IDUP / mars 21

Master 1 : Analyse statistique (J.-F. Léger)

Caractéristiques de dispersion / concentration

Exercice 6

Patients et durée d'hospitalisation

CORRIGÉ

Tableau 1 : distribution des patients d'un service hospitalier selon la durée de séjour

Durée de séjour en jours, en nombre de nuits (xi)	Nombre de malades (ni)	(1)	(2)	(3)	(4)	(5)	(6)
1	21	8,50%	8,50%	0,0%	21	21	0,0%
2	20	8,10%	8,10%	8,5%	40	61	0,4%
3	14	5,67%	5,67%	16,6%	42	103	1,2%
4	9	3,64%	3,64%	22,3%	36	139	2,0%
5	13	5,26%	5,26%	25,9%	65	204	2,7%
6	13	5,26%	5,26%	31,2%	78	282	4,0%
7	7	2,83%	2,83%	36,4%	49	331	5,6%
8	19	7,69%	7,69%	39,3%	152	483	6,5%
9	10	4,05%	4,05%	47,0%	90	573	9,5%
10-19	50	20,24%	2,02%	51,0%	750	1 323	11,3%
20-29	21	8,50%	0,85%	71,3%	525	1 848	26,2%
30-39	10	4,05%	0,40%	79,8%	350	2 198	36,5%
40-49	5	2,02%	0,20%	83,8%	225	2 423	43,5%
50-59	3	1,21%	0,12%	85,8%	165	2 588	47,9%
60-69	12	4,86%	0,49%	87,0%	780	3 368	51,2%
70-79	6	2,43%	0,24%	91,9%	450	3 818	66,6%
80-89	11	4,45%	0,45%	94,3%	935	4 753	75,5%
90-99	1	0,40%	0,04%	98,8%	95	4 848	94,0%
100-109	2	0,81%	0,08%	99,2%	210	5 058	95,8%
110 et plus				100,0%			100,0%
Total	247	100,00%			5 058		

1) Complétez le tableau 1 en nommant les colonnes (1) à (6) et en expliquant :
   - comment on obtient les résultats mentionnés dans chaque colonne ;
   - le sens qu'ont ces données.

Corrigé

Colonne 1 : il s'agit de la répartition en pourcentage des patients selon la durée de séjour. Les proportions (f_i) sont calculées en rapportant les effectifs (n_i) par la population totale (N). Par exemple, 8,5 % des personnes hospitalisées sont restées une nuit.

$$f_i = \frac{n_i}{N}$$

$$8{,}5\ \% = 100 \times \frac{21}{247}$$

Colonne 2 : il s'agit de la densité (d_i) de population par unité de référence de durée de séjour (ici une nuit). On la calcule en divisant la fréquence relative (f_i) par l'amplitude de l'intervalle de durée (a_i). Par exemple, la densité de population entre 10 et 20 nuits d'hospitalisation est de 2,02 %. Cela signifie qu'il y a, en moyenne, 2 % de personnes restées 10 nuits, 2 % de personnes restées 11 nuits, etc. La densité de population est utilisée pour la construction d'un histogramme dans le cas où les amplitudes des différentes modalités de la variable quantitative sont inégales, ce qui est le cas ici.

$$d_i = \frac{f_i}{a}$$

$$2{,}02\ \% = \frac{20{,}24}{10}$$

Colonne 3 : Il s'agit des fréquences relatives cumulées (fc_i). Elle indique la proportion de patients restés moins d'un certain nombre de nuit. Par exemple, compte tenu des conventions adoptées sur le tableau, 51,0 % des patients sont restés moins de 10 jours.

$$fc_i = \sum_{x=1}^{i-1} f_x$$

$$fc_{10} = \sum_{x=1}^{9} f_{x-1} = f_1 + f_2 + f_3 + \ldots + f_9$$

$$fc_{10} = 8{,}50\ \% + 8{,}10\ \% + \ldots + 4{,}05\ \% = 51{,}0\ \%$$

Colonne 4 : Il s'agit du produit du nombre de patients restés un certain nombre de nuits (n_i) par la durée moyenne de séjour de ces patients (x_i). Ce produit (la masse statistique m_i) correspond au nombre de nuits d'hospitalisation cumulées par les patients qui ont en commun d'être restés un même nombre de nuits à l'hôpital. Pour les classes de durées, on multiplie l'effectif correspondant de patients par le centre de la classe de durées. Par exemple, les 50 patients qui ont passé de 10 à 20 nuits à l'hôpital sont restés en moyenne 15 nuits. Au total, à ces 50 patients correspondent 750 nuits d'hospitalisation.

$$m_i = n_i \times x_i$$

$$ou\ m_i = n_i \times \left(\frac{x_i + x_{i+a}}{2} \right)$$

$$m_{10-19} = n_{10-19} \times \frac{x_{10} + x_{10+10}}{2} = 50 \times \left(\frac{10+20}{2} \right) = 50 \times 15 = 750$$

Colonne 5 : Il s'agit du cumul de la masse statistique ou masse statistique cumulée (mc_i). Elle se calcule à partir des masses statistiques (m_i) comme la fréquence cumulée à partir des fréquences relatives (f_i). Par exemple, le cumul des nuits passées à l'hôpital par les patients restés moins de 10 jours s'élève à 573 nuits.

$$mc_i = \sum_{x=1}^{i-1} m_x$$

$$mc_{10} = \sum_{x=1}^{9} m_{x-1} = m_1 + m_2 + m_3 + \ldots + m_9$$

$$mc_{10} = 21 + 40 + 42 + \ldots + 90 = 573$$

Colonne 6 : il s'agit de la masse statistique cumulée relative (mcr_i), c'est-à-dire, dans le cas présent, la proportion de nuits cumulées par les patients hospitalisés moins d'un certain nombre de nuits (i). Par exemple, les patients restés moins de 10 jours ont cumulé 573 nuits d'hospitalisation sur les 5 058 nuits passées par l'ensemble des patients. Cela représente 11,3 % du total des nuits passées par l'ensemble des patients dans ce service hospitalier.

$$mcr_i = \frac{\sum_{x=1}^{i-1} m_x}{\sum_{x=1}^{\omega} m_x}$$

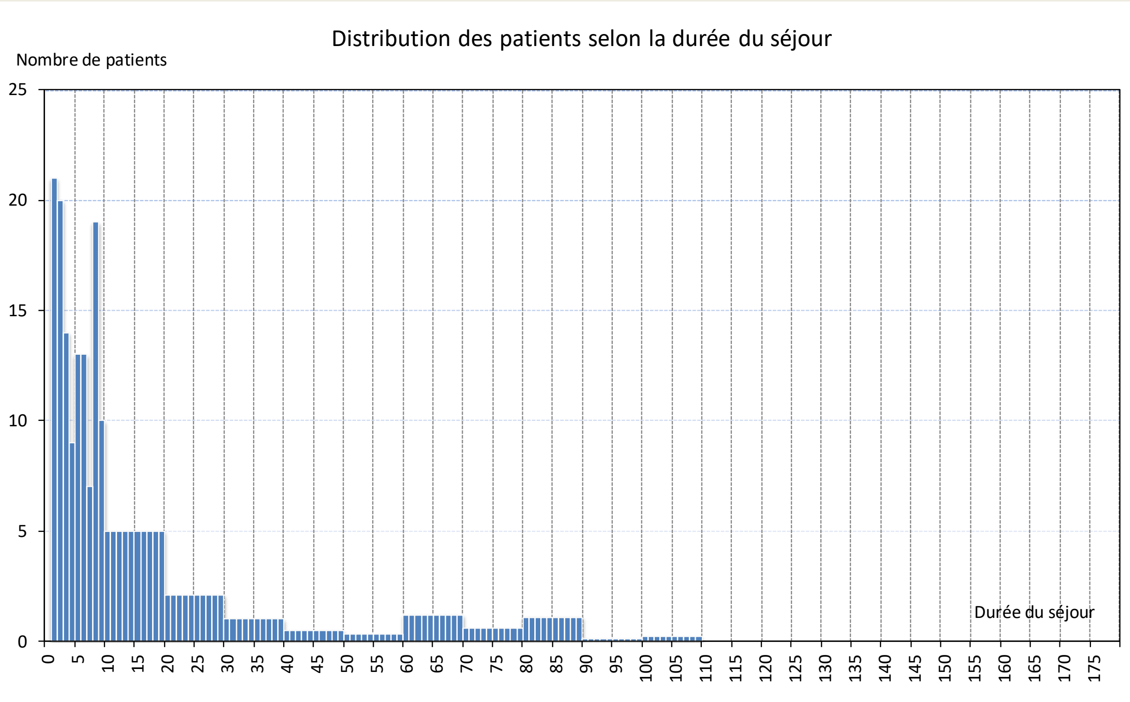
$$mc_{10} = \frac{\sum_{x=1}^{9} m_{x-1}}{\sum_{x=1}^{110} m_x} = \frac{m_1 + m_2 + m_3 + \ldots + m_9}{m_1 + m_2 + \ldots + m_{110}}$$

$$mc_{10} = \frac{21 + 40 + 42 + \ldots + 90}{5\,058} = 11{,}3\ \%$$

2) Représentez graphiquement la distribution des patients selon la durée de séjour.

Corrigé

La durée de séjour est une variable quantitative continue. La représentation graphique est donc un histogramme. Il faut tenir compte de l'inégale amplitude des intervalles de durée. Il faut donc représenter les densités de population.



3) Déterminez le mode, la médiane et la moyenne

Corrigé

a) Le mode correspond à la densité de population la plus importante. Ici, le mode vaut 1 jour.

b) La médiane est comprise entre 9 et 10 jours : en effet, 47 % des patients sont restés moins de 9 nuits, et 51 % sont restés moins de 10 nuits. En faisant l'hypothèse d'une variation linéaire des effectifs sur la durée comprise en 9 et 10 nuits exactes, on peut estimer la durée médiane d'hospitalisation :

$$\frac{Me-9}{10-9}=\frac{50\ \%-47\ \%}{51\ \%-47\ \%}$$

$$Me-9=(10-9)\times\left(\frac{50\ \%-47\ \%}{51\ \%-47\ \%}\right)$$

$$Me=9+\left(\frac{50\ \%-47\ \%}{51\ \%-47\ \%}\right)=9{,}8\ nuits$$

On peut arrondir ce résultat et considérer que la moitié des patients est restée moins de 10 nuits.

c) La durée moyenne correspond à la moyenne des durées ou centres des classes de durées pondérés par la répartition des patients selon la durée d'hospitalisation. Dans le cas présent, la durée moyenne d'hospitalisation est de 20,5 jours.

$$Moy = \sum_{i=1}^{110} (x_i \times f_i)$$

$$Moy = \frac{(8{,}50 \times 1) + (8{,}10 \times 2) + \dots + (20{,}24 \times 15) + \dots + (0{,}81 \times 105)}{100} = 20{,}5 \ nuits$$

4) Déterminez la médiale. Tracez la courbe de Lorenz et calculez l'indice de Gini en vous appuyant sur les couples de points cerclés en noir (cf. tableau 1).

Corrigé

a) La médiale (Ml) correspond à la valeur de la variable (ici la durée de séjour) qui partage la masse statistique (ici le nombre total de nuits d'hospitalisation) en deux parts égales. Elle se calcule comme la médiane. Dans le cas présent, les patients qui sont restés moins de 50 nuits ont passé 47,9 % des 5 058 nuits d'hospitalisation recensées dans ce service hospitalier. Quand on ajoute les patients restés de 50 à 60 nuits, on atteint une proportion de 51,2 %. La médiale se situe donc entre 50 et 60 jours :

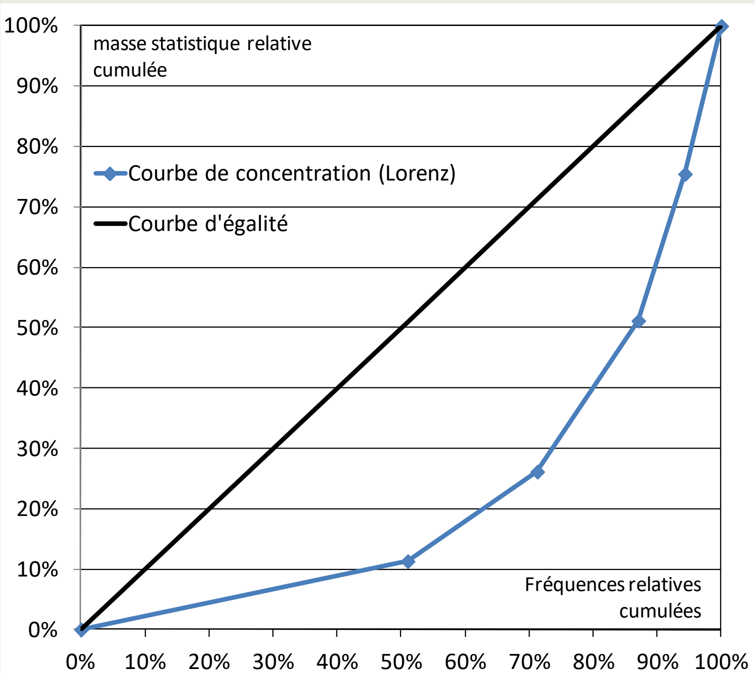
$$\frac{Ml - 50}{60 - 50} = \frac{50\ \% - 47{,}9\ \%}{51{,}2\ \% - 47{,}9\ \%}$$

$$Ml - 50 = (60 - 50) \times \left(\frac{50\ \% - 47{,}9\ \%}{51{,}2\ \% - 47{,}9\ \%} \right)$$

$$Ml = 50 + (60 - 50) \times \left(\frac{50\ \% - 47{,}9\ \%}{51{,}2\ \% - 47{,}9\ \%} \right) = 56{,}4 \ nuits$$

La moitié du nombre total de journées d'hospitalisation (5 058) est donc le fait de patients qui sont restés moins de 56,4 jours.

b) Courbe de Lorenz



c) Détermination de l'indice de Gini (I_G)

L'indice de Gini correspond au rapport entre la surface de concentration (SC) et la surface correspondant à concentration maximale (SC_{max}). Plus le rapport se rapproche de 1 et plus la concentration est forte, et vice versa.

$$I_{Gini} = \frac{S_{concentraiton}}{S_{concentraiton\ \max}}$$

La surface correspondant à la concentration maximale est celle d'un triangle rectangle. Il s'agit de la demi-surface du carré de Gini ($S_{\text{carré de Gini}}$), soit ½ ou 5 000) :

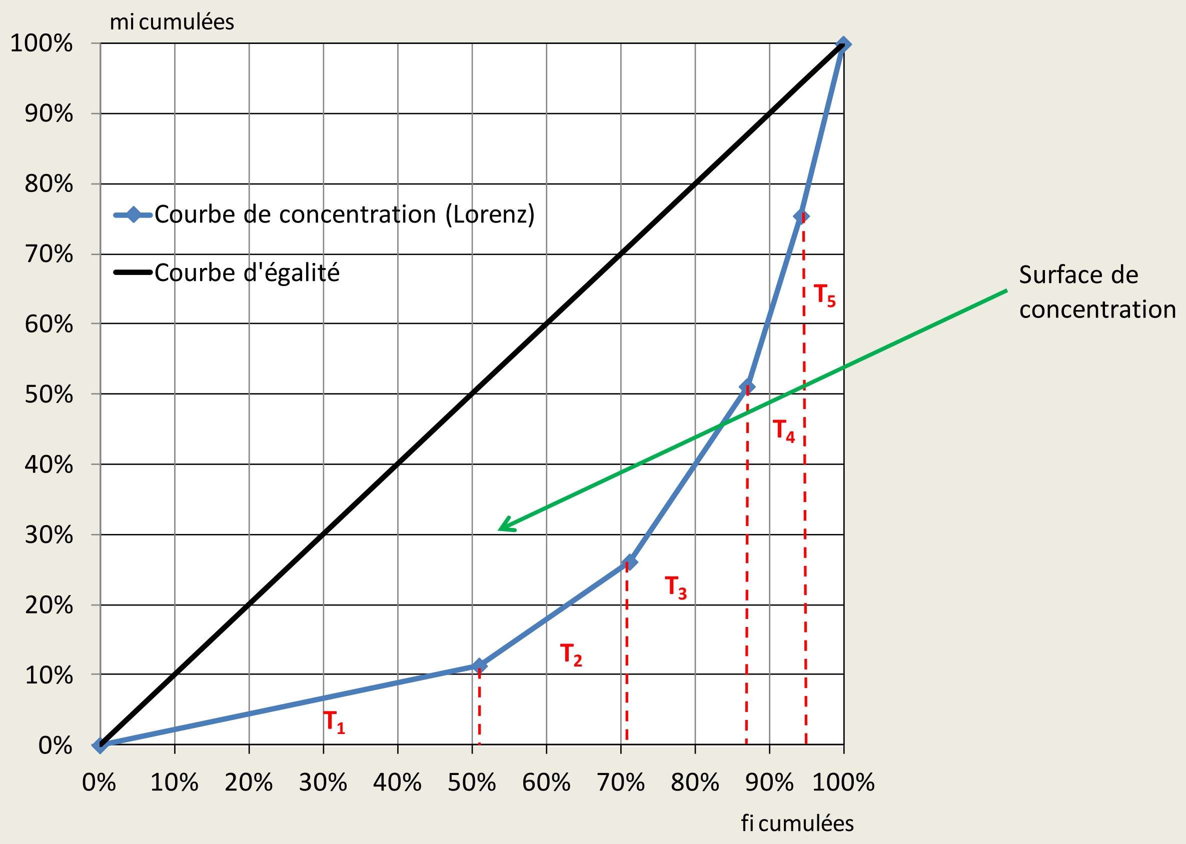
$$S_{Carré\,de\,Gini} = 1 \times 1 = 1 \; donc \; S_{Concentraiton\,\max} = \frac{1}{2}$$

$$ou$$

$$S_{Carré\,de\,Gini} = 100\ (\%) \times 100\ (\%) = 10\ 000$$

$$donc\ S_{Concentraiton\,\max} = \frac{10\ 000}{2} = 5\ 000$$

La surface de concentration maximale correspond au cas où la courbe de Gini se confond avec les bords horizontal inférieur et vertical droit du carré de Gini. Pour la déterminer, on peut calculer la surface se situant sous la courbe de Lorenz puis effectuer la différence entre la surface correspondant à la concentration maximale et celle se trouvant précisément sous la courbe de Lorenz.



On peut décomposer la surface se trouvant sous la courbe de Lorenz en une somme des surfaces de plusieurs trapèzes, ici les trapèzes T_1 à T_5.

La surface d'un trapèze correspond au produit de sa base (b) par la moyenne des hauteurs (h et H) :

$$S_{Trapèze} = b \times \left(\frac{h+H}{2} \right)$$

Dans le cas présent, la base de chaque trapèze correspond à la différence entre deux fréquences relatives cumulées et les hauteurs correspondent pour leur part aux masses statistiques relatives « délimitant » le trapèze.

$$S_{T1} = (51{,}0 - 0{,}0) \times \left(\frac{11{,}3 + 0{,}0}{2} \right) = 288{,}15$$

$$S_{T2} = (71{,}3 - 51{,}0) \times \left(\frac{26{,}2 + 11{,}3}{2} \right) = 380{,}625$$

$$S_{T3} = (87{,}0 - 71{,}3) \times \left(\frac{51{,}2 + 26{,}2}{2} \right) = 607{,}59$$

$$S_{T4} = (94{,}3 - 87{,}0) \times \left(\frac{75{,}5 + 51{,}2}{2} \right) = 462{,}455$$

$$S_{T5} = (100{,}0 - 94{,}3) \times \left(\frac{100{,}0 + 75{,}5}{2} \right) = 500{,}175$$

On peut maintenant déterminer la surface sous la courbe de Lorenz, en déduire la surface de concentration puis calculer l'indice de Gini :

$$S_{sous\,courbe\,Lorenz} = \sum_{i=1}^{5} S_{T_i} = 288{,}15 + \ldots + 500{,}175 = 2\,238{,}995 \cong 2\,239$$

$$S_{concentration} = S_{concentration\,\max} - S_{sous\,courbe\,Lorenz} = 5\,000 - 2\,239 = 2\,761$$

$$I_{Gini} = \frac{S_{concentration}}{S_{concentration\,\max}} = \frac{2\,761}{5\,000} = 0{,}5522 \cong 0{,}55$$

L'indice de Gini est de 0,55, c'est-à-dire que la concentration observée des durées de séjour correspond à 55 % de la concentration théorique maximale des durées de séjour (un patient occuperait la totalité des lits sur la période observée, ce qui est bien sûr impossible). L'indice traduit ici une concentration assez importante, ce que la lecture du tableau révélait d'ailleurs déjà très bien : 13 % des malades (ceux qui restent au moins 60 jours, soit deux mois) ont passé près de la moitié (48,8 %) des nuits d'hospitalisation assurés par ce service.

5) Synthétisez l’information contenue dans le tableau 1 dans une note de 10 lignes maximum à destination du directeur de ce service hospitalier.

Proposition de commentaires :

La plupart des patients de ce service effectue des séjours de courte durée :

- la durée modale d'hospitalisation est de 1 nuit : près de 10 % des personnes restent une nuit ;
- la moitié des patients ont passé moins de 10 nuits dans ce service (médiane).

Toutefois, la durée moyenne des séjours est nettement plus élevée (environ 20 nuits). Cet écart entre la distribution des patients caractérisée par une assez forte concentration des petites durées de séjour (70 % des patients restent moins de 20 jours) et la valeur moyenne est lié au poids très important que pèse une minorité de patients (qui eux restent longtemps) sur l'occupation des infrastructures.

Ainsi, la moitié de la durée totale des séjours (5 058 nuits) est due aux patients hospitalisés pendant plus de 60 jours. Or, ces derniers ne représentent que 13 % des patients.

En conséquence, une moitié des infrastructures est occupée par les rares patients de longue durée. En revanche, l'autre partie des infrastructures est utilisée par des patients dont la plupart (51 % des 87 %, soit 58 %) restent moins de 10 jours.