amplitude de la classe (ici, 0,3)

x_i = centre de la classe

z_i = numéro de la classe (ici, de 1 à 8)

On a $\bar{x} = \text{M} + \text{a} * \bar{z}$

<u>Démonstration</u> :

On a $\bar{x} = \frac{1}{n} \sum_i nixi = \frac{1}{n} \sum_i \text{ni(M} + \text{a} * \text{zi)} = \frac{1}{n} \sum_i \text{niM} + \frac{1}{n} \sum_i \text{ni} * \text{azi} = \text{M} * \frac{1}{n} \sum_i \text{ni} + \text{a} * \frac{1}{n} \sum_i \text{nizi} = \text{M} * \frac{n}{n} + \text{a} * \bar{z}$

Ex :

M = 2,05

a = 0,3

$$\bar{z} = \frac{1}{n} \sum_i \text{nizi} = \sum_i \text{fizi} = \frac{5 * 1 + 11 * 2 + \cdots + 8 * 6}{161} = \frac{728}{161}$$

$$\bar{x} = 2,05 + 0,3 * \frac{728}{161} = 3,406 \text{ kg}$$

2) Médiane

On va ordonner les valeurs de la série et on va prendre la valeur du « milieu ». Ou, cette valeur coupe la série en deux (autant de valeurs au-dessus qu'en-dessous).

<u>Série à caractère discret</u> :

Ex :

- 3, 3, 4, 5, 7, 7, 9 => Médiane = Me = 5
- 3, 3, 4, 5, 6, 7, 7, 9 => Médiane = Me = 5,5

Série à caractère continue :

La série est déjà ordonnée par classes donc la médiane (Me) va tomber dans une certaine classe notée [l_1 , l_2 [

Graphique

Avec : n = effectif total

F_1 = effectif cumulé en l_1

F_2 = effectif cumulé en l_2

$n_0 = F_2 - F_1$ = effectif de la classe

On va supposer que les n_0 valeurs de la classe [l_1 , l_2 [sont uniformément répartie dans cette classe.

Si on regarde une partie du diagramme des effectifs cumulés :

Graphique

On peut faire une estimation graphique

Graphique

On a $\tan \alpha = \frac{n'}{Me-l1} = \frac{n0}{l2-l1}$

On a aussi $\frac{n}{2} = F1 + n' = F1 + n0 * (\frac{Me-l1}{l2-l1})$

$\frac{n}{2} = F1 + (F2 - F1) * (\frac{Me-l1}{l2-l1})$

$\mathrm{Me} = \mathrm{l1} + \left(\frac{n}{2} - F1\right) * (\frac{l2-l1}{F2-F1})$

Cours du 30/01/14

Ex : Les nourrissons

Classe	[2,2;2,5[	[2,5;2,8[	[2,8;3,1[	[3,1;3,4[	[3,4;3,7[	[3,7;4,0[	[4,0;4,3[	[4,3;4,6[
n_i	5	11	24	40	42	20	13	6
$\sum n_i$	5	16	40	80	122	142	155	161

Médiane : valeurs des x_i qui correspond à $\frac{n}{2}$ = 80,5

Classe de la médiane : [3,4;3,7[

$$\mathrm{Me} = \mathrm{l1} + \left(\frac{n}{2} - F1\right) * (\frac{l2 - l1}{F2 - F1})$$

Ici, l_1 = 3,4 -> F_1 (effectif cumulé en l_1) = 80

l_2 = 3,7 -> F_2 (effectif cumulé en l_2) = 122

Me = 3,4 + (80,5 – 80)*$\frac{0,3}{42}$ = 3,4 + $\frac{1}{2} * \frac{0,3}{42}$ = 3,403 kg

Méthode graphique : A partir du diagramme cumulé, on va regarder la valeur de x_i qui correspond à $\frac{n}{2}$

Graphique

3) Quartiles

Les quartiles coupent la série en 4 groupes du même effectif

Cas discret

➔ 4, 5, 6, 11, 13, 14, 16

Q_1 = 5 (= $\frac{n}{4}$)

Me = Q_2 = 11 (= $\frac{n}{2}$)

Q_3 = 14 (=$\frac{3n}{4}$)

➔ 4, 5, 6, 7, 11, 13, 14, 15, 16

Q_1 = 6 (= $\frac{n}{4}$)

Me = Q_2 = 11 (= $\frac{n}{2}$)

Q_3 = 14 (=$\frac{3n}{4}$)

Cas continu

➔ Méthode graphique

Graphique

➔ Méthode analytique

$$Q_1 = l1 + \left(\frac{n}{4} - F1\right) * \left(\frac{l2-l1}{F2-F1}\right)$$

$$Q_3 = l1 + \left(\frac{3n}{4} - F1\right) * \left(\frac{l2-l1}{F2-F1}\right)$$

4) Mode ou valeur dominante

Mode = Valeur(s) la(les) plus fréquente(s). On peut en avoir plusieurs par série

Cas discret

2, 2, 5, 7, 8, 8, 9, 9, 10, 11, 14, 14

Modes : 2, 8, 9, 14

Cas continu

On va commencer par définir (trouver) la classe modale (classe avec l'effectif le plus important)

Histogramme :

« Zoom » sur la classe modale

Graphique

Mode : $M_0 = l_1 + \frac{\Delta 1}{\Delta 1 + \Delta 2} * (l_2 - l_1)$

Dans cette formule : [l1;l2[est la classe modale

Δ_1 est l'excédent entre la classe modale et la classe inférieure

Δ_2 est l'excédent entre la classe modale et la classe supérieure

On a deux segments qui ont la même pente. Ainsi :

$$\frac{\Delta 1}{M0 - l1} = \frac{\Delta 2}{l2 - M0}$$

On veut M_0

$$\Leftrightarrow \Delta 1(l2 - M0) = \Delta 2\,(M0 - l1)$$

$$\Leftrightarrow \Delta 1 l2 - \Delta 1 M0 = \Delta 2 M0 - \Delta 2 l1$$

$$\Leftrightarrow M0(\Delta 1 + \Delta 2) = \Delta 1 l2 + \Delta 2 l1$$

$$\Leftrightarrow M0 = \frac{\Delta 1 l2 + \Delta 2 l1}{\Delta 1 + \Delta 2}$$

$$\Leftrightarrow M0 = \frac{\Delta 1 l2 + \Delta 2 l1 + \Delta 1 l1 - \Delta 1 l1}{\Delta 1 + \Delta 2}$$

$$\Leftrightarrow M0 = \frac{l1(\Delta 1 + \Delta 2) + \Delta 1(l2 - l1)}{\Delta 1 + \Delta 2}$$

$$\Leftrightarrow M0 = l1 + \frac{\Delta 1}{\Delta 1 + \Delta 2} * (l2 - l1)$$

Ex : Les nourrissons

Classe modale : [3,4;3,7[

⇨ **Graphique**

Mode = $M_0 = 3{,}4 + \frac{2}{2+22} * (3{,}7 - 3{,}4) = 3{,}4 + \frac{2}{24} * 0{,}3 = 3{,}425$

Remarque

Séries unimodales : **Graphique**

Séries plurimodales : **Graphique**

II. Paramètres de dispersion

1) Etendue

L'étendue c'est $x_{max} - x_{min}$

2) Ecart moyen

Soit une série statistique, de valeurs x_i, i de 1 à n, de moyenne $\bar{x}$

Ecart de $x_i = e_i = |x_i - \bar{x}|$

Ecart moyen de la série :

$$\bar{e} = \frac{e1 + e2 + e3 + \cdots + en}{n} = \frac{1}{n}\sum_{i=1}^{n} ei$$

Cas discret

Valeurs x_i, i de 1 à p, effectif n_i, fréquence relative $f_i = \frac{ni}{n}$

Cas continu

p clases, x_i centre des classes, effectif n_i, fréquence relative $f_i = \frac{ni}{n}$

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} niei = \sum_{i=1}^{n} fiei$$

Ex : Nombres de garçons dans les familles de 4 enfants

$\bar{x} = 2{,}01 \sim 2$

$$\bar{e} = \frac{7}{100}|0-2| + \frac{20}{100}|1-2| + \frac{43}{100}|2-2| + \frac{25}{100}|3-2| + \frac{5}{100}|4-2|$$
$$= \frac{14 + 20 + 25 + 10}{100} = \frac{29}{100} = 0{,}69$$

3) Variances et écart-type

La variance est la moyenne arithmétique des carrés des écarts des x_i.
$\sqrt{o}_x$ représente un sigma minuscule indice x

$$\left(\sqrt{ox}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(xi - \bar{x})^2$$

L'écart type est la racine carrée de la variance.

Cas discret et continu

En reprenant les mêmes notations que pour l'écart moyen

$$\left(\sqrt{ox}\right)^2 = \frac{1}{n}\sum_{i=1}^{p} ni(xi - \bar{x})^2 = \sum_{i=1}^{p} fi(xi - \bar{x})^2$$

Formule de calcul :

$$\left(\sqrt{ox}\right)^2 = \frac{1}{n}\left(\sum_{i=1}^{p} nixi^2\right) - (\bar{x})^2 = \overline{x^2} - (\bar{x})^2$$

Démonstration

$$(\sqrt{o_x})^2 = \frac{1}{n}\sum_{i=1}^{n}(xi - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(xi^2 - 2xi\bar{x} + (\bar{x})^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n} xi^2 - \frac{2}{n}\sum_{i=1}^{n} 2xi\bar{x} + \frac{1}{n}\sum_{i=1}^{n}(\bar{x})^2$$

$$= \overline{x^2} - 2\bar{x}\frac{1}{n}\sum_{i=1}^{p} xi + (\bar{x})^2\frac{1}{n}\sum_{i=1}^{n} 1$$

$$= \overline{x^2} - 2(\bar{x})^2 + (\bar{x})^2$$

$$= \overline{x^2} - (\bar{x})^2$$

Méthode de calcul

1) On calcule $\bar{x}$
2) On calcule $\overline{x^2}$
3) On utilise la formule simple pour $(\sqrt{o_x})^2$

Ex : Nombres de garçons dans les familles de 4 enfants

$\bar{x} \sim 2$

$$\left(\sqrt{ox}\right)^2 = \frac{1}{n}\left(\sum_{i=1}^{5} nixi^2\right) - (\bar{x})^2$$

$$= \frac{7}{100} * 0^2 + \frac{20}{100} * 1^2 + \frac{43}{100} * 2^2 + \frac{25}{100} * 3^2 + \frac{5}{100} * 4^2 - 4$$

$$= \frac{20 + 172 + 225 + 80}{100} - 4 = 4{,}97 - 4 = 0{,}97$$