MSI – TP 3 : données suisses (19ème siècle)

CLOU Noémie
FONTAINE Pierre

Groupe 9
2014-2015

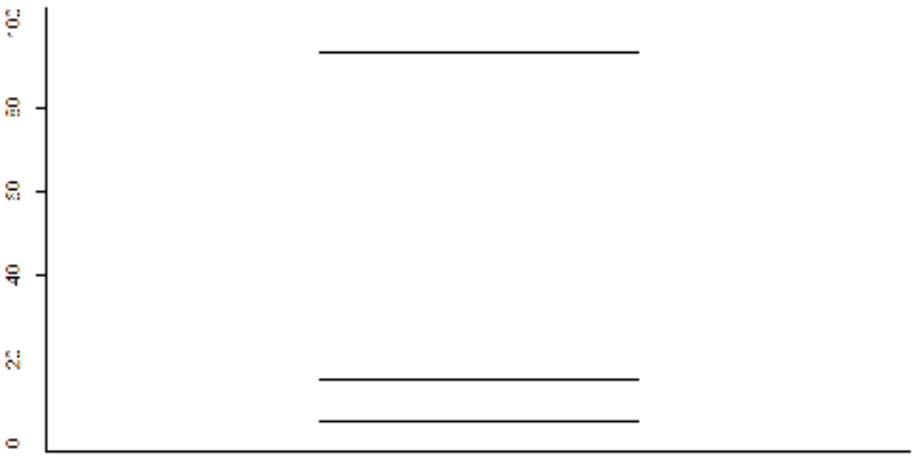
Question 1 :

On calcule la variance de chaque variable et on obtient :

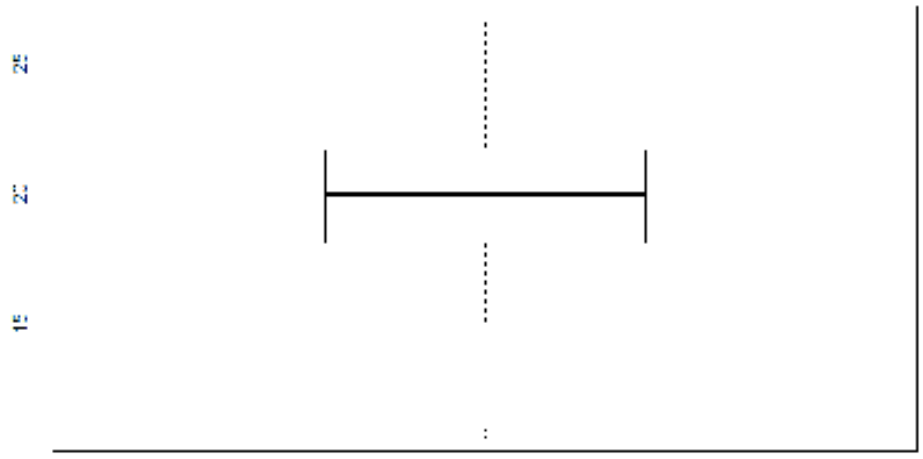
- Fertility : 156,04
- Examination : 63,65
- Education : 92,46
- Catholic : 1 739,30
- Infant Mortality : 8,48

La variable la plus dispersée est donc « Catholic » et la plus homogène est « Infant Mortality ».

Pour étayer notre réponse, on peut également dessiner des boîtes à moustaches et mesurer l'intervalle interquartile :
boxplot (donnees$Catholic)



boxplot (donnee$Infant.Mortality)



IQR (Catholic) : 87,93
IQR (Infant.Mortality) : 3,55

Question 2 :

On créé une variable « Protestant » : *Protestant = c(100-Catholic)*
On calcule ensuite sa moyenne : *mean (Protestant)* = 58,86
On vérifie notre réponse en calculant la moyenne de « Catholic » (41,14) et on remarque qu'elle est bien complémentaire de la moyenne de « Protestant ».

Question 3 :

On créé une nouvelle variable qui regroupe les provinces où au moins 4/5 (80 %) de la population est catholique : *Q3=which (donnees$Catholic > 80)*
On calcule ensuite la moyenne des agriculteurs : *mean (donnees [Q3,2])* = 65,52

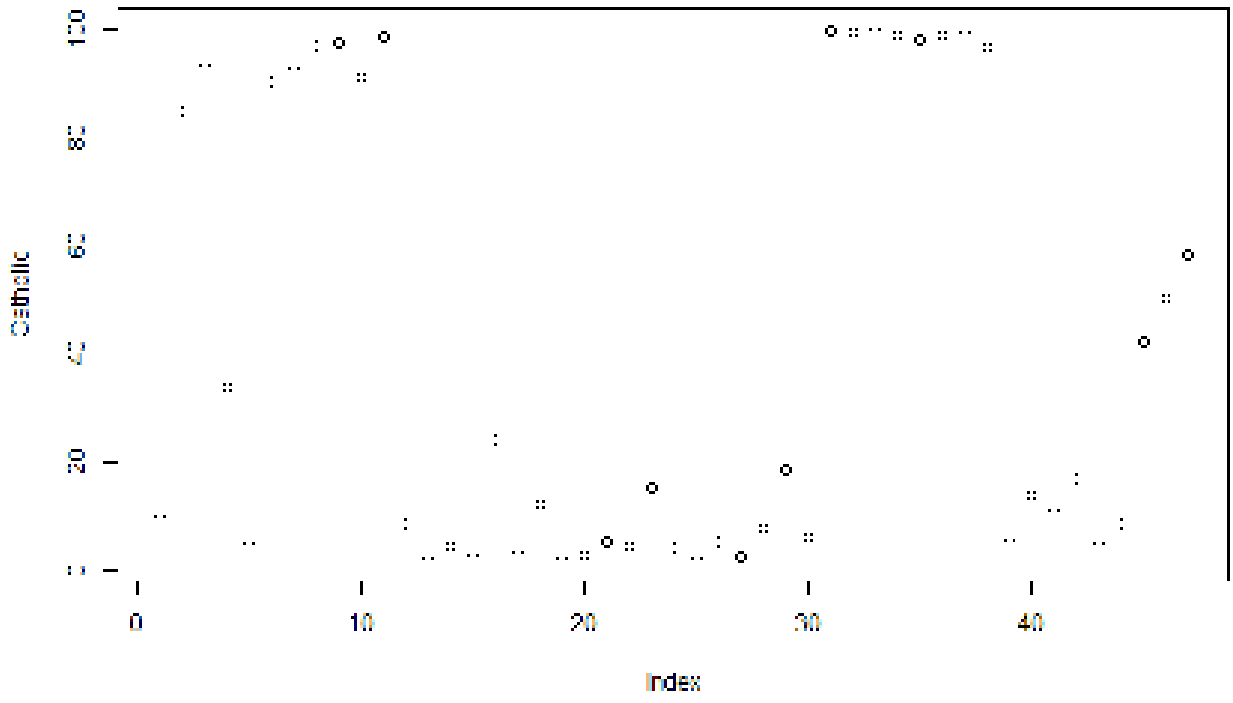
Pour déterminer le nombre de provinces concernées, on calcule la longueur du nouveau vecteur :
length (Q3) = 16

Pour comparer cette valeur à la moyenne de l'échantillon, on calcule le rapport des moyennes :
mean (donnees [Q3,2]) / mean (Agriculture) = 65,52/50,66 = 1,29

On remarque donc qu'il y a 1,29 fois plus d'agriculteurs dans les provinces où plus de 80 % de la population est catholique, que dans l'ensemble des provinces francophones.

Question 4 :

Pour traiter cette question, on dessine graphiquement le pourcentage de catholiques selon les provinces et on y ajoute deux lignes horizontales à 40 et 60 % :
plot (Catholic)
abline (h=40)
abline (h=60)



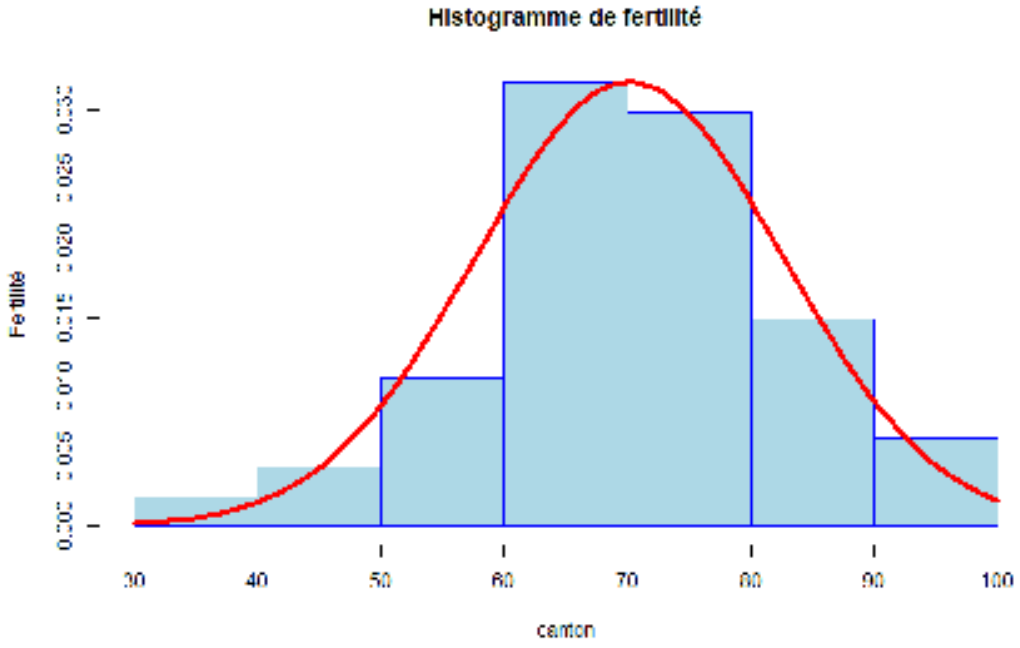
On remarque que les pourcentages sont très faibles (inférieurs à 20 %) ou très élevés (supérieurs à 80 %), ce qui témoigne d'une absence de brassage des religions dans une même province. Seules trois provinces ont des proportions de catholiques et de protestants équitablement réparties, comme le montrent les

calculs suivants, permettant de déterminer le nombre de provinces où on a entre 40 et 60 % de catholiques :
Q4 = which (40 < donnees$Catholic & donnees$Catholic > 60)
length (Q4) = 3

Question 5 :

Avec l'aide du code réalisé dans le TP 2, on écrit le code suivant :
xbar = mean (donnees$Fertility)
sigmabar = sd (donnees$Fertility)
hist (donnees$Fertility, freq = FALSE, border = "blue", col = "lightblue", xlab = "canton", ylab = "Fertility", main = "Histogramme de fertilité")
curve (dnorm (x, xbar, sigmabar), add = TRUE, col = "red", lwd = 4)

On obtient alors l'histogramme de densité demandé :

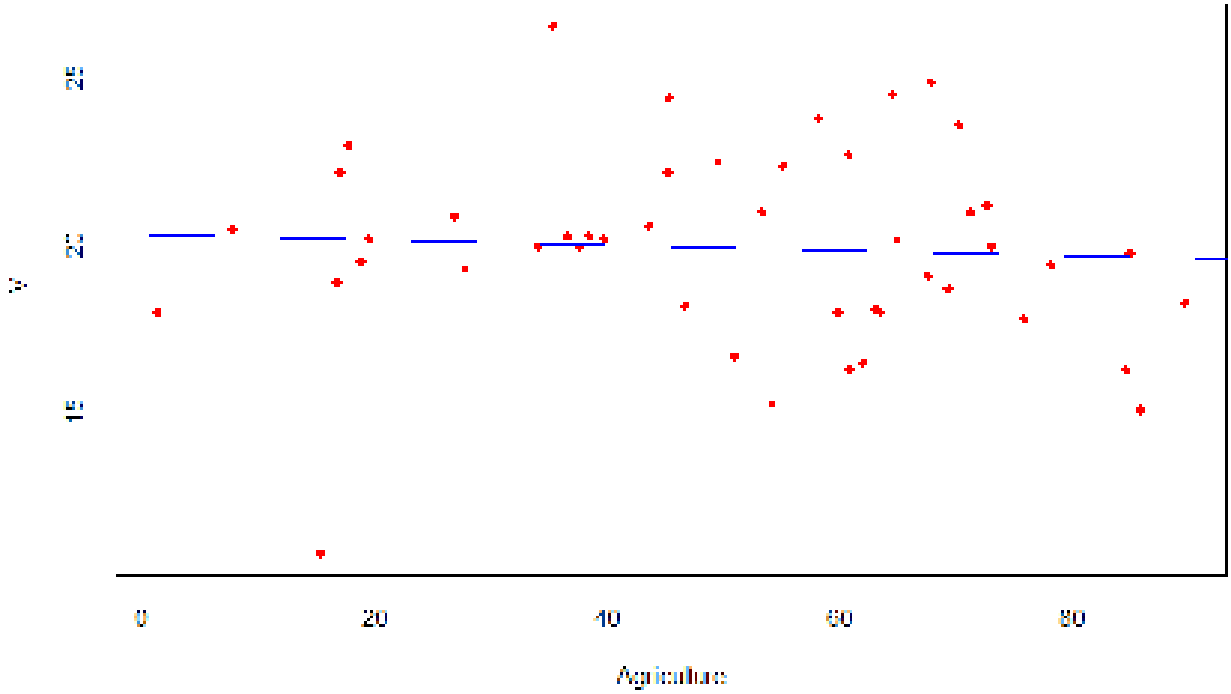


Question 6 :

Nous utilisons la fonction qui ne conserve qu'une seule fois chaque valeur : *length (unique (Examination))*
On a alors 22 variables différents sur l'ensemble de l'échantillon pour la variable « Examination ».

Question 7 :

Pour cette question, on trace le pourcentage d'agriculteurs en fonction du pourcentage de mortalité infantile, ainsi que la droite de corrélation entre ces deux variables :
plot (Agriculutre, Infant.Mortality, pch = 18, col = "red")
abline (lm (Infant.Mortality ~ Agriculture), col = "blue")
cor (Infant.Mortality, Agriculture) = -0,06



Le nuage de points observé est plutôt centré sur la droite en 0 et 40 % d'agriculteurs, alors qu'il est très dispersé au-delà. Cela montre également que le lien entre agriculteurs et mortalité infantile est inexistant.

Le coefficient de corrélation entre le taux de mortalité infantile et le taux d'agriculteurs étant légèrement négatif, la droite de corrélation est faiblement décroissante, ce qui traduit une légère diminution du taux de mortalité infantile avec l'augmentation du pourcentage d'agriculteur dans la province. Cependant, ce coefficient étant très faible, le lien entre les deux variables est quasiment nul.

Question 8 :

On calcule d'abord la moyenne (*moy8* = 19,94), la variance (*var8* = 8,48) et la longueur (*n* = 47) de la variable « Infant Mortality ».

Pour calculer les bornes de l'intervalle de confiance à 95 %, on utilise la loi de Student :
ICinf = moy8 + qt (0.025,47)(var8 / sqrt(n))* = 17,45
ICsup = moy8 - qt (0.025,47)(var8 / sqrt(n))* = 22,43

On fait de même pour l'intervalle de confiance à 90 % et on obtient un intervalle entre 17,86 et 22,02. On observe que cet intervalle se resserre très peu.

On calcule ensuite l'intervalle de confiance à 95 % en supposant que la variable « Infant Mortality » suit une loi normale :
*ICinfnorm = moy8 + qnorm (0.025, mean = moy8, sd = 3)*sqrt (moy8*(100-moy8)/n)*

Cela ne fonctionne pas, mais nous n'avons pas eu le temps d'y réfléchir plus longtemps…

Question 9 :

On détermine d'abord la proportion de provinces où le taux de mortalité infantile est supérieur à 20 % en calculant la longueur d'une nouvelle variable : *Q9 = which (Infant.Mortality > 20)*
On trouve alors *length (Q9)* = 23.