Statistique et Informatique — UFR STAPS
Licence 3 — Université de Paris X - Nanterre

TD n°11 : Exemple réel

Les étudiants de L2 ont mené une grande enquête "Fraude aux examens" auprès de 314 étudiants de l'université de Paris X. Le but de ce TD est d'analyser les données qu'ils ont collecté.

1 Data Management

Les bases de données réelles sont rarements "propres". La première étape d'une étude est donc de "nettoyer" les données.

1. Charger en mémoire le ficher `L2-Stat-ReponsesFraude2007.csv`. Stockez le dans `donnees`.
2. Á l'aide d'un `summary()`, examinez rapidement les variables. Plus précisément, examinez les variables `Age`, `Sexe`, `NiveauDEtude` et `Filière` Y a-t-il des problèmes ? Des valeurs aberrantes ?
3. Combien y a-t-il dans l'étude ?

La variable `[SEXE]` n'a pas été codée correctement : certains enquêteurs l'ont noté `(Homme)`, d'autres `(H)`, pareil pour `(Femme)` et `(F)`. Résultat, **R** nous donne 4 modalitées séparées alors qu'à l'évidence, il n'y en a que deux. Il va donc falloir corriger cela.

4. Fermez **R** sans sauvegarder. Ouvrez le fichier `L2-Stat-ReponsesFraude2007.csv` sous Excel ; remplacez tous les `H` par des `Homme` et le `F` pas des `Femme`.
5. En utilisant le *filtre automatique* ou le *tri*, trouvez la valeur aberrente du `niveauDEtude`. Remplacez-là par un `NA`
6. Que peut-on faire pour corriger la variable `Filière` ?
7. Souvegardez le fichier en concervant le format .csv (malgré les alertes d'Excel qui est fortement raciste et n'aime que le .xls...)
8. Ré ouvrez le fichier sous **R**.
9. Á l'aide d'un `summary()`, examinez rapidement les variables déjà éxaminé ci-dessus.

Un autres type de problème n'est réglable que sous **R**...

10. Quel est le type de la variable `Copier` ?
11. Tracez l'histogramme de cette variable. Constatez-vous un problème ?

La variable `Copier` est ordonnée mais **R** ne s'en est pas rendu compte. Du coup, il affiche les modalité de `Copier` dans n'importe quel ordre. Il faut donc lui indiquer que `Copier` est ordonnée.

12. Pour dire à **R** qu'une variable est ordonnée, on utilise la fonction `ordered` et on lui spécifie l'ordre des modalités grâce à `levels`. Quelles sont les modalités de `Copier` ? Quel vous semble être l'ordre croissant pour ces modalités (de la plus petite à la plus grande).
13. Transformez `donnee$Copier` en une variable ordonnée en utilisant l'instruction

```
donnees$Copier <- ordered(donnees$Copier,levels=c("Jamais","Rarement","Parfois","Souvent","Toujours"))
```

14. Recommancez pour toutes les autres variables de triche (de `Copier` à `VolerSujet`).

2 Analyse univarié

2.1 Analyse complète

Dans une étude réelle, on doit faire l'analyse univarié de toutes les variables. Ici, nous nous limiterons à quelques unes.

15. Faites l'analyse univarité de la variable `Copier`
16. Faites l'analyse univarité de la variable `UFR`
17. Faites l'analyse univarité de la variable `ScoreTricheTotal`

2.2 Représentation graphique

On souhaite comparer les différents type de triche. Pour cela, on veut représenter les histogrammes de triche côte à côte.

18. Pour spécifier à **R** que l'on veut plusieurs graphiques sur la même feuille, on utilise l'instruction `par(mfrow=c(a,b))` ou a est le nombre de lignes graphiques verticales et b est le nombre de colonnes graphiques. Par exemple `par( mfrow=c(2,3) )` permet d'afficher 6 graphiques, 3 sur une lignes, 3 autres en dessous. Pour vérifier, tapez `par(mfrow(2,3))` puis tapez 6 fois `barplot(c(1,2))` .
19. Dans notre cas, nous voulons afficher 10 graphiques côte à côte, 5 sur une ligne, 5 sur la suivante. Tapez la bonne instruction.
20. Afficher l'histogramme de la variable `donnees$Copier`
21. Afficher les histogrammes des 9 autres variable de mesure de triche.

3 Analyse bivarié

3.1 Représentation graphique

R permet de tracer des boites a moustaches côte à côte. C'est l'outil idéal pour comparer des variables... de quel type déjà ?

22. Pour quel type de variable utilise-t-on des boites à moustache ?

On souhaite par exemple représenter graphiquement l'âge en fonction de l'UFR d'origine. Pour cela, on utilise le symbole ~ qui signifie "en fonction de". Par exemple, `donnees$Age~donnees$UFR` signifie "Age en fonction de l'UFR" ou encore "Age classé par UFR".

23. Tapez `boxplot(donnees$Age~donnees$UFR)` .
24. Représentez graphiquement le ScoreTricheTotal en fonction de l'UFR.
25. Qui triche le plus ? Qui triche le moins ? La différence observée est-elle significative ?
26. Qui des hommes ou des femmes triche le plus ? (essayez de deviner avant, puis représentez graphiquement)

3.2 Tests statistiques

27. Y a-t-il un lien entre les variables `Sexe` et `TricheBac` ?
28. Y a-t-il un lien entre les variables `Sexe` et `ScoreTotalTriche` ?

4 Examen final

Les questions **2.1 Analyse complète** et **3.2 Test Statitiques** sont des questions d'examen.
Pour les questions **2.1**, n'oubliez pas les 4 étapes de l'analyse univarié.
Pour les questions **3.2**, n'oubliez pas les 3 étapes :

1. Determination du type de variable
2. Diagnotic (pour choisir entre paramétrique ou non paramétrique)
3. Faire le test(c'est à dire les 5 étapes : 1° H_0, 2° collecte, 3° calcul...)