Exercice pour le Mooc Fondamentaux en statistique

Avner Bar-Hen

Vous avez à votre disposition un jeu de données disponible dans un fichier Excel.

Lorsque vous avez terminé l'exercice vous déposez votre/vos documents dans l'espace réservé aux dépôts des productions (un par semaine).

C'est également à partir de cet espace de dépôt que vous pourrez analyser et évaluer les productions de quelques pairs.

La date limite de dépôt est fixé au **mercredi 29 janvier 2014. à 16h**

Ce fichier va vous suivre pendant l'ensemble de ce cours. À l'issue des cinq semaines, vous aurez réalisé une analyse de données de votre tableau de données

Cette semaine vous avez le corrigé de la semaine 1 et l'énoncé de la semaine 2
Bon courage et en avant !

Semaine 1 : Comment résumer l'information d'une variable

Énoncé de l'exercice

En quelques lignes vous décrirez le fichier de données **Geraud.xls** (nombre d'observations, nombre de variables, nombre de modalités). Pour chacune des variables vous calculerez les statistiques descriptives et vous proposerez une représentation graphique.

Merci de ne pas dépasser une page pour le rendu de cet exercice.

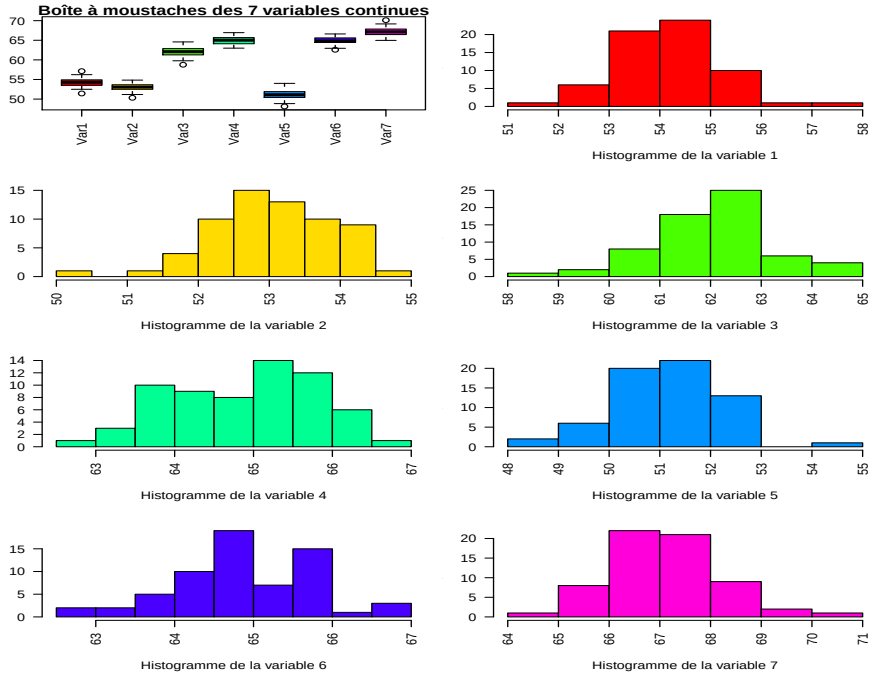
Solution

Le fichier Excel Geraud.xls contient 64 observations et 7 variables continues (dénommées Var1,. . .,Var7).
Le tableau ci-dessous donne les statistiques descriptives pour chacune des 7 variables continues.

	Moyenne	Médiane	Variance	Écart-Type	1er Quartile	3ème quartile	Min	Max
Var1	54.23	54.30	0.99	0.99	53.60	54.91	51.43	57.13
Var2	53.04	53.07	0.73	0.85	52.51	53.65	50.33	54.84
Var3	62.05	62.09	1.34	1.16	61.28	62.87	58.76	64.58
Var4	64.92	65.03	0.87	0.93	64.14	65.67	62.99	66.94
Var5	51.15	51.11	1.18	1.09	50.41	51.90	48.13	54.01
Var6	64.89	64.81	0.73	0.85	64.44	65.60	62.58	66.61
Var7	67.15	67.15	1.08	1.04	66.47	67.85	64.97	70.12

TABLE 1 – Statistiques descriptives des variables continues

On peut donc représenter les histogrammes pour voir la distribution des différentes variables et les boîtes à moustaches pour comparer les caractéristiques de chacune des variables.



On note que les moyennes et les médianes sont proches pour chacune des variables. On rappelle que l'égalité de la moyenne et de la médiane est une indication de la symétrie de la distribution des données de la variable étudiée.

Le tableau ci-dessous donne la liste des observations en dehors des moustaches.

Variable	Individu
Var1	Ind8, Ind19
Var2	Ind30
Var3	Ind40
Var5	Ind55
Var6	Ind6
Var7	Ind40

À première vue il ne semble pas y avoir de données aberrantes.

Les variables (Var4, Var6) ont des moyennes proches.

Les variances des 7 variables sont entre 0.73 et 1.34.

Semaine 2 : Comment résumer l'information d'une variable

Le but de l'exercice de cette semaine est de comprendre les résultats d'une analyse en composantes principales de votre fichier de données. Comme nous n'avons ni le volume horaire, ni les outils techniques pour inclure un enseignement logciel dans ce cours, je vous propose un ensemble de sorties que vous devrez analyser.

Énoncé de l'exercice

Dans un premier temps, vous calculerez les corrélations entre les variables (que vous pourrez compléter par des graphiques).

Si vous utilisez Excel, la fonction COEFFICIENT.CORRELATION devrait vous aider mais vous êtes libre d'utiliser le logiciel que vous souhaitez.

Les résultats devront être commentés.

Cette partie ne devra pas faire plus d'une page.

Ci-dessous vous avez les sorties d'une analyse en composantes principales.

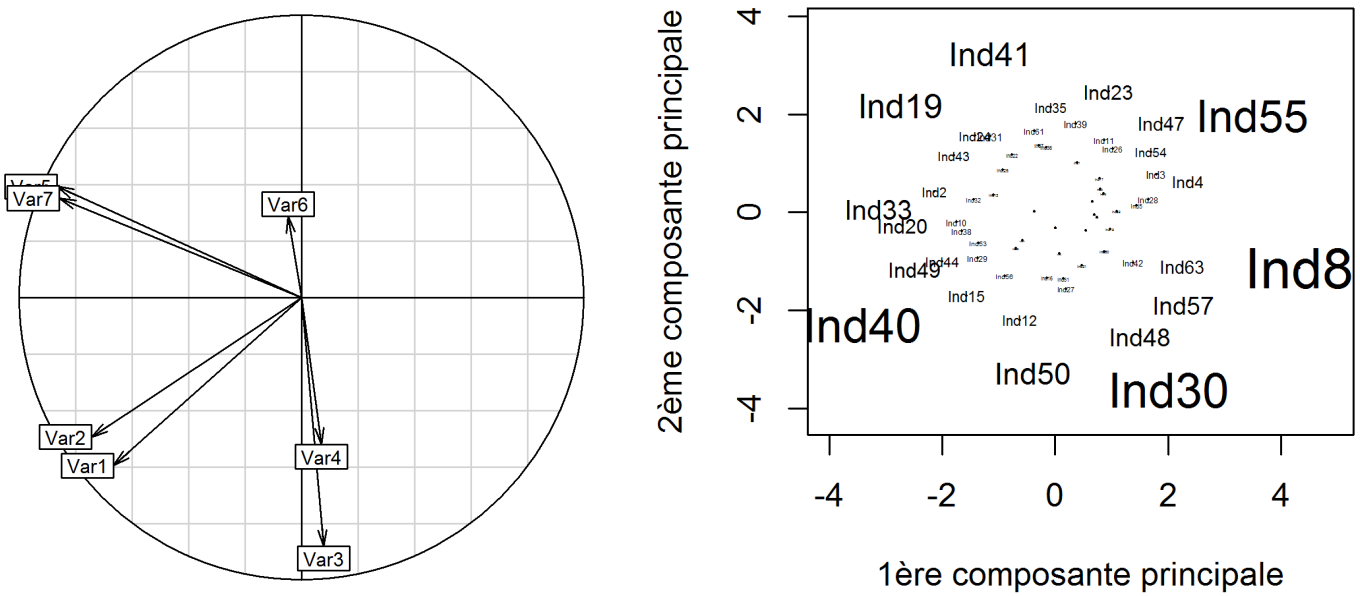
Les coordonnées des individus dans le plan (1,2) de l'ACP est dans la feuille "ACP" du fichier Geraud.xls

```
Importance of components:
                          PC1   PC2   PC3    PC4    PC5    PC6    PC7
Standard deviation     1.582 1.414 1.223 0.7095 0.4661 0.4221 0.3206
Proportion of Variance 0.357 0.286 0.214 0.0720 0.0310 0.0250 0.0150
Cumulative Proportion  0.357 0.643 0.857 0.9290 0.9600 0.9850 1.0000


       PC1   PC2   PC3   PC4   PC5   PC6   PC7
Var1 -0.42  0.42 -0.21  0.33 -0.02  0.66  0.24
Var2 -0.47  0.35 -0.09 -0.43  0.61 -0.29 -0.06
Var3  0.05  0.62 -0.26  0.21 -0.46 -0.50 -0.19
Var4  0.04  0.37  0.60 -0.55 -0.36  0.26 -0.04
Var5 -0.55 -0.28  0.12  0.12 -0.20  0.06 -0.74
Var6 -0.03 -0.20 -0.69 -0.58 -0.34  0.17  0.00
Var7 -0.54 -0.25  0.18  0.01 -0.36 -0.36  0.59
```

Corrélation entre les variables d'origine et les composantes principales

```
       PC1   PC2   PC3   PC4   PC5   PC6   PC7
Var1 -0.67  0.59 -0.25  0.23 -0.01  0.28  0.08
Var2 -0.74  0.49 -0.11 -0.31  0.28 -0.12 -0.02
Var3  0.08  0.88 -0.32  0.15 -0.22 -0.21 -0.06
Var4  0.07  0.52  0.73 -0.39 -0.17  0.11 -0.01
Var5 -0.87 -0.39  0.14  0.09 -0.10  0.03 -0.24
Var6 -0.05 -0.29 -0.85 -0.41 -0.16  0.07  0.00
Var7 -0.86 -0.35  0.22  0.00 -0.17 -0.15  0.19
```



Dans le graphique de ci-dessus, la taille des individus est proportionnelle à la qualité de la représentation.

Combien d'axes doit-on garder ?

Dans le plan (1,2), les individus (Ind33,Ind20) sont-ils proches ?

Dans le plan (1,2), les individus (Ind64,Ind14) sont-ils proches ?

Que peut-on dire des individus (Ind55,Ind40) dans le plan (1,2) ?

Quelle est la variable la mieux représentée sur l'axe 1 ?

Quelle est la variable la moins bien représentée sur l'axe 2 ?

Quelle est la variable la mieux représentée dans le plan (1,2) ?

À partir du cercle de corrélation, que peut-on dire du lien entre les variables dans le plan (1,2) ?

Question bonus : Détailler les calculs permettant d'obtenir les coordonnées de l'individu 8 ?

Merci de ne pas dépasser deux pages pour l'ensemble du travail de la semaine.