L3 STAPS **Université de Paris X - Nanterre**
Statistique et Informatique **Année 2006-2007**

TD n° 2
Analyse Univariée

0 Information sur l'examen

Lors de l'examen final, vous aurez le droit d'utiliser une fiche résumée comportant toutes les formules **R** que vous souhaitez. Commencer la rédaction de cette fiche dès à présent et noter les formules au fur et à mesure que vous les rencontrez aidera beaucoup à la rédaction de cette fiche... Enfin, nous, on dit ça, on dit rien !

Rappel : toutes les réponses doivent être notées dans un fichier Word.

1 La coupe du Comminges

1.1 Data.frame

1. Un fichier nommé *CoupeDuComminges.csv* se trouve sur le site Internet. Enregistrez-le dans le répertoire que vous avez créé dans *Mes Documents*
2. Importez ce fichier sous **R** et stockez le dans la variable `donnees` . Rappel : il y a trois choses à faire 1) sauvegarder le fichier dans un répertoire (vous venez de le faire) ; 2) dans **R**, changer le répertoire courant ; 3) lire le fichier avec la fonction `read.csv2()` . Si ça ne marche pas, c'est OBLIGATOIREMENT qu'une de ces étapes n'a pas été respectées. Enfin, si le fichier a une tête bizarre, c'est souvent parce que vous utilisez la fonction `read.csv()` au lieu de `read.csv2()` .
3. Après avoir ouvert un fichier contenant un data.frame et l'avoir stocké, la première chose à faire est de vérifier son contenu. Pour cela, on utilise la commande `summary()` . Tapez `summary(donnees)` . Qu'obtenez-vous ?
4. Déterminez (sans **R**) la liste des variables et leur type (à noter dans votre fichier Word)

De son coté, **R** a automatiquement déterminé la nature des variables. Pour connaître ses choix, on peut utiliser la fonction `str()` . `str(donnees)` donne la liste des colonnes de la data.frame en précisant leur type. Les types possibles sont :
– `int` (int pour *integer*) est une variable quantitative.
– `num` (num pour *numeric*) est une variable continue.
– `Factor` désigne une variable nominale.
– `Ord.Factor` est une variable ordonnée.
On remarque que **R** se trompe pour la variable `resultat` : en effet, celle-ci est ordonnée et **R** la classe nominale. On peut rectifier cela grâce à la fonction `ordered()` . Nous verrons cela lors de l'analyse univariée des variables ordonnées.

1.2 Variables discrètes

Quelle est la variable discrète de la data.frame ? Pour cette variable :

1. Dressez le tableau des effectifs (en utilisant la fonction `table()`)
2. Y a-t-il des valeurs aberrantes ? Si oui, remplacez les par `NA` (NA signifie *Non Available*).
   Rappel : pour accéder à une valeur dans un data.frame, on donne le numéro de la ligne, puis le numéro de la colonne entre crochet ; par exemple, `donnees[2,3]` est la valeur de la deuxième ligne et troisième colonne. Pour remplacer une valeur, il suffit de donner une nouvelle valeur à cette case. Par exemple, `donnees[2,3] <- NA` place NA dans la deuxième ligne et troisième colonne.
3. S'il y avait des valeurs aberrantes, recommencez le tableau des effectifs.

4. Calculez le nombre de valeurs manquantes (avec la fonction `summary()` : soit `summary(donnees)` , soit `summary(donnees[,i])` *i* étant le numéro de la colonne qui vous intéresse.)
5. Calculez le nombre total d'observations non manquantes. La fonction `nrow(donnees)` vous donne le nombre total de lignes de la data.frame. Pour avoir le nombre d'observations non manquantes, soustrayez les manquantes au total.
6. Calculez la moyenne, la médiane et les quartiles. Tous ces indices sont donnés par la fonction `summary()` .
7. Le calcul de la variance se fait grâce à la fonction `var()` . Tapez `var(donnees$carton)` . Que se passe-t-il ?
8. La fonction `var` ne fonctionne pas sur une variable dans laquelle il y a des valeurs manquantes. Il faut donc commencer par effacer les valeurs manquantes, au moins le temps de calculer la variance. Pour cela, tapez `na.omit(donnees$carton)` . Qu'obtenez-vous ? Calculez ensuite la variance de la variable carton.
9. De la même manière, calculez l'écart type à l'aide de la fonction `sd()` .
10. Tracez la boite à moustache de la variable carton. Pour cela, utilisez la fonction `boxplot(donnees$carton)` . Cette boite à moustache présente-t-elle un intérêt ?
11. Tracez de diagramme en bâtons de la variable carton. Pour cela, utilisez la fonction `barplot()` en lui donnant comme argument le tableau des effectifs de la variable carton.

1.3 Variables continues

Quelle est la variable continue de la data.frame ? Pour cette variable :

1. Doit-on dresser le tableau des effectifs ?
2. Y a-t-il des valeurs aberrantes ? Pour le savoir, on regarde le minimum et le maximum (fonction `summary()` . Si le maximum est trop grand ou le minimum trop petit (négatif par exemple), éliminez les valeurs aberrantes en les remplaçant par des `NA` .
3. S'il y avait des valeurs aberrantes, recommencez le tableau des effectifs.
4. Calculez le nombre de valeurs manquantes (avec la fonction `summary()`)
5. Calculez le nombre total d'observations non manquantes (fonction `nrow()`).
6. Calculez la moyenne, la médiane, l'écart type et les quartiles (`summary()` , `sd()` et éventuellement `na.omit()`).
7. Tracer la boite à moustache (`boxplot()`)
8. Tracer l'histogramme de la variable. Il est donné par la fonction `hist()` .

1.4 Variables nominales

Quelle est la variable nominale de la data.frame ? Pour cette variable :

1. Dressez le tableau des effectifs (en utilisant la fonction `table()`)
2. Y a-t-il des valeurs aberrantes ? Si oui, remplacez les par `NA`
3. S'il y avait des valeurs aberrantes, recommencez le tableau des effectifs.
4. Calculez le nombre de valeurs manquantes.
5. Calculez le nombre total d'observations non manquantes.
6. Calculez le mode (en vous basant sur les effectifs).
7. Tracer le diagramme en bâton des effectifs. Pour cela, utilisez la fonction `barplot()` et donnez lui comme argument les effectifs de la variable.

A ce stade, vous devriez avoir éliminé l'observation *illisible* (si ça n'est pas le cas, faites le !). Cette modalité est pourtant présente sur le diagramme, ce qui nous pose un problème. A quoi cela est-il du ?
En fait, **R** se souvient qu'initialement, *illisible* était une modalité de la variable. Dans son diagramme, il réserve donc un espace pour cette modalité, qui pourtant n'existe plus.
Pour corriger cela, il faut modifier la variable `donnees$meteo` en donnant à **R** la liste des modalités. Cela se fait avec la fonction `factor()`

1. Tapez `factor(donnees$meteo)` .
2. Tapez `table(factor(donnees$meteo))` puis `table(donnees$meteo)` . Conclusion ? Comment obtenir un diagramme à bâton sans la modalité *illisible* ?

Malheureusement, la manipulation que nous venons de faire n'a pas modifié le data.frame. En effet, si vous tapez `barplot(table(donnee$meteo))` , vous obtenez toujours un diagramme avec la modalité *illisible*. Pour que la correction soit définitive, il faut remplacer la colonne `donnees$meteo` par une nouvelle colonne, celle qu'on obtient en utilisant `factor()` (remplacement qui se fait, comme toujours, grâce au symbole d'affectation `<-`).

1. Tapez `donnees$meteo <- factor(donnees$meteo)`
2. Vérifiez que le data.frame est maintenant correct en traçant le diagramme : `barplot(table(dn$meteo))`

1.5 Variables ordonnées

Quelles sont les variables ordonnées de la data.frame ? Pour ce qui suit, on choisit *resultat*.

1. Dressez le tableau des effectifs.
2. Y a-t-il des valeurs aberrantes ? Si oui, traitez-les.
3. S'il y avait des valeurs aberrantes, recommencez le tableau des effectifs.
4. Calculez le nombre de valeurs manquantes.
5. Calculez le nombre total d'observations non manquantes.
6. Calculez le mode (en vous basant sur les effectifs).
7. Tracer le diagramme en bâton des effectifs. Pour cela, utilisez la fonction `barplot()` et donnez lui comme argument les effectifs de la variable.

La représentation graphique met un problème en évidence : **R** ne sait pas que la variable est ordonnée et il affiche les modalités de cette variable dans un ordre quelconque (en fait, il choisit l'ordre alphabétique). Pour corriger cela, il faut spécifier à **R** que la variable *resultat* est ordonnée. Cela se fait grâce à la fonction `ordered()` .

Cette fonction prend deux arguments. Le premier est le nom de la variable à ordonner, dans notre cas `donnees$resultat` . Le deuxième est la liste des modalités DANS l'ordre, de la plus petite à la plus grande, précédé par `levels=` . Dans notre cas, `levels=c("Perdu","Nul","Gagné","Gagné (bonus)")` . Au final, l'instruction pour ordonner notre variable est
`ordered(donnees$resultat,levels=c("Perdu","Nul","Gagné","Gagné (bonus)"))` .

1. Tapez la ligne de commande `ordered(donnees$resultat,c(levels=c("Perdu","Nul","Gagné","Gagné (bonus)"))` .
2. Vérifiez (grâce à `str(donnees)`) le type de la variable *resultat*. A-t-il changé ? En vous basant sur ce que vous avez fait au paragraphe précédent (problème de la modification de la variable meteo), comment faire pour qu'il change ?
3. Pour qu'il change, il faut modifier la colonne `donnees$resultat` dans le data.frame. Pour cela, on utilise le symbole d'affectation `<-` . Au final, on doit donc taper :
   `donnees$resultat <- ordered(donnees$resultat,levels=c("Perdu","Nul","Gagné","Gagné (bonus)"))`
4. Vérifiez (grâce à `str(donnees)`) le type de la variable *resultat*. A-t-il changé ?
5. Tracer le diagramme en bâton de la variable ordonnée.

2 Enquête ESPAD99

L'enquête ESPAD99[1] est un projet Européen visant à étudier la santé des jeunes et leur consommation de drogue (en collaboration avec l'OFDT[2]. L'enquête a été menée auprès de 12000 élèves (de la classe de 4^e à la fin des études secondaires) sur lesquels 564 variables ont été mesurées. L'équipe INSERM chargée de mener l'enquête en France a ajouté au questionnaire drogues deux questionnaires. Le premier porte sur la pratique sportive. Le second étudie la violence. L'idée est de voir s'il y a des liens entre la pratique sportive et la consommation de drogue et /

[1] European School Survey Project on Alcohol and other Drugs
[2] Observatoire Français des Drogues et Toxicomanie

ou les conduites violentes.

Votre tâche est de partir sur les traces des chercheurs INSERM et d'essayer de trouver des liens qui existeraient entre la pratique sportive, la consommation de drogue et les conduites violentes. Comme vous le savez (parce que ça a été dit en amphi), toute analyse statistique commence par une analyse univariée. Les données ESPAD99 sont dans le fichier *miniESPAD99.csv*. Nous n'avons retenu que 10 des 564 variables. A vous d'en faire l'analyse univarié.

Ceci n'est pas un exercice construit pour des étudiants,
c'est un cas réel avec des vrais gens et des vraies données.
Je répète : ceci n'est pas un exercice...

Cette analyse univariée doit être présentée sous Word. Choisissez 4 variables, une de chaque type. Pour chacune, vous devez faire l'analyse univariée la plus complète possible, à savoir :
- Définir son type
- Selon son type, présenter les effectifs, les données aberrantes, manquante, le nombre d'individu.
- Calculer les indices de centralité
- Selon son type, calculez les indices de dispersion
- Représenter graphiquement la variable.
- Éventuellement, quand le cas se présente, dire si la variable suit une loi particulière.