Scénario: Statistiques élémentaires d'une cohorte familiale

Résumé

Initiation à la pratique des techniques élémentaires de statistique par l'étude d'un jeu de données à l'aide du logiciel R. L'objectif est d'étudier quelques données épidémiologiques relatives à une famille autour de la naissance d'un enfant : description élémentaire, estimation, tests, régression linéaire, analyse de variance (ANOVA), analyse en composantes principales, régression multiple. Si nécessaire, un tutoriel de démarrage avec R est disponible.

1 Introduction

Une étude [1] réalisée entre 1961 et 1973 dans la maternité d'un hôpital d'Oakland (Californie) avait pour but de rechercher si certaines caractéristiques des parents avaient une influence sur le développement de l'enfant. Parmi les variables collectées, 19 variables décrites dans le tableau ci-dessous ont été observées sur 115 familles ou unités statistiques. Ces variables décrivent des informations médicales et socio-économiques concernant le bébé et ses parents au moment de la naissance puis dix ans plus tard. Ces données vont servir à illustrer la démarche classique d'une étude statistique.

Ces données permettent de se poser différentes questions de nature plutôt épidémiologique :

- Influence ou non de la consommation de cigarettes sur le sexe de l'enfant, sur son poids, sur sa taille,
- sur l'évolution du poids de la mère en 10 ans,
- sur les liaisons entre les caractéristiques des parents (poids, taille, rhésus) et celles de leur enfant,
- ...

1. J.L. Hodges, D. Krech et R. Crutchfield, *Statlab : an Empirical Introduction to Statistics*, 1975.

Code	Libellé	Unité ou modalités
ESx	sexe de l'enfant	M ou F
ERh	rhésus de l'enfant	Rh+ ou RH-
ET0	taille de l'enfant	à la naissance en cm
EP0	poids de l'enfant	à la naissance en kg
ET10	taille de l'enfant	à 10 ans en cm
EP10	poids de l'enfant	à 10 ans en kg
MRh	rhésus de la mère	Rh+ ou RH-
MA0	âge de la mère	à la naissance
MP0	poids de la mère	à la naissance
MCig0	consom. de cigarettes	0, 1 à 10, >10
MT	taille de la mère	
MP10	poids de la mère	10 ans après
MCig10	consommation de cigarettes	10 ans après
PA0	âge du père	à la naissance
PCig0	consommation de cigarettes	à la naissance
PT	taille du père	
PP10	poids du père	10 ans après
RF0	revenus familiaux	à la naissance
RF10	revenus	10 ans après

TABLE 1 – Statlab : liste des variables

2 Exploration statistique élémentaire

2.1 Lire les données

Les données sont disponibles dans le répertoire `http://wikistat.fr/data` sous la forme d'un fichier `statlab.csv` construit à partir de Excel en choisissant " ;" comme séparateur et "," comme marque décimale. Télécharger ce fichier dans le répertoire courant de R avant d'exécuter les commandes :

```
# Data frame à partir d'un fichier csv
famil=read.csv2("statlab.csv")
# vérification
summary(famil)
```

2.2 Description unidimensionnelle

Décrire chacune des variables en précisant ses caractéristiques :

Variable quantitatives

Décrire chaque variable (moyenne, écart-type, quantiles, diagramme boîte, histogramme) afin d'identifier les problèmes potentiels : valeurs atypiques, hétérogénéité des variances, distributions dissymétriques...

```
summary(famil)
sapply(famil, mean)   # moyennes
sapply(famil, sd)     # écarts-types
```

Les commandes suivantes s'intéressent à deux variables quantitatives, elles peuvent être appliquées à chacune d'elles.

```
boxplot(famil$ET0)
boxplot(famil$EP0)
hist(famil$ET0)
hist(famil$EP0)
```

Commenter les résultats obtenus en terme de symétrie des distributions, de présence de valeurs atypiques.

Variables qualitatives

Fréquences des modalités des variables qualitatives.

```
barplot(table(famil$ESx))
barplot(table(famil$MCig0))
pie(table(famil$MCig10))
```

2.3 Description bidimensionnelle

Variables quantitatives

Une matrice de nuages de points donne un aperçu rapide des structures de corrélation :

```
pairs(famil[,c(3:6,8,9,11,12,14,16:19)])
plot(EP10~PP10,data=famil)
plot(EP10~ET10,data=famil)
```

Variables qualitatives

Calcul de la table de contingence et graphe des profils colonnes dans un *mosaic plot*.

```
table(famil$ESx,famil$ERh)
# avec les marges
addmargins(table(famil$ESx,famil$ERh))
# fréquences relatives
prop.table(table(famil$ESx,famil$ERh))
mosaicplot(table(famil$ESx,famil$ERh))
addmargins(table(famil$MCig0,famil$ESx))
mosaicplot(table(famil$MCig0,famil$ESx))
```

Commenter la nature des liaisons entre certaines variables.

Variables qualitative et quantitative

Représenter une possible liaison entre les variables principales et celles qualitatives par des diagrammes boîtes.

```
boxplot(EP0~ESx,data=famil)
boxplot(EP0~MCig0,data=famil)
```

Commenter.

Bien d'autres options permettent de modifier les apparences des graphiques (titres, légendes...). Consulter l'aide en ligne si nécessaire.

3 Tests de comparaison

Important : Lors de l'exécution de chaque test préciser explicitement :

1. la question posée,
2. l'hypothèse H_0 en relation avec la question et associée au test,
3. la p-valeur calculée et la décision du test,
4. la réponse à la question.

3.1 Cas gaussien

Beaucoup des outils ci-dessous nécessitent de vérifier le caractère gaussien ou non de la distribution. En fait, le nombre important d'observations dans l'échantillon permet de s'affranchir de cette hypothèse mais il est utile de savoir la vérifier et éventuellement de sélectionner la transformation la plus appropriée des données notamment pour des variables de concentration.

Normalité d'une distribution : Shapiro-Wilks

La droite de Henri ou graphe quantile-quantile donne déjà un aperçu graphique de la normalité de la distribution avant de calculer le test.

```
# qq-plots
qqnorm(famil$EP0)
qqline(famil$EP0,col=2)
qqnorm(famil$ET0)
qqline(famil$ET0,col=2)
qqnorm(famil$ET10)
qqline(famil$ET10,col=2)
# Test de shapiro-Wilks
shapiro.test(famil$EP0)
shapiro.test(famil$ET0)
shapiro.test(famil$ET10)
```

Le test de Kolmogorov-Smirnov de comparaison à une distribution théorique pourrait également être utilisé (`ks.test`).

Intervalle de confiance d'une moyenne : Student

Il est important de savoir estimer l'intervalle de confiance d'une moyenne ; celui-ci permet de tester l'égalité de cette moyenne à une valeur théorique selon l'appartenance ou non de cette valeur à l'intervalle. L'effectif étant suffisamment grand, il n'est pas nécessaire de supposer la normalité des données. L'intervalle de confiance est calculé par défaut avec un seuil à 95% mais ce paramètre peut être précisé (`conf.level=.95`) de même que la moyenne théorique testée (`mu=0.0`, par défaut à 0).

```
t.test(famil$EP0, conf.level=.95)
```

Comparaison de deux variances : Fisher

On s'intéresse à l'influence du sexe sur la taille à la naissance. Tester l'égalité des deux moyennes nécessite de vérifier péralablement plusieurs points :

1. la normalité des distributions dans chaque classe à moins que l'échantillon soit considéré de taille suffisamment grande,
2. le caractère indépendant ou appariés des échantillons,
3. l'égalité ou non des variances à l'intérieure de chaque groupe.

On dispose de deux échantillons *indépendants* : les garçons et les filles. Testons les autres hypothèses.

```
# Normalité des distributions
# (facultatif car $n$ grand)
shapiro.test(famil[famil$ESx=="M","ET0"])
shapiro.test(famil[famil$ESx=="F","ET0"])
# égalité des variances (test de Fisher)
var.test(ET0~ESx,data=famil)
```

Commenter les résultats.

Comparaison de deux moyennes

Le test de comparaison des moyennes à utiliser (Student *vs.* Welsh) dépend du résultat précédent concernant l'égalité des variances.

Échantillons indépendants Si les variances sont différentes, il s'agit d'un test de Welch.

```
t.test(ET0~ESx,var.equal=F, data=famil)
```

Dans le cas où elles sont considérées égales, c'est un test de Student.

```
t.test(ET0~ESx,var.equal=T, data=famil)
```

Commenter.

Échantillons appariés On se propose d'étudier l'évolution du poids de la mère au moment de la naissance et dix ans après. La mesure étant observée

sur les mêmes personnes à deux instants différents, les échantillons sont cette fois appariés. Quelle "intuition" pouvons nous avoir du résultat à partir de graphique ci-dessous :

```
# nuage de points et première bissectrice
plot(famil$MP10~famil$MP0)
abline(a=0,b=1)
# test
t.test(famil$MP0, famil$MP10,paired=TRUE)
```

3.2 Cas non-paramétrique

Si l'hypothèse de normalité des distributions n'est pas vérifiée et si l'échantillon est trop réduit, c'est un test non-paramétrique qu'il faut mettre en œuvre. Les tests non-paramétriques sont basés sur les rangs des observations et donc sur les comparaisons des médianes des échantillons. Une transformation des variables par une fonction monotone (*i.e.* log) qui ne changent pas leur ordonnancement n'a donc pas d'effet sur le calcul d'un test non paramétrique.

Comparaison de deux médianes : Wilcoxon

Echantillons indépendants La même influence potentièle que le le test paramétrique est testée.

```
tapply(famil$ET0, famil$ESx, median)
wilcox.test(famil$ET0 ~ famil$ESx, data=famil)
```

Echantillons appariés Idem.

```
median(famil$MP0-famil$MP10)
wilcox.test(famil$MP0, famil$MP10,paired=TRUE)
```

Comparer avec les résultats des tests paramétriques.

4 Tests de liaison

4.1 Indépendance de 2 variables qualitatives

Le test du $\chi 2$ est adapté à ce problème.

```
chisq.test(table(famil$ESx,famil$ERh))
chisq.test(table(famil$ESx,famil$MCig0))
```

Remarque : un avertissement peut signaler que les effectifs théoriques (sous hypothèse d'indépendance) de certaines cellules sont trop faibles pour justifier des propriétés asymptotiques du test du $\chi 2$. Il est dans ce cas nécessaire de regrouper des modalités.

4.2 Une quantitative et une qualitative

L'ANOVA associée à un test de Fisher adapté à cette situation est sans doute le test le plus utilisé ; il revient au test de Student lorsque la variable qualitative n'a que deux modalités. L'ANOVA nécessite de vérifier :

1. le caractère indépendant des échantillons,
2. la normalité des distributions (ou une taille suffisante d'échantillon) dans chaque classe ou plutôt la normalité des résidus au modèle,
3. l'égalité des variances internes à chaque groupe.

Même si la normalité des résidus est vérifiée *a posteriori*, c'est *a priori* qu'il faut prendre en compte ce résultat pour statuer sur la légitimité du test.

Si la normalité n'est pas vérifiée pour un petit échantillon ou si l'égalité des variances n'est pas acceptable, un test non-paramétrique (Kruskal-Wallis) doit être envisagé.

Cas gaussien : ANOVA - Fisher

Deux tests permettent de comparer les variances des groupes. Le test de Bartlett dans le cas gaussien et celui de Levene si l'hypothèse de normalité n'est pas admissible. Le test de Bartlett est le plus utilisé compte tenu du contexte tandis que le test de Levene nécessite le chargement d'une autre librairie de R.

```
# test de Bartlett
bartlett.test(EP0~MCig0, data=famil)
# ANOVA à un facteur
res.anova=aov(EP0~MCig0, data=famil)
# normalité des résidus
qqnorm(res.anova$residuals)
```

```
qqline(res.anova$residuals)
shapiro.test(res.anova$residuals)
# Résulats du test
summary(res.anova)
```

Commenter.

Cas non-paramétrique : Kruskal-Wallis

```
kruskal.test(EP0~MCig0, data=famil)
```

Comparer les résultats et discuter de la validité de la décision à prendre.

Suivre la même démarche pour le taille de l'enfant à la naissance.

4.3 Deux variables quantitatives

La régression simple permet de tester l'influence éventuelle d'une variable sur une autre et plus précisément, dans le cas de cet exemple, d'expliquer et même de chercher à prévoir par exemple la taille de l'enfant à 10 ans. La commande `lm` produit un ensemble de résultats sous la forme d'une liste de matrices et vecteurs.

Estimation du modèle

```
res1.reg=lm(ET10 ~ PT, data = famil)
# liste des résultats
names(res1.reg)
```

Diagnostic des résidus

Des graphiques précédents permettent de s'assurer de la validité du modèle ; statuer sur l'homoscédasticité des résidus, leur normalité, la bonne linéarité du modèle.

```
# nuage de point
# normalité des résidus
qqnorm(res1.reg$residuals)
qqline(res1.reg$residuals)
shapiro.test(res1.reg$residuals)
# Repérage d'une structure particulière du nuage
# ou de la présence de "grands" résidus
res.student=rstudent(res1.reg)
ychap=res1.reg$fitted.values
plot(res.student~ychap,ylab="Résidus")
# ajouter des lignes
abline(h=c(-2,0,2),lty=c(2,1,2))
# repérage des points influents
cook=cooks.distance(res1.reg)
plot(cook~ychap,ylab="Distance de Cook")
abline(h=c(0,1),lty=c(1,2))
```

Les résidus sont "grands" si, une fois normalisés ou plutôt "studentisés", ils sont de valeur absolue plus grande que 2. Une observation est influente si elle a un grand résidu est est associée à une grande valeur sur la diagonale de la *hat matrix*. Cela correspond à une valeur élevée (plus grande que 1) de la distance de Cook.

Significativité du modèle

```
summary(res1.reg)
```

Que dire de l'influence de la taille du père ? Que dire également de la présence d'observations à effet levier potentiel ? Que dire de la qualité d'ajustement de ce modèle et donc de la qualité attendue de la prévision ?

Interpréter les tests.

Refaire ces calculs pour étudier le poids de l'enfant à la naissance en fonction de sa taille à la naissance et sa taille à 10 ans en fonction de celle à sa naissance.

5 ACP et régression multiple

5.1 Analyse en composantes principales

Cette description élémentaire permet de se familiariser avec la structure de corrélation particulière des variables. Il faut sélectionner les seules variables quantitatives et l'ACP est réduite.

```
# extraction des variables quantitatives
```

```
data=famil[,c(3:6,8,9,11,12,14,16:19)]
# liste des varaibles quantitatives
noms=dimnames(data)[[2]];noms
res.pca=prcomp(data,scale=T)
# décroissance des valeurs propres
plot(res.pca)
# Combien de dimension seraient à retenir ?
# parts de variance expliquée
summary(res.pca)
# biplot du premier plan principal
biplot(res.pca)
```

Comment s'interprètent les axes 1 et 2 ? D'autres commandes sont utilisées pour "colorier" les étiquettes selon la consommation de cigarettes.

```
plot(res.pca$x,col=as.integer(famil$MCig0))
text(10*res.pca$rotation,noms,col="blue")
abline(h=0,v=0,lty=2)
```

5.2 Régression multiple

Modèle linéaire complet

La régression linéaire simple conduit à un modèle très mal ajusté. Le modèle linéaire multiple ci-dessous, plus complexe, recherche un meilleur ajustement des données.

```
# estimation
res2.reg=lm(ET10~ET0+EP0+MA0+MP0+MT+MP10+PRA0+PT+
   PP10+RF0+RF10, data = famil)
# diagnostics
plot(res2.reg)
# résultats
summary(res2.reg)
```

Commenter les résultats obtenus sur la validité du modèle et la qualité de l'ajustement par rapport au modèle précédent. Que dire à propos de la significativité des tests de Student sur la nullité des paramètres ? Que penser alors de la présence de variables présentant de fortes colinéarités ?

Sous-modèle

Une procédure de sélection de modèle non détaillée (*stepwise*) conduit à considérer le modèle ci-dessous :

```
res3.reg=lm(ET10 ~ ET0+MT+PT, data = famil)
# diagnostics
plot(res3.reg)
# résultats
summary(res3.reg)
```

Commenter à nouveau les résultats.

Meilleure prévision

L'objectif est de rechercher le meilleur modèle de prévision de la taille de l'enfant à 10 ans. Ceux-ci sont comparés en considérant le PRESS (predicted residual sums of squares) ou *leave one out cross validation*. Une fonction élémentaire est définie pour calculer le PRESS dans le cas élémentaire de la régression linéaire.

```
# définition de la fonction PRESS
press=function(model) {
h=influence(model)$hat
e=influence(model)$wt.res
n=length(e)
sum((e/(1-h))^2)/n}
# application aux différents modèles
press(res1.reg)
press(res2.reg)
press(res3.reg)
```

Le meilleur modèle de prévision est-il celui qui ajuste le mieux les données ?

Attention, cete analyse se limite volontairement aux outils les plus élémentaires. D'autres modèles seraient à tester, notamment une analyse de covariance associant les variables qualitatives au modèle, la présence ou non d'interaction... pour tenter d'améliorer la qualité de prévision. C'est l'objet d'autres scénarios.