Probabilité et statistique I

Correction de l'examen du 3 juin 2020

1 Exercices de statistique descriptive

1. (a) Plusieurs répartitions étaient possibles et pertinentes. Voici la description d'une démarche de construction possible.

   Premièrement, utilisons la formule de Sturges pour déterminer combien de classes celle-ci préconise de construire. Le nombre d'employés qui travaillent en tant que `Sales Representatives` étant de 78, on a

   $$1 + \frac{10}{3}\log_{10}(78) = 7.31.$$

   Ainsi, la formule de Sturges suggère de construire un nombre de classes égal au plus petit nombre entier supérieur ou égal à 7.31, à savoir 8. Ensuite, le diagramme en tiges et feuilles nous informe que la plus petite valeur vaut 1 et la plus grande vaut 29. Il paraît donc adéquat d'utiliser 0 et 30 comme bornes inférieure et supérieure des classes que nous allons construire. Par ailleurs, on note beaucoup d'observations en début de distribution, et beaucoup moins en fin de distribution. Ainsi, les premières classes devraient avoir de petites amplitudes et les observations les plus grandes devraient être regroupées en une ou deux classes. Par exemple, on peut considérer 4 classes d'amplitude 2.5 pour séparer les employés parcourant moins de 10 km en 4 groupes, et deux classes d'amplitude 10 pour séparer les autres employés en deux groupes. Cela mène à une décomposition en 6 classes (nombre inférieur à ce qui est préconisé par Sturges mais qui peut se justifier par la faible masse d'individus au delà de 10 km). Une autre option pourrait être d'utiliser des classes d'amplitude 2 km entre 0 et 10, tout en gardant deux classes d'amplitude 10 km pour le reste de la distribution (cela mènerait à 7 classes). A titre d'illustration, la répartition en 6 classes proposée dans le tableau ci-dessous sera utilisée dans la suite.

Classes	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
$[0; 2.5]$	17	17	0.22	0.22
$]2.5; 5]$	13	30	0.17	0.38
$]5; 7.5]$	4	34	0.05	0.44
$]7.5; 10]$	24	58	0.31	0.74
$]10; 20]$	11	69	0.14	0.88
$]20; 30]$	9	78	0.12	1

   (b) L'histogramme construit à partir de cette répartition en classes est représenté à la Figure 1 (vu que les classes sont d'amplitudes variables, il convenait d'ajuster les hauteurs des rectangles de manière à tenir compte de ces amplitudes différentes). On y observe une certaine concentration d'individus dans les deux premières classes, parcourant donc moins de 5 km pour se rendre au travail, ainsi qu'un pic important dans la classe $]7.5; 10]$. On constate un certain étalement sur la droite après ce pic, les distances supérieures à 10 km étant associées à moins d'individus, mais s'étalant sur un intervalle plus important, allant jusqu'à 30 km.

   (c) La classe médiane étant, par définition, la première classe dont la fréquence cumulée dépasse 0.5, il s'agit ici de la classe $]7.5; 10]$. Dans l'ogive des fréquences cumulées, les points de coordonnées $(7.5; 0.44)$ et $(10; 0.74)$ sont alignés. La droite passant par ces deux points a pour équation

   $$y - 0.44 = \frac{0.74 - 0.44}{10 - 7.5}(x - 7.5).$$

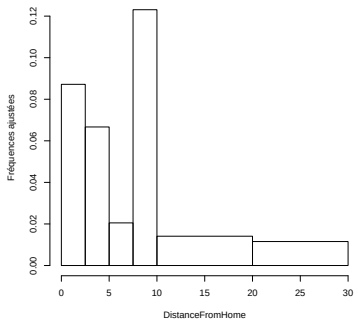

FIGURE 1 – Histogramme d'aire unitaire de la variable `DistanceFromeHome` dans le groupe des employés travaillant comme `Sales Representative`

La médiane $\tilde{x}$ étant associée à une fréquence cumulée de 0.5, elle vérifie l'équation

$$\begin{aligned} 0.5 - 0.44 &= \frac{0.74 - 0.44}{10 - 7.5}(\tilde{x} - 7.5). \\ \Leftrightarrow \tilde{x} &= 0.06 \times \frac{2.5}{0.3} + 7.5 \\ \Leftrightarrow \tilde{x} &= 8. \end{aligned}$$

La médiane calculée sur la série brute valant 9, on la sous-estime un peu. En effet, dans la classe médiane, 4 observations sont égales à 8, 9 observations sont égales à 9 et 11 observations sont égales à 10. Ainsi, la vraie répartition est plutôt concentrée à droite, et, en faisant l'hypothèse de répartition uniforme pour estimer la valeur de la médiane, on la sous-estime.

2. (a) Les distributions conditionnelles de la variable `Attrition` en fonction des deux modalités de la variable `Gender` s'obtiennent en se focalisant sur chacune des lignes du tableau de contingence donné. Ces distributions sont donc caractérisées par les tableaux statistiques suivants :

Dist cond d'`Attrition` sachant que l'employé est une femme :

Modalités	Effectifs	Fréq
`No`	86	0.835
`Yes`	17	0.165
	103	1

Dist cond d'`Attrition` sachant que l'employé est un homme :

Modalités	Effectifs	Fréq
`No`	142	0.79
`Yes`	37	0.21
	179	1

Les diagrammes en secteurs de la Figure 2 représentent les deux distributions conditionnelles. Des diagrammes en barres étaient aussi possibles, tout en utilisant les fréquences et non les effectifs puisque ceux-ci ne sont pas directement comparables vu les effectifs totaux différents.

Deux constats sont visibles. Tout d'abord, les deux distributions sont similaires en ce sens qu'elles correspondent toutes deux à une fréquence en faveur de la modalité `No` nettement supérieure à celle relative à la seconde modalité. Par ailleurs, on constate que le secteur associé à la modalité `Yes` chez les hommes est un peu plus important que celui relatif à cette modalité chez les femmes (les tableaux donnent plus précisément les informations concernant ce léger "avantage" des hommes en faveur du départ de la société). L'impact est cependant peu marqué et il reste intéressant d'investiguer plus en détail l'existence d'une association éventuelle entre les deux variables.

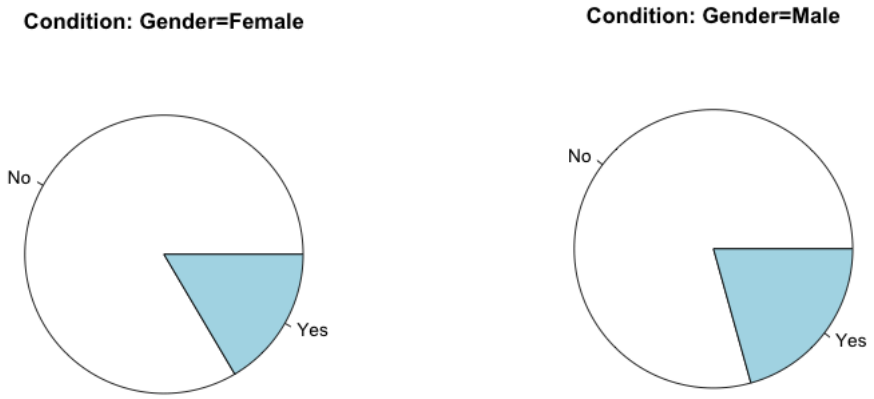

FIGURE 2 – Diagrammes en secteurs représentant les deux distributions conditionnelles

(b) Selon l'énoncé, les deux variables qualitatives considérées ici seraient "indépendante" si l'effectif bivarié n_{ij} était égal à

$$nf_{i\bullet}f_{\bullet j}$$

pour les valeurs de i et j correspondant au tableau de contingence donné. Afin de pouvoir déterminer si cette égalité est valable pour toutes valeurs de i et j possibles, il convient de calculer d'abord les fréquences marginales. Celles-ci s'obtiennent facilement à partir du tableau de contingence complet (avec l'ajout des distributions marginales) :

	`Attrition`		Dist. marg.
`Gender`	`No`	`Yes`	`Gender`
`Female`	86	17	103
`Male`	142	37	179
Dist marg `Attrition`	228	54	282

En utilisant les notations habituelles du cours, les fréquences marginales de la variable `Gender` (en ligne, avec la modalité 1 correspondant aux femmes et la modalité 2 correspondant aux hommes) sont données par

$$f_{1\bullet} = \frac{103}{282} = 0.37 \text{ et } f_{2\bullet} = \frac{179}{282} = 0.63$$

tandis que celles de la variable `Attrition` (en colonne, la modalité 1 étant `No` et la 2 `Yes`) :

$$f_{\bullet 1} = \frac{228}{282} = 0.81 \text{ et } f_{\bullet 2} = \frac{549}{282} = 0.19$$

Le tableau ci-dessous reprend dès lors pour les diverses associations des valeurs 1 et 2 des indices i et j, la quantité $nf_{i\bullet}f_{\bullet j}$

	`Attrition`	
`Gender`	`No`	`Yes`
`Female`	84.5	19.8
`Male`	143.9	33.7

On constate que même si les quantités calculées ci-dessus ne sont pas exactement égales aux effectifs bivariés du tableau de contigence de l'énoncé, les écarts sont très faibles (écart maximal égal à 3.3, par rapport à un effectif total égal à 282).

(c) A l'aide des deux analyses, on peut en conclure qu'il n'y a pas indépendance au sens strict entre les deux variables (puisque les égalités en (b) ne sont pas vérifiées) mais que le lien de dépendance (ou l'association) semble extrêmement faible (vu les écarts très faibles observés et vu le caractère fort similaire des deux distributions conditionnelles). Le lien de dépendance se traduit par une association légèrement plus forte entre les modalités `Male` et `Yes` que ce qui serait observé en cas d'indépendance (on voit que c'est pour ce couple que l'écart entre l'effectif bivarié et la valeur calculée sous indépendance est le plus important). L'effet est limité mais va néanmoins dans le sens de l'intuition du service GRH.

2 Analyse de données

1. (a) Les distributions de la variable `YearsAtCompany` lorsque les employés sont décomposés en les deux groupes définis par la variable `Attrition` sont comparées à l'aide de boîtes à moustaches et de polygones des fréquences aux Figures 3 et 4.

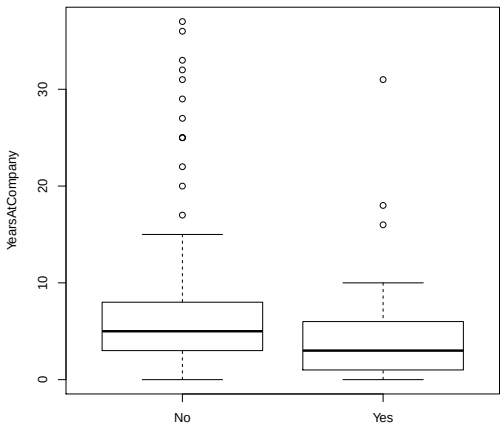

FIGURE 3 – Boîtes à moustaches de la variable `YearsAtCompany` lorsque les employés sont décomposés en les deux groupes définis par la variable `Attrition`

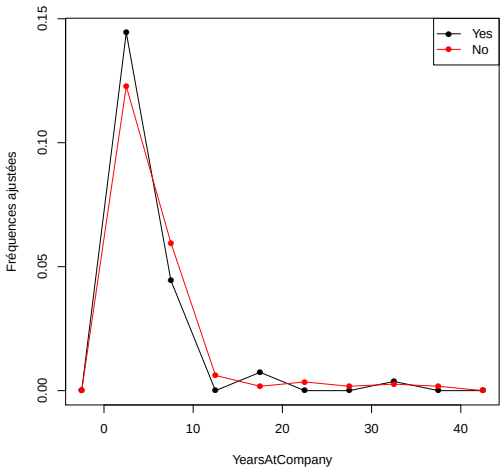

FIGURE 4 – Polygones des fréquences cumulées de la variable `YearsAtCompany` lorsque les employés sont décomposés en les deux groupes définis par la variable `Attrition`

(b) Le tableau suivant reprend certains paramètres statistiques calculés sur la variable `YearsAtCompany` dans les deux groupes.

	`Attrition = Yes`		`Attrition = No`
Moyenne	4.39	$<$	6.22
Médiane	3	$<$	5
Ecart-Type	5.31	$\approx$	6.24
EIQ	5	$=$	5
Dissymétrie (Fisher)	2.88	$\approx$	2.72

On peut observer que les deux distributions diffèrent légèrement en tendance centrale, puisque la moyenne et la médiane sont toutes les deux plus importantes dans le groupe des employés qui n'ont pas quitté l'entreprise. Ceci se remarque également grâce aux

boîtes à moustaches sur lesquelles on peut visualiser la différence entre les deux médianes. En termes de dispersion, ni les graphiques ni les valeurs des paramètres statistiques ne permettent de détecter une différence flagrante entre les deux distributions. En effet, les longueurs des deux boîtes, traduites par les écarts interquartiles, sont exactement les mêmes. L'écart-type est un tout petit peu plus élevé dans le groupe des employés n'ayant pas quitté l'entreprise, mais la différence reste néanmoins négligeable. Enfin, les deux distributions se valent en termes de dissymétrie, les deux distributions ayant des coefficients de dissymÃ©trie de Fisher semblables, traduisant un important étalement sur la droite que l'on visualise très clairement sur les polygones des fréquences ainsi que sur les boîtes à moustaches.

(c) Bien que les deux distributions diffèrent quelque peu en tendance centrale, et que les employés ayant quitté l'entreprise soient en moyenne un peu plus jeunes que ceux qui ne l'ont pas quittée, la différence reste minime, et les deux distributions se recouvrent assez bien. L'intuition du service GRH n'est donc pas clairement soutenue par les analyses précédentes.

2. (a) Le diagramme de dispersion représentant la variable `MonthlyIncome` en fonction de la variable `TotalWorkingYears` repris à la Figure 5 montre qu'il y a un lien linéaire assez fort entre les deux variables. En effet, comme attendu, le salaire augmente au fur et à mesure que le nombre d'années d'expérience professionnelle augmente. Cependant, le coefficient de corrélation entre ces deux variables n'est que de 0.55. Cette valeur, assez faible par rapport à ce que l'on aurait pu espérer, est dûe à la présence d'un petit nombre d'observations atypiques, situées dans le coin inférieur droit du diagramme de dispersion et qui ne suivent pas la tendance linéaire générale. Les individus concernés sont les membre du personnel retraités. Ceux-ci sont donc caractérisés par un nombre d'années de travail élevé, mais par un salaire nul ou presque.

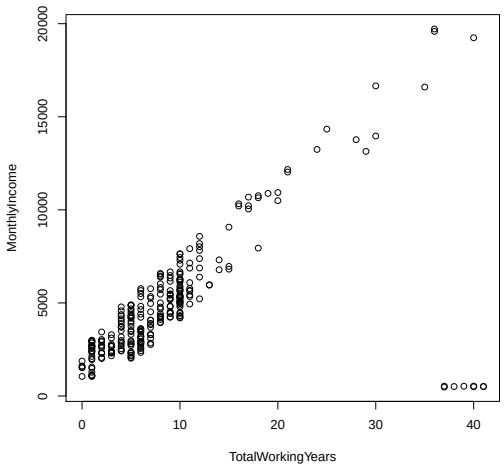

FIGURE 5 – Diagramme de dispersion de la variable `MonthlyIncome` en fonction de la variable `TotalWorkingYears`

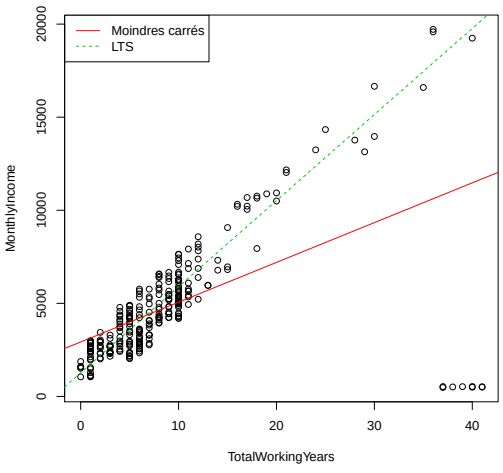

FIGURE 6 – Diagramme de dispersion de la variable `MonthlyIncome` en fonction de la variable `TotalWorkingYears` et droites de régression

(b) Le diagramme de dispersion de la variable `MonthlyIncome` en fonction de la variable `TotalWorkingYears` sur lequel sont ajoutées la droite de régression des moindres carrés ainsi qu'une droite de régression robuste est repris à la Figure 6. La droite de régression des moindres carrés, représentée en rouge, est très clairement attirée par le groupe

d'observations atypiques, et ne traduit pas la tendance linéaire du nuage de points principal. Pour tenter de remédier à ce problème, on peut par exemple utiliser une régression LTS, afin de ne pas tenir compte des résidus les plus grands dans l'estimation des paramètres de la droite de régression. Puisque le nombre d'employés retraités est de 8, nous pouvons décider de ne considérer que les $282 - 8 = 274$ plus petits résidus. La droite ainsi obtenue est représentée en vert sur le diagramme de dispersion, et l'on remarque qu'elle colle beaucoup mieux à la tendance linéaire globale.

(c) La droite de régression des moindres carrés est très sensible à la présence de données atypiques. Dès lors, cette droite est attirée par ces données atypiques, qui, dans ce contexte, ne suivent pas la tendance linéaire générale. Utiliser une technique robuste permet donc de fournir une droite qui traduit mieux le comportement de la majorité des observations de la base de données.