Quelques rappels en analyse uni et bidimensionnelle

1 Analyses préliminaires

=> consiste à cadrer l'information contenue dans les données.

— Variables : leur signification, leur nature/format, présence de valeurs manquantes, ... ?

— Observations : combien ?

2 Statistique univariée

=> consiste à étudier chaque variable l'une après l'autre.

2.1 pour une variable quantitative

• Numériquement : min, max, $1er$ et $3e$ quartiles, médiane, moyenne, écart-type.
• Graphiquement : histogramme, boîte à moustaches.
=> peut servir à éliminer une variable si elle est quasiment constante, à éliminer des individus s'ils ont des valeurs aberrantes ou du moins à repérer des individus ayant des valeurs particulières, à transformer une variable pour la rendre plus régulière.
• Outils : le meilleur moyen d'analyser une variable quantitative sous SAS est d'utiliser le module `distribution` de `SAS Insight`. En l'absence de `SAS Insight`, utiliser les procédures `means` et `univariate`.

2.2 pour une variable qualitative

• Numériquement : tableau d'effectifs et de fréquences.
• Graphiquement : diagramme en bâtons.
=> peut servir à éliminer une variable si elle est presque constante ou à regrouper des modalités proches si effectifs trop faibles.
• Outils : pour analyser une variable qualitative sous SAS, on peut utiliser le module `distribution` sous `SAS Insight` en affichant les tableaux de fréquences (frequency counts), ou la procédure `freq`.

3 Statistique bivariée

=> permet d'étudier la relation entre deux variables.

3.1 pour 2 variables quantitatives, notées x et y

• Graphiquement : Nuage de points entre x et y
• Numériquement : Cœfficient de corrélation linéaire :

$$r(x,y) = \frac{cov(x,y)}{\sigma(x)\sigma(y)}$$

Si $r(x,y)$ proche de 0 => pas de corrélation entre les 2 variables.
Si $r(x,y)$ proche de -1 => présence d'une corrélation linéaire négative entre x et y.
Si $r(x,y)$ proche de $+1$ => présence d'une corrélation linéaire positive entre x et y.

• Test de nullité du cœfficient de corrélation :
$H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$, où ρ est la vraie valeur inconnue du cœfficient de corrélation.

On utilise la statistique de test :

$$Tcal = r(x,y)\sqrt{\frac{n-2}{1-r(x,y)^2}}$$

à comparer au fractile de la loi de Student à $n-2$ ddl.
Si $Tcal > t_{n-2,1-\alpha/2}$ ou p-value< 5% => on rejette H_0 et on conclut à la présence d'une corrélation significative entre x et y.
=> peut être complétée par une régression linéaire simple si nécessaire à condition que la corrélation soit significative et que l'on puisse définir une relation de cause à effet entre les 2 variables.

• Outils :
  — sous `SAS Insight`, utiliser les modules `Scatter plot` et `Multivariate` (en affichant le tableau des p-values).
  — procédures `gplot` et `corr`.

3.2 pour 2 variables qualitatives

Utilisons 2 variables qualitatives, respectivement, à L et C modalités.

• Graphiquement : Diagramme en bâtons.
• Numériquement : Tableau de contingence croisant les 2 variables, constituée de $L \times C$ cellules.
Dans chaque cellule (l,c) :
  — les effectifs n_{lc},
  — le pourcentage de la cellule par rapport au nombre total d'observations n_{lc}/n,
  — le pourcentage en ligne (appelé rowpct sous SAS) : $n_{lc}/n_{l.}$,
  — le pourcentage en colonne (appelé colpct sous SAS) : $n_{lc}/n_{.c}$.
Les totaux des lignes et colonnes donnent la répartition des individus selon chaque variable.

• Test d'indépendance du Chi-deux
=> permet de détecter s'il existe une liaison significative entre les 2 variables, c'est-à-dire si la répartition des individus selon une variable sera la même ou non selon les modalités de l'autre variable.
On utilise la statistique de test du Chi-deux :

$$\chi^2_{cal} = \sum_{l,c} \frac{(O_{l,c} - E_{l,c})^2}{E_{l,c}}$$

où $O_{l,c}$ est l'effectif observé pour la cellule (l,c) et $E_{l,c}$ est l'effectif attendu pour la cellule (l,c) sous l'hypothèse d'indépendance entre les deux variables de la façon suivante :

$$E_{l,c} = \frac{Total_l \times Total_c}{Total}$$

Cette statistique de test est à comparer au fractile de la loi du Chi-deux à $(L-1)\times(C-1)$ ddl, à condition que tous les effectifs attendus soient supérieurs à 5.
Si $\chi^2_{cal} > \chi^2_{(L-1)(C-1),1-\alpha}$ ou p-value< 5%, on rejette l'hypothèse d'indépendance et on conclut à une liaison significative entre les deux variables.

• Outils : procédure `freq` avec l'option `chisq`.

3.3 pour une variable quantitative et une variable qualitative

• Graphiquement : Boîte à moustaches ou histogramme de la variable quantitative par groupe, c'est-à-dire selon les modalités de la variable qualitative.
• Numériquement : Calcul des indicateurs statistiques habituels pour variable quantitative, par groupe.
• Test : Comparaison de moyennes par une analyse de variance à un facteur, par un test de Student ou des tests non-paramétriques.
• Outils : proc `means` avec l'instruction `class`, proc `ttest` (pour comparer deux moyennes), proc `npar1way` (test non-paramétrique de comparaison de moyennes), proc `glm`.