Statistique descriptive

Année académique 2020–2021

Carole.Baum@uliege.be

TP6 : Analyse d'une base de données réelles

Description des données :

Le fichier de données (disponible sur eCampus) s'intitule `DataTP6Info.csv` et correspond à des indicateurs économiques mesurés sur 90 pays, ces indicateurs ayant vraisemblablement un lien avec un *indice de bien-être* calculé sur le pays. Les données proviennent du rapport *World Happiness Report* de 2020, rapport rédigé par le réseau *Sustainable Development Solutions Network* (voir le site https ://worldhappiness.report).

Plus précisément, les variables sont les suivantes :
`Generosite` : indicateur qualitatif à deux modalités caractérisant le niveau de générosité des habitants du pays (critère basé sur les montants versés à diverses associations caritatives). Les niveaux sont PeuG (pour peu généreux) et G (pour généreux).
`BienEtre` : indicateur de bien-être mesuré sur une échelle continue de 1 à 10.
`LogPIB` : logarithme du produit intérieur brut du pays (calculé en dollars).
`EspVie` : espérance de vie dans le pays, à la naissance (en années, avec décimales).
`Gini` : indice de Gini du pays (mesure comprise entre 0 et 1, décrivant les inégalités de revenu dans le pays ; plus la valeur est proche de 0, plus la répartition des revenus est égalitaire, plus l'indice approche 1, plus les inégalités sont fortes).

La colonne intitulée Pays reprend les noms des pays. Certaines données sont manquantes, indiquées par NA.

Remarque : Il est supposé dans la suite du correctif que les données ont été importées via les commandes

```
data <- read.table("DataTP6Info.csv", header=TRUE, row.names=1, sep=";")
data$Generosite<-as.factor(data$Generosite)
attach(data)
```

où l'option `header=TRUE` permet de spécifier que la première ligne du fichier contient les noms des variables, l'option `row.names=1` spécifie que la première colonne contient les noms des individus (dans ce cas-ci, les pays), et l'option `sep=";"` précise que le symbole qui sépare les colonnes dans le fichier initial est le point-virgule.

Exercice 1. Passer en revue les variables de la base de données et pour chacune, préciser l'intervalle des valeurs observées, dénombrer le nombre de valeurs manquantes et indiquer si des valeurs clairement aberrantes sont présentes (il faut dans ce cas, indiquer le pays correspondant, la valeur aberrante et préciser la justification de son caractère aberrant). Les valeurs aberrantes avérées doivent ensuite être remplacées [1] par `NA`.

Solution :

1. Attention : Les valeurs doivent être remplacées directement dans l'ensemble de données et ces données doivent ensuite être détachées et réattachées avant de répondre à la suite. Par exemple, si l'observation n° 1 de la variable `LogPIB` doit être remplacée, il faut effectuer les étapes suivantes :

```
data$LogPIB[1] ←NA
detach(data)
attach(data)
```

L'intervalle des valeurs observées ainsi que le nombre de valeurs manquantes pour toutes les variables, qu'elles soient quantitatives ou qualitatives, peuvent être obtenus directement via la commande `summary(data)`, dont l'output est le suivant.

```
> summary(data)
   BienEtre         LogPIB          EspVie       Generosite      Gini
Min.   :3.160   Min.   : 7.680   Min.   : 50.50   G   :38   Min.   :-0.210
1st Qu.:5.188   1st Qu.: 9.090   1st Qu.: 64.33   PeuG:52   1st Qu.: 0.310
Median :5.845   Median : 9.880   Median : 67.85             Median : 0.350
Mean   :5.842   Mean   : 9.751   Mean   : 67.52             Mean   : 0.369
3rd Qu.:6.497   3rd Qu.:10.565   3rd Qu.: 72.33             3rd Qu.: 0.410
Max.   :7.890   Max.   :11.320   Max.   :115.87             Max.   : 1.120
                NA's   :2                                   NA's   :1
```

La variable `Generosite` est une variable qualitative. Elle ne compte pas de donnée manquante, ni de modalité "additionnelle" encodée par erreur.

Pour les variables quantitatives, les valeurs minimales (`Min.`), maximales (`Max.`) ainsi que les nombres de données manquantes (`NA's`) sont précisés ci-dessus.

A partir de cet output, on obtient les informations suivantes :

— Deux variables sont complètes (pas de donnée manquante), à savoir `BienEtre` et `EspVie`. Les deux autres variables comptent une (`Gini`) et deux (`LogPIB`) observations manquantes.

— A partir des valeurs minimales et maximales observées, certains problèmes sautent aux yeux.

  — L'indice de Gini est, par définition, une mesure comprise entre 0 et 1. Il y a donc un souci des deux côtés de l'intervalle de définition observé (valeur minimale inférieure à 0 et valeur maximale supérieure à 1. Afin de déterminer les pays correspondant à des valeurs aberrantes, on peut utiliser les commandes

  ```
  row.names(data)[Gini<0]
  row.names(data)[Gini>1]
  ```

  qui permettent de déduire que la `Tanzanie` correspond à une valeur aberrante inférieure à 0, et que la `Slovénie` correspond à une valeur aberrante supérieure à 1[2].

  — Une espérance de vie de 115.87 ans est impossible à imaginer car elle est trop grande. En utilisant la commande `row.names(data)[EspVie==max(EspVie)]`, on se rend compte qu'il s'agit du `Danemark`.

  — Les intervalles de définition des deux autres variables ne sont pas suspects.

  En analysant la boîte à moustaches représentée à la Figure 1, obtenue via la commande `boxplot(EspVie, xlab='EspVie')`, on peut par ailleurs se convaincre que seule la valeur maximale de l'espérance de vie est très clairement erronée.

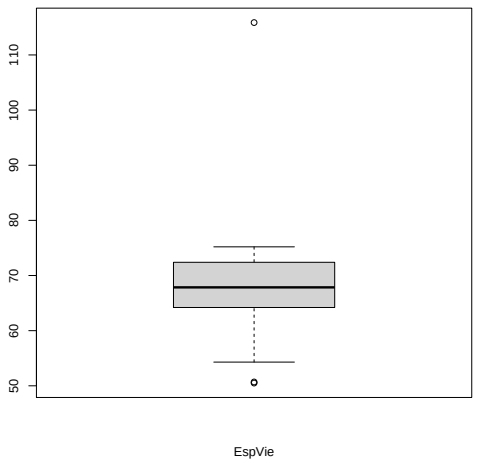

FIGURE 1 – Boîte à moustaches de la variable `EspVie`

2. Notez que, pour chacune de ces deux commandes, le logiciel renvoie comme première observation une valeur manquante. Celle-ci provient du fait que la variable `Gini` contient elle-même une valeur manquante. La valeur rendue par le booléen `Gini>0` est alors une valeur manquante, qui se retrouve dans le résultat final.

Afin de remplacer les valeurs aberrantes par des valeurs manquantes, on peut utiliser les commandes suivantes

```
data$Gini[c(72,79)] <- NA
data$EspVie[22] <- NA
detach(data)
attach(data)
```

car les pays `Danemark`, `Slovénie` et `Tanzanie` correspondent respectivement aux observations 22, 72 et 79.

Exercice 2. Les procédures implémentées dans `R` ne traitent pas toutes les données manquantes de la même manière, mais la plupart (exemple `mean`, `median`, `var`...) retourne `NA` dès qu'elles sont appliquées à des données contenant des valeurs manquantes. Des options (du type `na.rm=TRUE` ou `use= "complete.obs"`) permettent d'imposer que les calculs se fassent uniquement sur les données complètes. Cependant, lorsqu'il y a beaucoup de données manquantes (ce qui n'est pas le cas ici), éliminer les observations manquantes peut mener à une grosse perte d'information. Dans ce cas, il est classique d'essayer de leur donner une valeur plausible en utilisant les autres observations disponibles (c'est la technique dite d'imputation des valeurs manquantes).

Par exemple, on peut remplacer une donnée manquante par la moyenne (ou la médiane, le mode) de la variable, celle-ci étant calculée sur les données observées. Cela s'appelle la technique d'imputation non conditionnelle. Une autre technique habituelle consiste à remplacer une valeur manquante d'une variable Y par sa valeur ajustée par un modèle de régression linéaire basé sur une variable X potentiellement explicative pour Y : Il s'agit dans ce cas d'une technique d'imputation conditionnelle. En vous focalisant sur les données manquantes de la variable `LogPIB` mais sans modifier la base de données, déterminer quelles valeurs pourraient leur être imputées si les deux stratégies ci-dessus étaient exploitées. Lorsque des choix doivent être faits pour l'application de la technique d'imputation, il faut justifier l'option privilégiée.

Solution :

— *Imputation non conditionnelle :*
Les valeurs manquantes de la variable `LogPIB` peuvent être remplacées par la moyenne ou par la médiane calculées sur les observations disponibles, puisque cette variable est une variable quantative continue. Ces deux valeurs sont obtenues via les commandes

```
mean(LogPIB, na.rm=TRUE)
median(LogPIB, na.rm=TRUE)
```

Dans les deux cas, on constate que les valeurs imputées seraient assez similaires (9.75 en utilisant la moyenne et 9.88 en exploitant la médiane). L'écart ne serait que de 0.13 entre les deux valeurs (ce qui, sur une échelle logarithmique, n'est tout de même pas complètement négligeable).

— *Imputation conditionnelle :*
Afin de choisir de manière objective la variable explicative à utiliser dans ce modèle linéaire, une analyse préliminaire de corrélation et la visualisation d'une éventuelle présence d'un lien linéaire sont utiles.
La matrice de corrélations entre toutes les variables quantitatives peut être obtenue via la commande suivante.

```
cor(data[,c(1,2,3,5),], use="pairwise.complete.obs")
```

L'option `use="pairwise.complete.obs"` permet d'utiliser les observations complètes sur les deux variables considérées dans le calcul de la corrélation (par exemple, pour le calcul de la corrélation entre les variables `BienEtre` et `EspVie`, 89 observations sont utilisées alors que seules 87 observations sont utilisées dans le calcul de la corrélation entre `BienEtre` et `Gini`). Les corrélations entre la variable dépendante `LogPIB` et les autres variables quantitatives sont les suivantes : 0.8421 avec `BienEtre`, 0.8426 avec `EspVie` et -0.3998 avec `Gini`. La variable `EspVie` est donc celle qui a le lien linéaire le plus fort avec la variable `LogPIB`, lien que l'on peut en effet observer sur le diagramme de dispersion repris à la Figure 2 et obtenu via la commande `plot(EspVie, LogPIB)`.
C'est donc la variable `EspVie` qui va être privilégiée. Le modèle linéaire ajusté par la technique des moindres carrés est défini via la commande

```
reg <- lm(LogPIB ~ EspVie)
```

Afficher le résumé (voir Figure 3) permet ensuite de visualiser les valeurs des paramètres et de déterminer l'équation de la droite de régression, qui est donnée par

$$\texttt{LogPIB} = 0.67 + 0.14\ \texttt{EspVie}$$

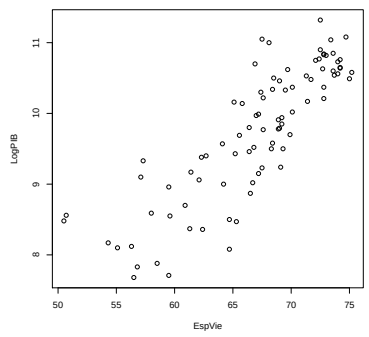

FIGURE 2 – Nuage de points de la variable `LogPIB` en fonction de la variable explicative `EspVie`

```
> summary(reg)

Call:
lm(formula = LogPIB ~ EspVie)

Residuals:
     Min       1Q   Median       3Q      Max
-1.35930 -0.25693 -0.02703  0.28661  1.23133

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.673088   0.630835   1.067    0.289
EspVie      0.135490   0.009393  14.424   <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5138 on 85 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.7099,    Adjusted R-squared:  0.7065
F-statistic: 208.1 on 1 and 85 DF,  p-value: < 2.2e-16
```

FIGURE 3 – Résumé de la régression de la variable `LogPIB` en fonction de la variable `EspVie`

Les pays ayant une valeur manquante pour la variable `LogPIB` sont `Chypre` et l'`Iran` et correspondent à des espérances de vie égales à 74.1 et 66.6 respectivement. Dès lors, les valeurs imputées seraient données par $0.67 + 0.14 \cdot 74.1 = 11.04$ et $0.67 + 0.14 \cdot 66.6 = 9.99$ respectivement.

La suite du correctif est basée sur la base de données de laquelle les données manquantes sont éliminées. Pour ce faire, on peut effectuer les opérations

```
data2 <- na.omit(data)
detach(data)
attach(data2)
```

Exercice 3. Analyse de la variable `Generosite`

1. Représenter la distribution des effectifs de la variable qualitative `Generosite` à l'aide d'un graphique adéquat et en déduire le mode de cette variable.
2. Certains économistes considèrent que le niveau de générosité d'un pays dépend de l'importance des inégalités de revenus observées dans le pays en question. Après avoir construit un indicateur binaire permettant de distinguer les pays ayant un indice de Gini inférieur à la médiane des indices de Gini et les pays ayant un indice supérieur ou égal à la médiane, trouver un moyen graphique ainsi qu'un résumé statistique permettant d'illustrer et quantifier la dépendance éventuelle de la variable `Generosite` sur l'indice de Gini (via sa version binarisée).

Solution :

1. La variable `Generosite` étant une variable qualitative, la distribution des effectifs peut être représentée à l'aide d'un diagramme en barres ou en secteurs. Ces deux diagrammes sont représentés aux Figures 4 et 5 respectivement et sont obtenus en utilisant les commandes

   ```
   effGen <- table(Generosite)
   pie(effGen)
   barplot(effGen)
   ```

   On en déduit que le mode est la modalité `PeuG`, ce mode ayant été observé sur 50 des 84 pays considérés, soit 60%.

2. L'indicateur binaire peut être défini de la façon suivante.

   ```
   GiniBin <- as.integer(Gini >= median(Gini, na.rm=TRUE))
   ```

   On peut alors commencer par construire la table de contingence entre les deux variables, via la commande `table(Generosite, GiniBin)`. Cependant, ici, on souhaite déterminer si la distribution de `Generosite` est globalement la même dans les deux groupes définis par l'indicatrice binaire. Ce sont donc les distributions conditionnelles de `Generosite` qui nous intéressent. Celles-ci peuvent se calculer "à la main" à partir du tableau de contingence mais elles peuvent aussi être obtenues via la commande

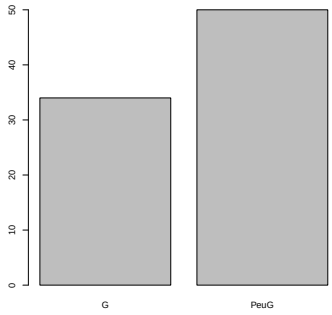

FIGURE 4 – Diagramme en barres construit sur la distribution des effectifs de la variable `Generosite`

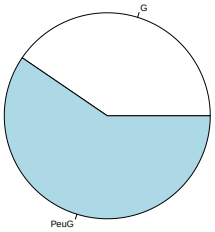

FIGURE 5 – Diagramme en secteurs construit sur la distribution des effectifs de la variable `Generosite`

```
prop.table(table(Generosite, GiniBin), margin=2)
```

où l'option `margin=2` spécifie que les fréquences doivent être calculée en colonnes. On obtient alors la table suivante :

```
> prop.table(table(Generosite, GiniBin), margin=2)
          GiniBin
Generosite         0         1
     G    0.4146341 0.3953488
     PeuG 0.5853659 0.6046512
```

qui nous indique que, parmi les pays ayant un indice de Gini inférieur à la médiane (valeur 0 de `GiniBin`), 59% sont peu généreux, tandis que parmi ceux dont l'indice de Gini est supérieur à la médiane, 60% ont cette caractéristique. On constate que les deux distributions sont assez similaires. On pourrait donc imaginer que, que les habitants de pays soient confrontés à de fortes inégalités ou non, ils ont tout autant tendance à participer à des oeuvres caritatives.

Exercice 4. Les économistes imaginent également un lien entre le Produit intérieur brut d'un pays et le fait qu'il soit ou pas généreux.

1. A l'aide de deux graphiques adéquats (représentant les deux distributions sur le même repère dans les deux cas), comparer les distributions de la variable `LogPIB` dans les deux groupes de la variable `Generosite`.
2. En se basant non seulement sur les graphiques représentés, mais également sur au moins un paramètre statistique de chaque type (position, dispersion et dissymétrie), déterminer si les distributions diffèrent plutôt en tendance centrale, en dispersion ou en forme (voire en plusieurs caractéristiques).
3. Déterminer la part de la variabilité de la variable `LogPIB` expliquée par la variabilité dans les groupes. Commenter en vous référant aux graphiques et analyses effectuées.

Solution :

1. Les distributions conditionnelles du produit intérieur brut (en logarithme) dans les deux groupes de la variable `Generosite` peuvent être comparées à l'aide de boîtes à moustaches et de polygones des fréquences. Ces deux graphiques sont repris à la Figure 6 et sont obtenus grâce aux commandes

```
boxplot(LogPIB ~ Generosite)
h1 <- hist(LogPIB[Generosite=="G"], freq=FALSE)
h2 <- hist(LogPIB[Generosite=="PeuG"], freq=FALSE)
plot(c(7.25,h1$mids,11.75), c(0,h1$density,0), type="o", pch=16, col=2)
lines(c(7.25,h2$mids,11.75), c(0,h2$density,0), type="o", pch=16, col=3)
```

On constate que les deux distributions se situent globalement dans le même intervalle de définition. On voit aussi que le produit intérieur brut des pays généreux présente une distribution légèrement plus étendue et "uniforme" (avec cependant une concentration un peu plus forte parmi les valeurs supérieures à la médiane). L'autre distribution a une structure plus symétrique (même si l'étalement à gauche est plus

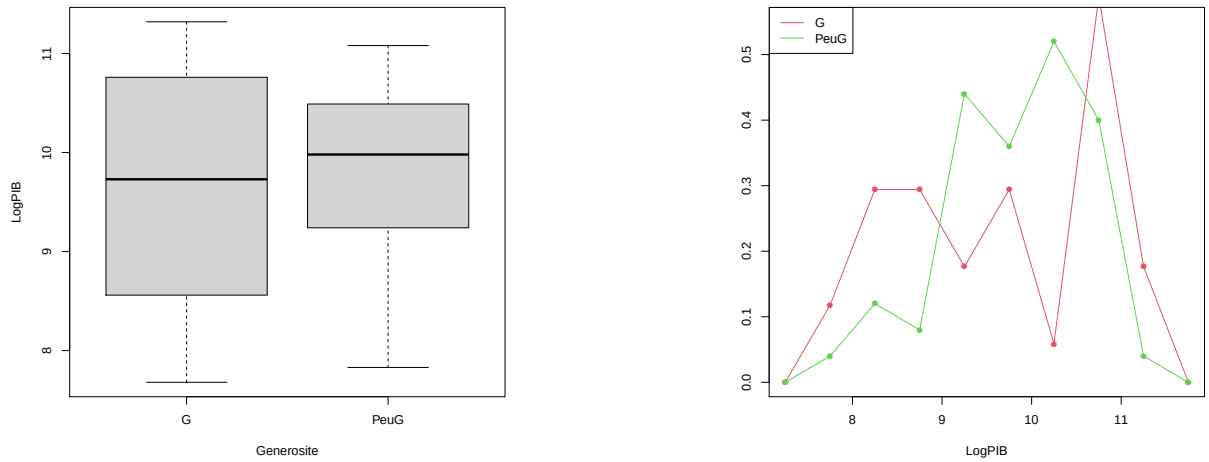

FIGURE 6 – Boîtes à moustaches et polygones des fréquences de la variable `LogPIB` en fonction des deux modalités de la variable `Generosite`

prononcé qu'à droite) et présente une plus forte concentration des valeurs entre 8.5 et 10.5. En ce qui concerne les tendances centrales, on constate, grâce aux boîtes à moustaches, qu'elles sont très proches.

2. Les statistiques reprises dans la Table 1 sont obtenues via les fonctions `mean, median, sd, IQR` et `skewness` (de la librairie `moments`), appliquées aux deux vecteurs `LogPIB[Generosite=="G"]` et `LogPIB[Generosite=="PeuG"]`. On peut confirmer que les deux distributions se valent en termes de tendance centrale (moyennes et médianes presque égales), tandis que la distribution correspondant aux pays généreux est en effet plus dispersée (ainsi qu'indiqué par l'écart-type et l'écart interquartile). En ce qui concerne la forme, les deux paramètres de dissymétrie sont négatifs (ce qui traduit une prédominance de petites observations ou étalement à gauche). De manière attendue vu la forme des polygones des fréquences, la dissymétrie est plus prononcée dans la distribution des pays peu généreux.

	Généreux		Peu généreux
Moyenne	9.66	$\approx$	9.84
Médiane	9.73	$\approx$	9.98
Ecart-Type	1.12	$>$	0.78
EIQ	2.19	$>$	1.22
Dissymétrie (Fisher)	-0.20		-0.72

TABLE 1 – Statistiques descriptives de la variable `LogPIB` en fonction des deux modalités de la variable `Generosite`

3. Étant donné le caractère très similaire des moyennes mis en évidence ci-dessus (les écarts entre les moyennes locales, égales à 9.66 et 9.84, et la moyenne globale de 9.77 sont négligeables), on s'attend bien à ce que la variabilité dans les groupes soit la source de variabilité la plus importante. Pour quantifier l'analyse, il faut d'abord calculer la variance globale de la variable `LogPIB` (elle vaut 0.86) ainsi que la variance dans les groupes donnée par

$$\frac{n_{\mathtt{G}} s^2_{\mathtt{G}} + n_{\mathtt{PG}} s^2_{\mathtt{PG}}}{n_{\mathtt{G}} + n_{\mathtt{PG}}} = \frac{34 \cdot 1.26 + 50 \cdot 0.6}{84} = 0.87$$

où l'indice `G` (resp. `PG`) correspond à la modalité `Genereux` (resp. `PeuGenereux`).

On constate cependant que l'on arrive à une absurdité car le terme correspondant à la variance dans les groupes est plus grand que la variance totale, ce qui est impossible mathématiquement.

La raison de cette incohérence est la suivante : `R` ne calcule pas la variance et l'écart-type selon les formules définies au cours, à savoir

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ et } s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

mais remplace les dénominateurs n par $n-1$ (pour une raison statistiquement pertinente, comme cela sera enseigné dans les cours suivants de statistique). Dans ce cas, la formule de décomposition n'est plus vérifiée. Pour corriger les variances, il suffit de les multiplier par $\frac{n_G-1}{n_G}$ et $\frac{n_{PG}-1}{n_{PG}}$ respectivement. Après avoir effectué cette opération, les variances deviennent 1.22 et 0.59. On a alors

$$\frac{n_G s_G^2 + n_{PG} s_{PG}^2}{n_G + n_{PG}} = \frac{34 \cdot 1.22 + 50 \cdot 0.59}{84} = 0.85$$

ce qui permet de dire que pratiquement 100% de la variabilité totale est expliquée par le caractère non homogène des pays dans les deux groupes.

Exercice 5. Pour mieux connaître la variable `LogPIB`,

1. Représenter cette variable à l'aide d'un histogramme adéquat (ne pas nécessairement se contenter du choix par défaut de R) et construire le tableau statistique correspondant à la découpe privilégiée.
2. Choisir un paramètre statistique permettant de quantifier la dissymétrie observée et commenter.
3. A partir de la série groupée, la moyenne de la distribution de `LogPIB` peut être approximée. Calculer l'écart entre la moyenne exacte et la moyenne approximée. Commenter en développant en détail le calcul de l'erreur commise par l'approximation dans une des classes à spécifier.

Solution :

1. Le nombre de classes suggéré par la formule de Sturges est 8 et le diagramme en tiges et feuilles de la variable, obtenu via la commande `stem(LogPIB)`, est représenté ci-dessous.

```
> stem(LogPIB)

  The decimal point is at the |

   7 | 778
   8 | 111244
   8 | 5566679
   9 | 0001122223444
   9 | 555566778888999
  10 | 0001222233344
  10 | 55555666666777888888899
  11 | 00113
```

   Fixer la première borne inférieure en 7.5 et la dernière borne supérieure en 11.5 semble idéal vu les valeurs minimale et maximale observées sur la variable. Construire des classes d'amplitude constante et égale à 0.5 mènerait à un total de 8 classes. Ce faisant, la première classe ne compterait que peu d'observations (3). Il serait donc pertinent de regrouper cette classe avec la suivante. Il pourrait ensuite être intéressant de découper l'intervalle $]9;11]$ en 5 classes en non 4. Cela mènerait alors à des amplitudes de classes de 0.2, 0.5 et 1. Il serait plus harmonieux de choisir des amplitudes de classes qui ont comme diviseur commun l'amplitude 0.2. L'histogramme est finalement obtenu via la commande

   ```
   h <- hist(LogPIB, breaks=c(7.6,8.4,9,9.4,9.8,10.2,10.6,11,11.4))
   ```

   et représenté à la Figure 7.
   Pour construire le tableau statistique correspondant à cette découpe, les effectifs sont obtenus grâce à `h$counts`, les effectifs cumulés grâce à la fonction `cumsum` et les fréquences et fréquences cumulées peuvent être facilement déduites. La Table 2 reprend l'ensemble de ces valeurs.

2. Les trois paramètres de dissymétrie vus au cours sont décrits ci-dessous :

Nom	Formule	Valeur
Fisher	$\gamma_1 = \frac{m_3}{s^3}$	-0.48
Pearson	$S_k = \frac{\bar{x}-\tilde{x}}{s}$	-0.12
Yule-Kendall	$Y_k = \frac{Q_1+Q_3-2\tilde{x}}{Q_3-Q_1}$	-0.04

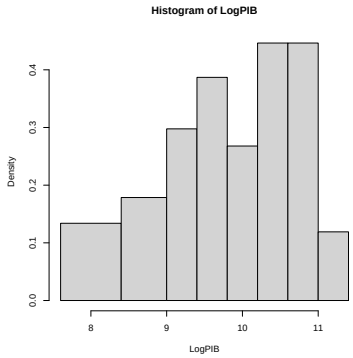

FIGURE 7 – Histogramme d'aire unitaire de la variable `LogPIB`

Classes	Effectifs	Eff. cum.	Fréquences	Fréq. cum.
[7.6, 8.4]	9	9	0.11	0.11
]8.4, 9]	9	18	0.11	0.22
]9, 9.4]	10	28	0.12	0.34
]9.4, 9.8]	13	41	0.15	0.49
]9.8, 10.2]	9	50	0.11	0.60
]10.2, 10.6]	15	65	0.18	0.77
]10.6, 11]	15	80	0.18	0.95
]11, 11.4]	4	84	0.05	1.00

TABLE 2 – Tableau statistique de la version groupée de `LogPIB`

Le coefficient de Fisher est disponible via la fonction `skewness` de la librairie `moments`. Les autres peuvent être calculés à partir des moyenne, médiane, 1er et 3ème quartiles (qui sont eux-mêmes obtenus grâce à la fonction `quantile`, ou via le `summary`).

Dans tous les cas, les valeurs obtenues sont négatives, ce qui confirme la présence d'un étalement sur la gauche. Les valeurs ne sont pas "comparables" de façon brute.

3. A partir de la série groupée reprise dans la Table 2 et en utilisant par exemple la commande `sum(h$counts*h$mids)` on obtient la moyenne approximée qui vaut 9.761, alors que la moyenne exacte vaut 9.766. On voit donc que l'approximation est de très grande qualité. Pour s'en convaincre, considérons les 9 observations de la première classe, que l'on retrouve en triant les observations de la variable `LogPIB` grâce à la fonction `sort`. Puisque le centre de la classe vaut 8, les erreurs commises sont données par

```
> sort(LogPIB)[1:9] - 8
[1] -0.32 -0.29 -0.17  0.08  0.10  0.12  0.17  0.36  0.37
```

On voit donc qu'il y a trois erreurs négatives et 6 erreurs positives. La somme des erreurs est donc égale à 0.42, valeur qui doit ensuite être "relativisée" par le diviseur habituel du nombre d'observations, à savoir 84. La contribution de la première classe à l'erreur totale est donc donnée par 0.005.

Exercice 6. Les économistes pensent que le niveau de bien-être dépend du produit intérieur brut dans le pays.

1. Commenter à propos de la présence éventuelle d'un lien linéaire entre ces deux variables et compléter votre analyse par le calcul du coefficient de corrélation.
2. Déterminer la droite de régression de la variable `BienEtre` en fonction de la variable `LogPIB` (préciser son équation) et représenter le diagramme de dispersion de la variable `BienEtre` en fonction de `LogPIB` en y ajoutant la droite de régression ajustée.
3. Représenter les résidus en fonction des valeurs observées de la variable explicative et calculer la variance des résidus. Commenter et déterminer quel pays est celui dont le résidu est le plus grand en valeur absolue.
4. En exploitant les réponses précédentes, commenter à propos de l'adéquation du modèle estimé et considérer les points suivants (pour chaque question, mettre au moins un élément concret - calcul statistique ou graphique par exemple - en évidence) :

- La connaissance des valeurs observées de la variable `Generosite` pourrait-elle changer l'étude du lien linéaire entre les variables `BienEtre` et `LogPIB` ?
- Aurait-il été plus opportun de sélectionner une autre variable parmi celles disponibles pour expliquer linéairement la variable `BienEtre` ?

Solution :

1. Le diagramme de dispersion (sur lequel est déjà superposée la droite de régression, voir question suivante) représenté à gauche sur la Figure 8 montre qu'il semble y avoir en effet une dépendance linéaire croissante entre ces variables. Le coefficient de corrélation vaut 0.83, ce qui est une valeur "proche" de 1. Le nuage de points et le coefficient de corrélation sont obtenus grâce aux commandes

   ```
   plot(LogPIB, BienEtre)
   cor(LogPIB, BienEtre)
   ```

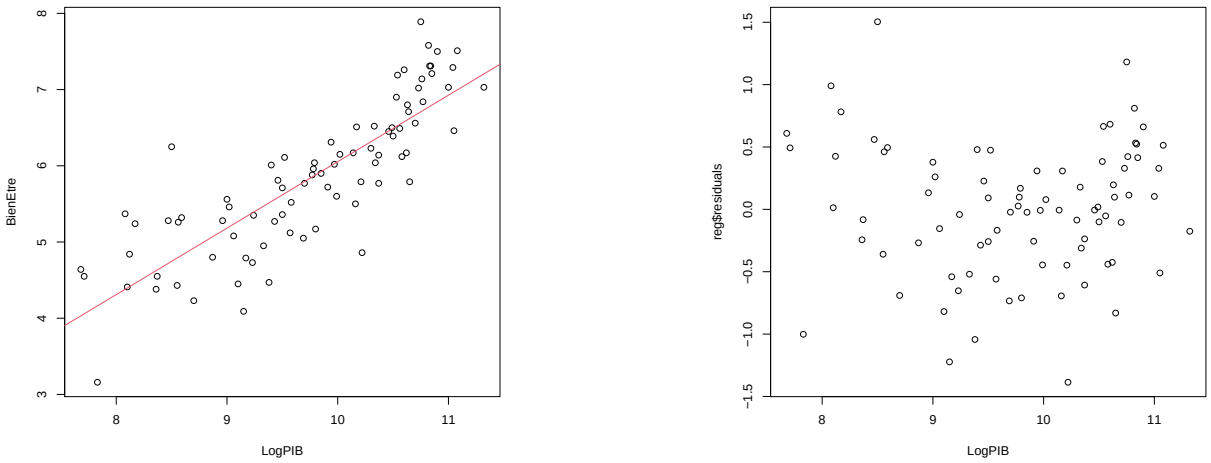

FIGURE 8 – A gauche : Diagramme de dispersion de la variable `BienEtre` en fonction de `LogPIB` sur lequel est superposée la droite de régression. A droite : Graphique des résidus en fonction des valeurs de `LogPIB`

2. Les commandes

   ```
   reg <- lm(BienEtre ~ LogPIB)
   summary(reg)
   ```

```
> summary(reg)

Call:
lm(formula = BienEtre ~ LogPIB)

Residuals:
     Min       1Q   Median       3Q      Max
-1.38597 -0.32284  0.00386  0.39139  1.50414

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.66747    0.62360  -4.278 5.08e-05 ***
LogPIB       0.87216    0.06357  13.719  < 2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5383 on 82 degrees of freedom
Multiple R-squared:  0.6965,    Adjusted R-squared:  0.6928
F-statistic: 188.2 on 1 and 82 DF,  p-value: < 2.2e-16
```

   permettent d'obtenir la droite de régression de la variable `BienEtre` en fonction de la variable `LogPIB` estimée par la technique des moindres carrés. Elle est donnée par

$$\texttt{BienEtre} = -2.67 + 0.87\ \texttt{LogPIB}.$$

   Cette droite peut être ajoutée au nuage de points précédent en utilisant la commande `abline(reg, col=2)` et elle est représentée sur le nuage de points discuté ci-dessus.

3. Le graphique des résidus en fonction de la variable `LogPIB` est obtenu via la commande `plot(LogPIB, reg$residuals)` et est visible à droite de la Figure 8. On constate que ce graphique ne présente aucune structure (ce qui est souhaité). On repère aussi le résidu le plus grand en valeur absolue (valeur supérieure à 1.5) et, en utilisant, par exemple, les commandes

   ```
   row.names(data)[which.max(abs(reg$residuals))]
   max(abs(reg$residuals))
   ```

   on constate qu'il s'agit du résidu de la `Jordanie` qui vaut 1.504. La variance des résidus est, quant à elle, égale à 0.29 (elle est obtenue grâce à la commande `var(reg$residuals)`).
4. La droite de régression s'ajuste raisonnablement bien au nuage de points mais la part de variance expliquée par la régression ne vaut que 70%. Cette part est donnée par le coefficient de détermination, le coefficient R^2, disponible directement dans le `summary` de la régression repris ci-dessus, ou via la commande `cor(LogPIB, BienEtre)^2`. On pouvait déjà conclure que la variabilité des résidus n'était pas négligeable à l'exercice précédent.
   — Le lien entre les deux variables est plus prononcé parmi les pays généreux que parmi les autres. Ceci s'exprime par des coefficients de corrélation différents dans les deux groupes : 0.88 d'un côté et 0.81 de l'autre. Ces valeurs sont obtenues via les commandes

   ```
   cor(LogPIB[Generosite=="G"], BienEtre[Generosite=="G"])
   cor(LogPIB[Generosite=="PeuG"], BienEtre[Generosite=="PeuG"]))
   ```

   Cependant, les droites estimées séparément dans les deux groupes sont "assez" similaires. Ces droites sont représentées à la Figure 9 et sont obtenues via les commandes

   ```
   regG <- lm(BienEtre[Generosite=="G"] ~ LogPIB[Generosite=="G"])
   regPeuG <- lm(BienEtre[Generosite=="PeuG"] ~ LogPIB[Generosite=="PeuG"])
   plot(LogPIB, BienEtre, col=as.integer(Generosite)+1)
   abline(regG, col=2)
   abline(regPeuG, col=3)
   legend("topleft", c("G","PeuG"), col=2:3, pch=16)
   ```

   Pour une même valeur en abscisse, la valeur attendue en ordonnée est sensiblement la même dans les deux groupes. En "moyenne", les analyses sont donc similaires, mais cette moyenne est plus représentative de l'ensemble des pays parmi les pays généreux que dans l'autre groupe, vu les variabilités différentes.

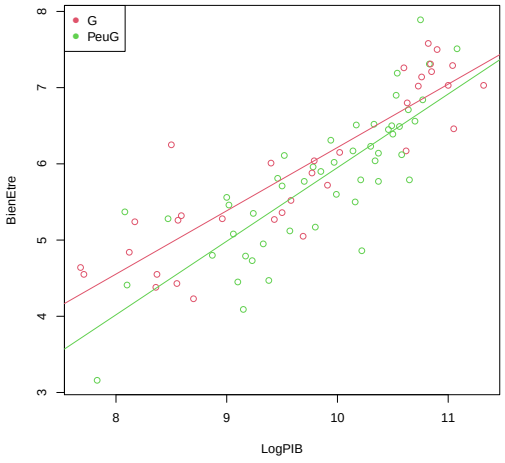

FIGURE 9 – Droites de régression des moindres carrés estimées sur les deux groupes définis par leur modalité de la variable `Generosite`

— Selon les valeurs du coefficient de corrélation, le lien linéaire le plus prononcé entre la variable `BienEtre` et une autre est celui basé sur la variable `LoGPIB` (la corrélation, vaut 0.83, tandis que les deux autres corrélations sont égales à 0.76 et -0.42). Pour rappel, ces corrélations peuvent être obtenues via la commande `cor(data2[,c(1,2,3,5),])`.