TD 2. Préparation des données

1 Données

Nous allons travailler sur l'enquête **EPO2009 Fraude à l'université**. Pour cela, chargez le fichier *Fraude2009.csv* et stoquez le dans le data.frame `donnees` (Rappel : pour cela, enregistrez le ficher dans un repertoire, changez le répertoire courant puis utilisez l'instruction `read.csv2` , comme vous l'avez fait au TD1).
Si le fichier contient des valeurs manquantes (comme c'est le cas dans tous les fichiers réels), il faut le préciser à R. Cela se fait avec l'option `na.string=""` .

1. Chargez le fichez le fichier et stoquez-le grâve à l'instruction

```
donnees <- read.csv2("Fraude2009.csv",na.string="")
```

2 Type de variable

Chaque colonne d'un data.frame correspond à une variable. Chaque colonne a donc un type. Les différents type de variables statistiques correspondent aux types R suivant :

En statistique	Sous R
Nominale	`factor`
Ordonnée	`ordered`
Discrète	`numeric` (ou `integer`)
Continue	`numeric` (ou `integer`)

Quand R charge un ficher en mémoire (dans `donnees`), il donne a chaque variable un type. Pour connaître le type d'une variable, on utilise `str` . Cela liste toutes les variables avec leur type, leurs modalités et les premières observations.

2. En utilisant l'instruction `str(donnees)` , affichez le type des variables. Choisissez une variable de chaque type.

Dans certain cas, R se trompe : il n'a aucun moyen d'identifier les variables ordonnées (il les prend pour des `factor`) car il ne connait pas la relation d'ordre qui s'applique. C'est par exemple le cas de la variable `[Copier]`. Nous allons donc devoir corriger ses choix.

2.1 Ordonner une variable

La transformation d'une variable nominale en ordonnée se fait grâce à l'insctruction `ordered` . On doit également préciser la relation d'ordre grâce à `levels` . Ainsi, l'instruction
`donnees$Copier <- ordered(donnee$Copier,levels=c("Jamais","Rarement","Parfois",`
`"Souvent","Toujours"))` indique à R que la colonne Copier est une variable ordonnée et que l'ordre des modalité est Jamais < Rarement < Parfoit < Souvent < Toujours

3. Ordonnez une variable que vous choisissez parmi : Copier, Communiquer, EchangeBrouillon, QuestionProf, Antiseche, SMS, Internet, Cours, GarderCopie, RetardExam, PreparerSalle, VolerSujet, Autres
4. Ordonnez les variables NiveauDEtude et MentionBac
5. Á l'aide de `str` , vérifiez le type des colonnes de donnees.

2.2 Transformer en nominale

R se trompe également pour la variable `Identifiant` : il pense que c'est une variable numérique (`int`) alors qu'en fait, c'est une nominale (l'identifiant est un numéro unique qui désigne un patient, impossible de faire des opérations dessus). Pour corriger cela, on utilise la fonction `as.factor`

6. Grâce à l'instruction `donnees$Identifiant <- as.factor(donnees$Identifiant)` , corrigez le type de Identifiant.
7. Vérifiez que la correction est correcte.

3 Analyse univariée

Nos variables sont maintenant prêtes, l'analyse univariée peut commencer. L'instruction `summary` a pour effet de calculer automatiquement une partie de cette analyse en l'adaptant au type de variable : effectifs pour les `factor` et les `ordered` , moyenne et quartiles pour les `numeric` :

8. Utilisez `summary(donnees)` pour avoir un résumé global des données.
9. Combien y a-t-il d'homme dans l'étude ? Quelle est la moyenne d'age ?

Cela permet de jeter un premier oeil sur nos variables. Des instructions plus spécifiques permettent une analyse plus précise.

3.1 Effectifs

Les effectifs se calculent pour les variables nominales, ordonnées et discrètes. Cela se fait grâce à l'instruction `table` :

10. Choisissez une variable nominale. Grâce à l'instruction `table(donnees$MaVariable)` , dressez le tableau des effectifs.
11. Même question pour une variable ordonnée.

3.2 Centralité

3.2.1 Mode

Le mode s'obtient par lecture de la table des effectifs, en prenant tout simplement le plus grand.

3.2.2 Médiane

La médiane se calcule avec `median` . Quand la variable contient des valeurs manquantes, il faut préciser à R de les supprimer en ajoutant l'option `na.rm=TRUE` :

12. Avec l'instruction `median(donnees$Age,na.rm=TRUE)` , calculez la médiane de la variable Age.
13. Calculez la médiane de GraviteCopier

3.2.3 Moyenne

La moyenne se calcule avec `mean` .

14. Calculez la moyenne de la variable Age.
15. Calculez la moyenne de GraviteCopier

3.3 Dispersion

3.3.1 Variance

La variance se calcule avec `var` .

16. Calculez la variance de la variable Age.
17. Calculez la variance de GraviteCopier

3.3.2 Écart type

L'écart type se calcule avec `sd`.

18. Calculez l'écart type de la variable Age.
19. Calculez l'écart type de de GraviteCopier

3.3.3 Écart type

L'écart type se calcule avec `sd`.

20. Calculez l'écart type de la variable Age.
21. Calculez l'écart type de GraviteCopier

3.3.4 Quartiles

Les quatre quartiles se calculent tous d'un seul coup avec `quantile`.

22. Calculez les quartiles de la variable Age.
23. Calculez les quartiles de GraviteCopier

4 Représentation graphique

4.0.5 Diagramme en baton

Un diagramme en baton se trace pour les variables nominales, ordonnées et discrètes à partir des effectifs grâce à la fonction `barplot`.

24. Grâce à l'instruction `barplot(table(donnees$NiveauDEtude))`, représentez graphiquement la variable NiveauDEtude
25. Représentez graphiquement la variable MentionBac

4.0.6 Histogramme

Un histogramme, pour les variables continues, s'obtient à l'aide de l'instruction `hist`. Á noter, `hist` s'utilise directement sur la variable et non sur les effectifs de la variable.

26. Grâce à `hist(donnees$Age)`, représentez graphiquement la variable Age.
27. Représentez graphiquement la variable GraviteCopier.
28. Pour ces deux variables, comparez la représentation graphique obtenue avec `barplot` et avec `hist`. Quelle représentation vous semble le plus adapté ?

4.0.7 Boite à moustache

Une boite à moustache s'obtient à l'aide de l'instruction `boxplot`. Á noter, `boxplot` s'utilise directement sur la variable et non sur les effectifs de la variable.

29. Grâce à `boxplot(donnees$Age)`, représentez graphiquement la boite à moustache de la variable Age.
30. Représentez graphiquement la boite à moustache de la variable GraviteCopier.

Fin de l'analyse univarié !