Statistique descriptive

Bachelier en sciences informatiques
Correction de l'examen écrit du 7 juin 2019

QCM

Les questions des différents questionnaires n'étaient pas nécessairement exactement égales à celles développées ci-dessous mais correspondaient au même raisonnement/développement.

1. Dans un club de sport il y a 60% de garçons. On sait que 40% des garçons de ce club font du tennis et 30% des filles de ce club font du tennis. Le pourcentage global de personnes pratiquant le tennis dans ce club est de
   ☐ 24% ■ 36% ☐ 70% ☐ On ne peut pas le calculer.
   `Justification` : $0.6 \times 0.4 + 0.4 \times 0.3 = 0.4 \times 0.9 = 0.36$.

2. A la fin du mois d'avril, une tablette coûtait 999 EURO après avoir subi une baisse de 10% depuis le début du mois d'avril. Quel était le prix de cette tablette (arrondi à l'unité la plus proche) au début du mois d'avril ?
   ☐ 1099 EURO ■ 1110 EURO ☐ 899 EURO
   `Justification` : Si x_0 est le prix au début du mois d'avril, le prix à la fin du mois vaut $999 = x_0 - 0.1x_0$, ce qui implique $x_0 = \frac{999}{0.9} = 1110$.

3. Soit une variable quantitative mesurée en fonction d'une certaine unité de mesure (mètre, minute,...). En quelle unité l'écart interquartile s'exprime-t-il ?
   ■ unité ☐ nombre pur (sans unité) ☐ unité au carré

4. Lorsque l'on calcule le coefficient γ_1 de Fisher (moment centré d'ordre 3 divisé par l'écart-type au cube), quelle caractéristique de la série essaye-t-on de quantifier ?
   ☐ la tendance centrale ☐ la dispersion ■ la dissymétrie

5. Une enquête menée auprès d'un grand nombre de jeunes s'intéressait à leur loisir préféré (un seul choix possible entre les quatre modalités : `Shopping`, `Jeux Vidéos`, `Sport`, `Télé/Netflix`). Les résultats sont illustrés sur le diagramme en secteurs [1] ci-dessous :
   Quel est l'effectif du mode de cette série ?
   ☐ 60% ☐ 60 ■ On n'a pas assez d'information pour le calculer.

6. Après avoir corrigé les examens de sa classe de 25 élèves en attribuant des cotes sur 20 points, un professeur se rend compte que la moyenne vaut 9, avec un écart-type égal à 3 tandis que sa meilleure cote est un 13/20. Il décide, en vue de ne pas démoraliser ses élèves, de modifier ses cotes à l'aide de la transformation affine suivante : $x'_i = \frac{2}{3}x_i + 10$.

   - L'écart-type des cotes transformées
     ☐ suffit la même transformation
     ■ est multiplié par 2/3
     ☐ reste inchangé
     ☐ On n'a pas assez d'information pour conclure.

1. Il y avait une erreur dans les étiquettes associées aux secteurs, avec une somme de pourcentages qui excédait 100%, ce que certains ont signalé à juste titre.

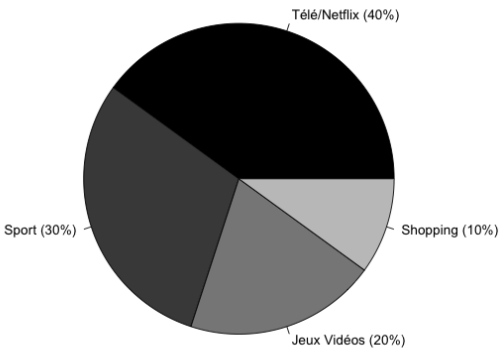

FIGURE 1 – Diagramme en secteurs des loisirs d'un grand nombre de jeunes

- En considérant que tous les élèves doivent pouvoir bénéficier de cette transformation, quelle(s) contrainte(s) faut-il imposer sur les cotes initiales minimale et maximale pour que chaque cote transformée soit compris en 0 et 20 ?
  ☐ Pas de contrainte sur $x_{(1)}$ mais $x_{(25)} \leq 10$
  ☐ $x_{(1)} \geq 3$ et $x_{(25)} \leq 15$
  ■ Pas de contrainte sur $x_{(1)}$ mais $x_{(25)} \leq 15$
  ☐ Aucune contrainte
- Avant la modification des cotes de ce professeur, la corrélation entre ses points et ceux attribués aux mêmes élèves par sa collègue enseignant les mathématiques était égale à 0.85. Que devient la corrélation après la transformation du professeur de physique, les cotes de mathématique restant inchangées ?
  ☐ Elle est multiplié par 2/3
  ■ Elle reste inchangé
  ☐ On n'a pas assez d'information pour conclure.

  `Justification` : Lors d'une transformation affine $x'_i = ax_i + b$, le nouvel écart-type est donné par $|a|s$. Il faut que chaque observation soit comprise entre 0 et 20. Ainsi, $0 \leq \frac{2}{3}x_i + 10 \leq 20$ pour tout i. Cela signifie $-15 \leq x_i \leq 15$.

7. Une enquête a été menée auprès de 1000 personnes afin de déterminer le temps consacré au sport par semaine (les données étaient enregistrées en minutes). L'ogive des fréquences cumulées construite après avoir regroupé les observations en 5 classes (dont les bornes s'expriment en heures) est représentée à la Figure 2.

   - La proportion de personnes pratiquant entre 1h30 et 4h de sport par semaine est égale à
     ☐ 0.35 ■ 0.40 ☐ 0.65 ☐ On ne peut pas la calculer.
   - Sur l'histogramme d'aire unitaire représentant cette série groupée, la hauteur du rectangle correspondant à la classe $]3; 5]$ serait égale à
     ■ 0.15 ☐ 0.30 ☐ 0.5 ☐ On n'a pas assez d'information pour la calculer.

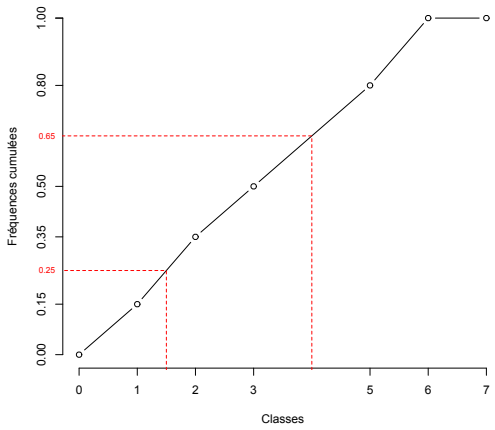

FIGURE 2 – Ogive d'une série décomposée en 5 classes

`Justification` : Grâce à l'ogive, nous pouvons retrouver les fréquences cumulées associées à 1h30 et 4h. La fréquence recherchée est donnée par $F(4) - F(1.5) = 0.65 - 0.25 = 0.4$. La hauteur de la 4ème classe est donnée par le rapport entre sa fréquence et sa largeur, à savoir $0.30/2 = 0.15$.

8. Soit une série statistique quantitative bivariée constituée des observations (x_i, y_i), $1 \leq i \leq n$, dont les moyennes marginales sont nulles et dont les variances marginales sont égales à 1. On précise également que l'équation de la droite de régression de Y en X, estimée par la technique des moindres carrés, est donnée par $y = x/2$. Le coefficient de détermination associé à cet ajustement est égal à

   ☐ 1 ☐ 1/2 ■ 1/4 ☐ On n'a pas assez d'information pour le calculer.

   `Justification` : La pente de la droite de régression estimée par la technique des moindres carrés est donnée par $\hat{a} = \frac{s_{xy}}{s_x^2}$. Comme les variances marginales sont égales à 1, $s_{xy} = 1/2$. Ainsi, le coefficient de détermination vaut $R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{1/4}{1 \times 1}$.

9. La distribution du nombre d'écoles implantées dans 100 communes belges est décrite par la courbe cumulative des fréquences cumulées à la Figure 3.

   - Selon la convention imposée dans le cours, le 9ème décile de cette série vaut :

     ☐ 4 ■ 4.5 ☐ 0.9 ☐ Il n'est pas défini

   - La moyenne tronquée de paramètre $\alpha = 0.25$ de la série vaut

     ■ 2.9 ☐ 3 ☐ Il manque des informations pour calculer cette moyenne

`Justification` : Comme la fréquence cumulée de 90% correspond à un "palier" dans la courbe cumulative, suivant la convention vue au cours, le 9ème décile est donné par le milieu du palier, à savoir 4.5.
Par le graphique, comme nous considérons 100 communes, nous pouvons retrouver les effectifs cumulés ($N_i = 100F_i$) et les effectifs de chaque modalité. Celles-ci sont reprises ci-dessous :

X	1	2	3	4	5
F_i	0.15	0.35	0.70	0.90	1
N_i	15	35	70	90	100
n_i	15	20	35	20	10

À partir de ces effectifs, nous pouvons calculer la moyenne tronquée où les 25 plus petites observations et les 25 plus grandes observations ne sont pas prises en compte : $\bar{x}_{0.25} = \frac{10\times2+35\times3+5\times4}{50} = 2.9$.

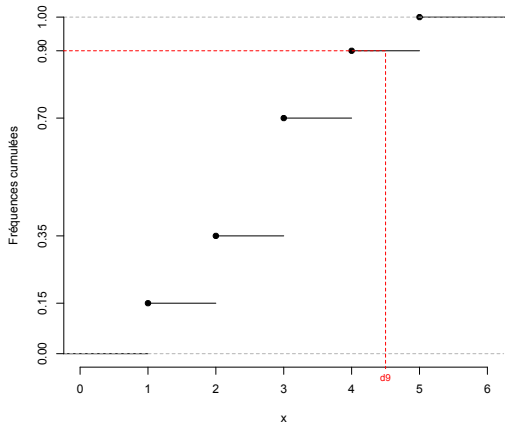

FIGURE 3 – Représentation de la distribution des fréquences cumulées

10. Une droite de régression estimée par la technique des moindres carrés a été ajustée au nuage de points de la Figure 4. Des graphiques de résidus représentés en fonction de la variable explicative X sont disponibles à la Figure 5. Quel graphique des résidus représente effectivement les résidus de la droite ajustée sur les données de la Figure 4 ?
   ☐ Le graphique de gauche
   ☐ Le graphique du milieu
   ■ Le graphique de droite
   ☐ Aucun des trois graphiques

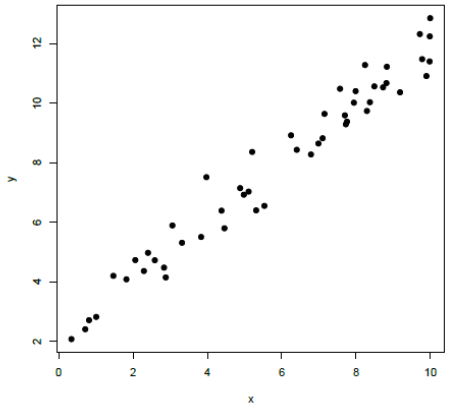

FIGURE 4 – Diagramme de dispersion construit sur les deux variables X et Y

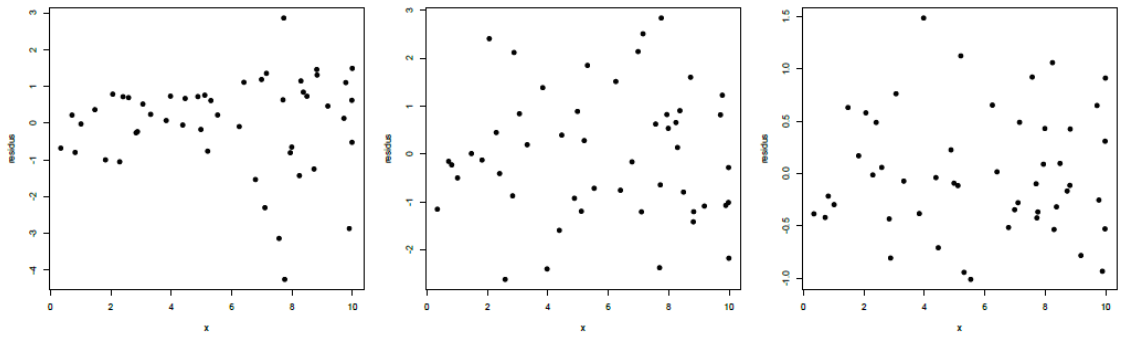

FIGURE 5 – Graphique des résidus représentés en fonction des valeurs de X

`Justification` : Au vu du diagramme de dispersion, aucune structure ne peut être présente dans le graphique des résidus en fonction des valeurs de X, ce qui élimine le graphique de gauche. Ensuite, les résidus étant visiblement trop grands par rapport aux écarts obtenus entre les valeurs observées et les valeurs ajustées sur le diagramme de dispersion, le graphique du milieu doit également être éliminé. Le graphique de droite est enfin confirmé lorsque l'on tient compte du fait que les valeurs en abscisse et en ordonnée sont telles que celles naturellement attendues pour le modèle ajusté.

11. Que vaut la valeur pivot supérieure (ou de droite) de la série représentée par la boîte à moustaches de base (les moustaches vont jusqu'aux valeurs minimale et maximale) de la Figure 6 ?

☐ 7 ■ 11.5 ☐ 14 ☐ Il manque des informations pour calculer cette valeur

`Justification` : Par définition, la valeur pivot supérieure se situe à 1.5 écart-interquartile du 3ème quartile. L'écart-interquartile est égal à $7 - 4 = 3$, ce qui donne $7 + 1.5 \times 3 = 11.5$.

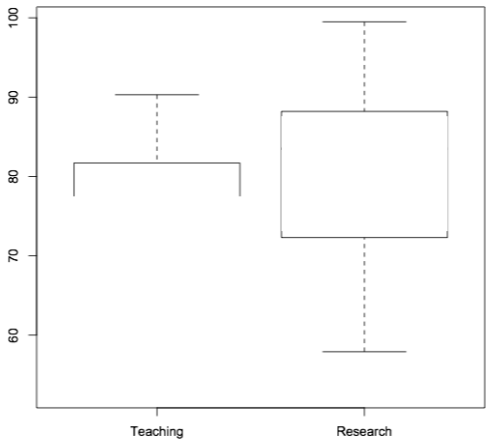

FIGURE 6 – Boîte à moustaches d'une série quelconque

Exercices

1. (a) Comme l'effectif total est de 82, nous obtenons une première équation :

$$n_{22} + n_{43} = 17.$$

D'autre part, le nombre moyen d'enfants nés de femmes ayant au moins 3 frères et soeurs vaut 3.3. Cette information nous permet d'écrire une seconde équation :

$$\begin{aligned} 3.3 &= \frac{1 \times 1 + 2 \times 2 + 3n_{43} + 4 \times 3 + 5 \times 2}{n_{45} + 8} \\ &\Leftrightarrow 8 \times 3.3 + 3.3n_{45} = 27 + 3n_{45} \\ &\Leftrightarrow n_{45} = 2. \end{aligned}$$

Pour conclure, grâce à la première équation, on obtient que le dernier effectif vaut 15.

(b) Les distributions conditionnelles de Y sachant que $X = 0$ et $X = 2$ sont reprises Table 1

$Y_{\mid X=0}$	Effectif	Fréq.	Fréq. cumulées	$Y_{\mid X=2}$	Effectif	Fréq.	Fréq. cumulées
1	4	0.22	0.22	1	1	0.06	0.06
2	9	0.50	0.72	2	5	0.29	0.35
3	4	0.22	0.94	3	9	0.53	0.88
4	1	0.06	1	4	1	0.06	0.94
5	0	0	1	5	1	0.06	1
Total	18	1		Total	17	1	

TABLE 1 – Tableau statistique des distributions conditionnelles de Y sachant que $X = 0$ et $X = 2$

i. Le mode de la distribution vaut 2 lorsque les femmes sont enfants uniques et 3 lorsque les femmes ont 2 frères et soeurs.

ii. Les courbes cumulatives des deux distributions peuvent être comparées à l'aide des fréquences cumulées reprises également Table 1. Ces courbes sont représentées Figure 7.

iii. Les courbes cumulatives et les modes sont concordants avec l'hypothèse des sociologues. En effet, les femmes ayant 2 frères et soeurs ont globalement plus d'enfants que les femmes étant enfants uniques.

2. (a) Le facteur 2 répartit les $n = 18$ observations (mois d'activité) en deux groupes (A : plus de 10 tubes - B : moins de 10 tubes), d'effectif $n_A = n_B = 9$. Cette répartition nous permet de retrouver facilement la moyenne globale :

$$\bar{x} = \frac{n_A \bar{x}_A + n_B \bar{x}_B}{n} = \frac{9 \times 33 + 9 \times 41}{18} = 37.$$

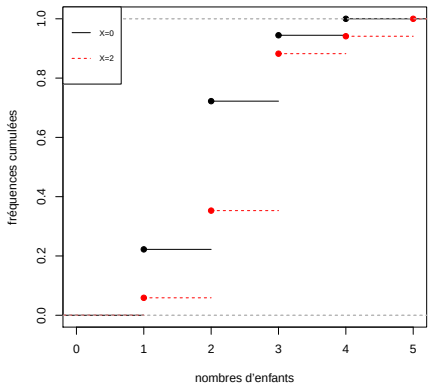

FIGURE 7 – Courbes cumulatives des fréquences cumulées.

La variance totale est quant à elle décomposée en deux parties : la variance dans les groupes et la variance entre les groupes :

$$\begin{aligned} s^2 &= \text{variance dans les groupes} && + && \text{variance entre les groupes} \\ &= \frac{n_A s_A^2 + n_B s_B^2}{n} && + && \frac{n_A(\bar{x} - \bar{x}_A)^2 + n_B(\bar{x} - \bar{x}_B)^2}{n} \end{aligned}$$

Chacune des quantités exploitées dans cette formule est disponible dans l'énoncé. On obtient donc

$$s^2 = \frac{9 \times 172 + 9 \times 176}{18} + \frac{9\ (\bar{x} - 33)^2 + 9\ (\bar{x} - 41)^2}{18} = 190.$$

(b) Considérons maintenant le facteur 1. Celui-ci permet également de répartir les 18 observations en deux autres groupes (P : avec promotion - S : sans promotion), dont les effectifs sont cette fois égaux à 6 et 12. Par définition, la moyenne globale et la variance totale ne changent pas. La part de la variance entre les groupes associée au facteur 1 est donc donnée par :

$$x = \frac{\text{variance entre les groupes}}{\text{variance totale}} = \frac{\frac{6(20-37)^2+12(37-45.5)^2}{18}}{190} = \frac{144.5}{190} = 0.76.$$

Pour conclure, le facteur 1 est le plus déterminant des deux car il permet d'expliquer une plus grande part de la variabilité : 76% de la variabilité peut être expliquée par la présence ou l'absence de promotion alors que la quantité de tubes ne permet d'expliquer que 8.42% de la variance globale.

3. Pour commencer, nous pouvons observer que le sondage a permis de bien prédire les résultats des différents partis, excepté pour Ecolo. En effet, toutes les prédictions se situent à au plus 3.1 % des vrais résultats, sauf pour le parti Ecolo pour lequel le sondage avait surestimé le score de 7.5%, ce qui dépasse la marge d'erreur annoncée.

   Il était également intéressant de déterminer à quel point la hiérarchie entre les partis, telle qu'observée à partir des résultats du sondage, s'est trouvée confirmée par les élections. On constate que le mode des élections est celui qui avait été prédit par le sondage,

à savoir le parti PS qui récolte 26.2% des voix. Plus globalement, l'ordre des partis est presque similaire à celui prédit par le sondage, il y a juste une inversion entre Ecolo et le MR. Cela est notamment du au fait que le sondage à surestimé le score d'Ecolo et sous-estimé celui du MR. En effet, le sondage a surestimé les scores pour seulement deux partis, à savoir Ecolo et le CDH.

En ce qui concerne le sondage, on pouvait également s'étonner à propos du fait que les pourcentages des sondés des trois régions ne respectent pas les pourcentages "théoriques" de répartition de la population belge.

Analyse de données

Les questions des différents questionnaires n'étaient pas nécessairement exactement égales à celles développées ci-dessous mais correspondaient au même raisonnement/développement.

1. Dans les données initiales, 4 données sont manquantes (codée `NA` par le logiciel `R`). Il s'agit des observations n° 30, 31, 32 et 38. De plus, deux erreurs grossières sont facilement détectables aux lignes 9 et 23 car des scores ne sont pas compris entre 0 et 10 comme prévu. Ces différentes lignes sont reprises Table 2 :

	Vitesse	Prix	Service	Qualite	Taille	Strategie	TypeClient
9	4.00	3.50	-3.70	8.70	grande	telephone	occasionnel
23	2.40	1.50	1.90	11.20	grande	telephone	nouveau
30	NA	3.80	3.30	8.20	petite	telephone	regulier
31	5.20	2.00	3.70	4.60	petite	NA	regulier
32	3.40	3.70	3.50	8.40	NA	telephone	nouveau
38	2.40	2.00	NA	8.80	grande	telephone	nouveau

TABLE 2 – Lignes contenant des données manquantes ou atypiques. Elles sont supprimées de la base de données pour les questions suivantes.

2. (a) Différents graphiques peuvent être utilisés pour comparer l'indice de satisfaction `Qualite` en fonction de la taille de l'entreprise. Par exemple, des boites à moustaches, des fonctions de répartitions empiriques ou des ogives, des polygones ou des histogrammes. Ces différents graphiques sont repris Figure 8. Tous ceux-ci permettent de conclure qu'il y a une distinction assez claire entre les deux groupes, les grandes entreprises ont globalement un indice de qualité supérieure à celui des petites entreprises.

   (b) Différents paramètres de localisation (moyenne, médiane, quartiles, ...), de dispersion (étendue, variance, ...) et de dissymétrie (Fisher, Yule et Kendall) sont calculés pour chacun des deux groupes. Les résultats sont repris Table 3. Ceux-ci indique qu'il y a clairement une différence de localisation (pour tous les paramètres cités). La dispersion est plus importante pour les petites entreprises. Pour finir, la série des grandes entreprises est légèrement plus dissymétrique que celle des petites entreprises. Ces différentes observations sont également visibles sur les différents graphiques du point précédent.

Groupe	Localisation				Dispersion		Dissymétrie	
	moyenne	médiane	minimum	maximum	étendue	variance	Fisher	Y_k
Grandes	8.33	8.4	6.2	10	3.8	0.91	-0.36	7.35
Petites	6.08	6.1	3.7	8.5	4.8	1.61	-0.04	3.17

TABLE 3 – Résumés statistiques des deux groupes.

   (c) Comme indiqué Table 2, nous ne connaissons pas la taille de la ligne 32. Cependant, son indice de qualité vaut 8.4, ce qui est supérieur à la médiane du groupe `Grande` et très proche du maximum pour l'autre groupe. Il est donc très probable que cette observation corresponde à une grande entreprise.

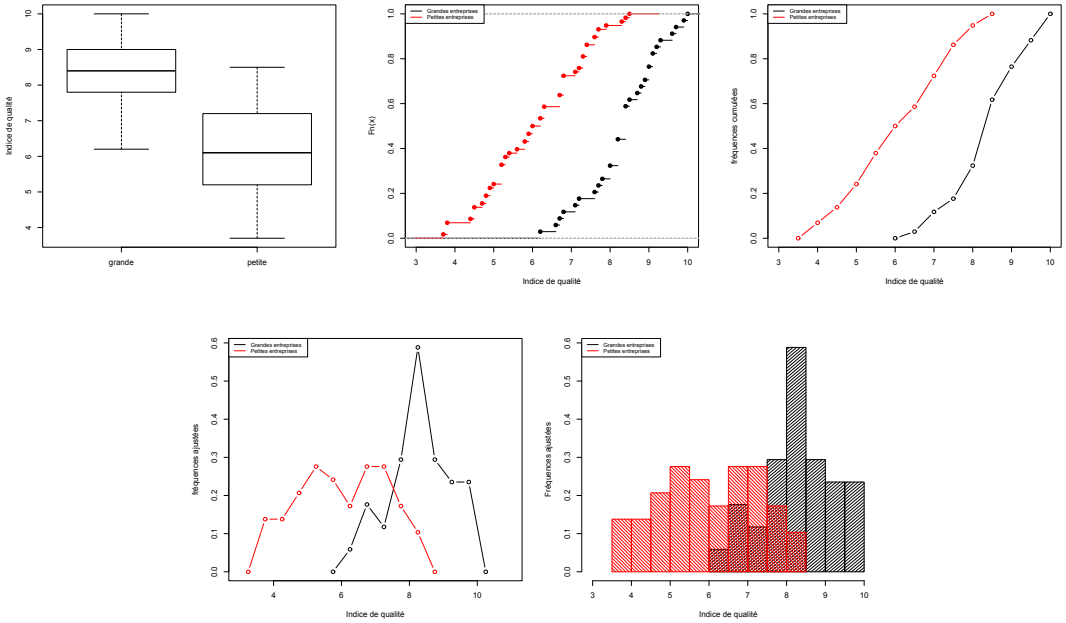

FIGURE 8 – Différents graphiques pour comparer l'indice de qualité en fonction de la taille des entreprises.

(d) La corrélation entre la variable `Qualite` et la variable binaire prenant une valeur 1 lorsque l'entreprise est `Grande` vaut 0.69. Cela signifie qu'il y a un lien assez important entre la taille de l'entreprise et son indice de qualité, l'indice de qualité augmente lorsque la taille de l'entreprise est plus grande. Cette observation est en adéquation avec les commentaires des points précédents.

3. (a) Le diagramme de dispersion de la variable `Service` expliquée par la variable `Prix` est repris Figure 9. La droite de régression estimée par la méthode des moindres carrés y est représentée et est d'équation

$$\text{Service} = 2.166 + 0.318\ \text{Prix}.$$

(b) Le graphique indexé des résidus est repris Figure 9. Celui-ci permet d'observer que la plupart des résidus sont assez faibles. En effet, seules deux observations ont un résidu qui est en valeur absolue plus grand que 2 fois l'écart-type (observations 32 et 88). Ces deux observations sont indiquées en rouge sur le diagramme de dispersion.

(c) Le coefficient de détermination permet de mesurer la qualité d'ajustement. Celui-ci est le coefficient de corrélation au carré, à savoir dans ce cas-ci, 0.26. Cela signifie que 26% de la variabilité de l'indice de Service peut être expliquée par la régression. Cette ajustement linéaire n'est donc pas très performant pour prédire l'indice de service.

(d) Dans l'ensemble de données de départ, l'indice de service de la ligne 38 est manquant. Celui-ci peut être estimé à l'aide de la régression mentionnée aux points

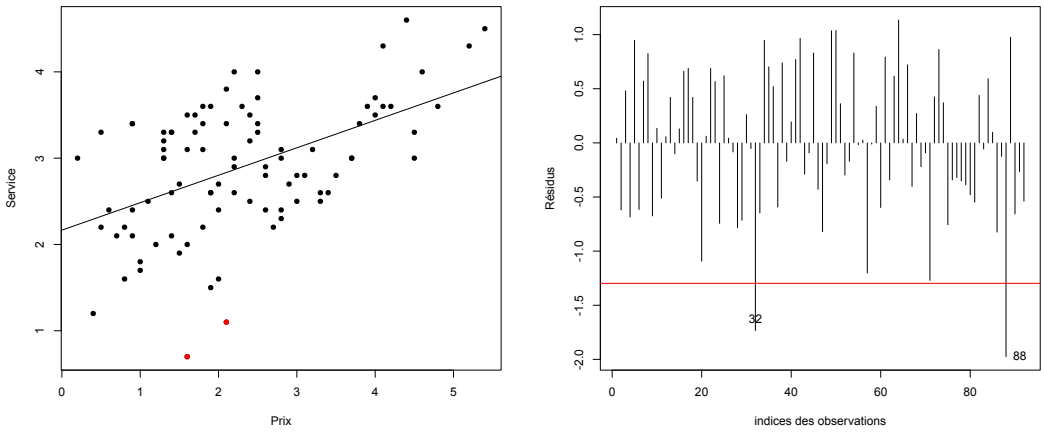

FIGURE 9 – Régression linéaire : diagramme de dispersion et graphique indexé des résidus.

précédents :

$$\hat{y}_{38} = 2.166 + 0.318 \times 2 = 2.802.$$

D'autre part, l'indice de service de la ligne 9 (erreur d'encodage) peut également être estimée à l'aide de son prix :

$$\hat{y}_9 = 2.166 + 0.318 \times 3.5 = 3.279.$$