Initiation aux statistiques univariées sur SAS

Année Universitaire 2019-2020

Par

Cédric Cally-Caballero, Aurélien Dautois, Alexandre Fernandez &Théo Ripoche

Encadrant Universitaire : **Monsieur Ricco Rakotomalala** : Enseignant chercheur à l'université Lumière Lyon 2

Introduction

Statistiques Descriptives

- La **statistique descriptive** est la branche des statistiques qui regroupe les nombreuses techniques utilisées pour décrire un ensemble relativement important de données.

- Si les données ne sont relatives qu'à une seule variable, on parle de statistique descriptive univariée. Dans le cas où l'on s'intéresse à deux variables simultanément, on met en œuvre la statistique descriptive bivariée. Si l'ensemble de données provient de l'observation de plusieurs variables, on doit faire appel aux méthodes de la statistique descriptive multivariée.

Introduction

Pourquoi

- L'objectif des statistiques univariées est de décrire, c'est-à-dire de résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses en se concentrant uniquement sur une variable à la fois.

- C'est utilisé pour avoir une idée plus précise des données et permet de trier l'information dès le départ.

Comment

- La statistique univariée utilise des techniques de calcul tel que la moyenne, médiane, quartiles pour les variables quantitatives et des comptages pour les variables qualitatives.

- Les représentations graphiques comme les boxplots, histogrammes sont également utilisées afin de représenter au mieux les données.

Statistiques Univariées sous SAS : PROC FREQ

Syntaxe :

```
PROC FREQ <options>;
by variable;
tables liste des croisements requis </options>;
weight variable;
```

La proc FREQ permet de réaliser des statistiques univariées sur des variables **nominales** en les représentant dans des tableaux de fréquence.

Diverses options statistiques peuvent s'appliquer après le « data » telles que « order » ou « nlevels ».

Les Instructions :

- « TABLE » pour choisir la ou les variables à afficher dans les tableaux.
- « BY » pour regrouper les résultats par modalités.
- « WEIGHT » permet de pondérer avec une autre variable.

```
/*Répartition du nombre de cylindres pour les automatiques*/
proc freq data=malib.mtcars ;
where am = 0 ;
table cyl ;
RUN ;
```

Statistiques Univariées sous SAS : PROC MEANS

Syntaxe :

```
PROC MEANS DATA=nomtab1 optnum;
  VAR var1 var2 var3 var4 var5 var6 ;
  CLASS var2 …;
  WEIGHT var3;
  ID var4;
  FREQ var5 … ;
  BY var7 …;
  OUTPUT OUT=nomtab2 optnum=lvar ;
```

Calcul les indicateurs statistiques basiques des variables indiquées dans l'instruction « VAR » , toutes les statistiques descriptives sont produites par défaut (print activé) au contraire de la proc Summary.

SI « DATA » n'a pas d'arguments, la dernière table SAS connue sera utilisée. On dispose ensuite de diverses instructions statistiques (N, NMISS,MEAN, STD, MIN, MAX, VAR,KURTOSIS).

Les Instructions :

-« Class » permet de réaliser les calculs par modalités de variables (type de moteur). L'instruction « BY » est similaire mais ne trie pas les classes et sort une table par modalité.

-« Weight » permet de pondérer avec une autre variable.

-« ID » ajoute un identifiant à partir d'une variable.

-« FREQ » donne les fréquences pourcentages simples et cumulés des variables choisies (tableau croisé pouvant être à n modalités).

- « OUTPUT OUT » afin de désigner la table de sortie.

```
/*Principales statistiques sur les cylindres et gear
 en fonction du type de moteur*/
proc means data=malib.mtcars;
class vs;
var cyl gear;
run;
```

Statistiques Univariées sous SAS : PROC SUMMARY

Syntaxe :

```
PROC SUMMARY DATA=nomtab1 optnum;
  VAR var1 var2 var3 var4 var5 var6 ;
  CLASS var2 …;
  WEIGHT var3;
  ID var4;
  FREQ var5 … ;
  BY var6 …;
  OUTPUT OUT=nomtab2 optnum=lvar ;
```

La procédure SUMMARY est identique à la procédure MEANS à l'exception qu'elle ne produit pas automatiquement un rapport en sortie. Ainsi, pour calculer les indicateurs statistiques basiques il faut indiquer les variables dans l'instruction « VAR », il faut appeler les statistiques descriptives que l'on souhaite à la suite du « DATA ».

SI « DATA » n'a pas d'argument, la dernière table SAS connue sera utilisée. On dispose ensuite de diverses instructions statistiques (N, NMISS,MEAN, STD, MIN, MAX, VAR,KURTOSIS).

Les Instructions :

- « Class » permet de réaliser les calculs par modalités de variables (boite de transmission). L'instruction « BY » est similaire mais ne trie pas les classes et sort une table par modalité.
- « Weight » permet de pondérer avec une autre variable.
- « ID » ajoute un identifiant à partir d'une variable.
- « FREQ » donne les fréquences pourcentages simples et cumulés des variables choisies (tableau croisé pouvant être à n modalités).
- « OUTPUT OUT » afin de désigner la table de sortie.

```
/*Valeurs manquantes, nombre d'observations
et moyenne des cylindres et gear
 en fonction du type de transmission*/
proc summary data=TD_STAT.mtcars print nmiss n mean;
var cyl gear;
class am;
run;
```

Statistiques Univariées sous SAS : PROC UNIVARIATE

C'est la principale procédure pour réaliser des analyses statistiques univariées sur des variables quantitatives. Elle reprend toutes les options et instructions des procédures MEANS et SUMMARY et permet également de réaliser des graphiques et des test paramétriques.

Syntaxe :

```
PROC UNIVARIATE <options>;
var liste de variables;
by variable;
class variable;
where (variable1=5) or (variable2=5);
weight variable;
output <out=table sas> <liste de statistiques>;
```

Plusieurs options sont disponibles au début de la procédure :
- « ALL » pour demander toutes les statistiques,
- « FREQ » pour obtenir un tableau de contingence des données,
- « NORMAL » afin d'effectuer un test de normalité,
- « PLOT » afin d'obtenir une visualisation graphique,
- « NO PRINT » pour ne pas avoir une vue de la sortie.

Les Instructions :

Les instructions restent les mêmes que pour les proc MEANS ET SUMMARY.

Liste (non exhaustive) des statistiques :

- MIN, MAX, RANGE
- MEAN, STD, VAR
- Q1, MEDIAN, Q3, QRANGE, P95
- T, PROBT etc.

```
/*Poids moyen et répartition selon le nombre de cylindre*/
proc univariate data=malib.mtcars noprint;
class cyl ;
var wt ;
output out=malib.exemple n=Nobs mean=wt  ;
run ;
```

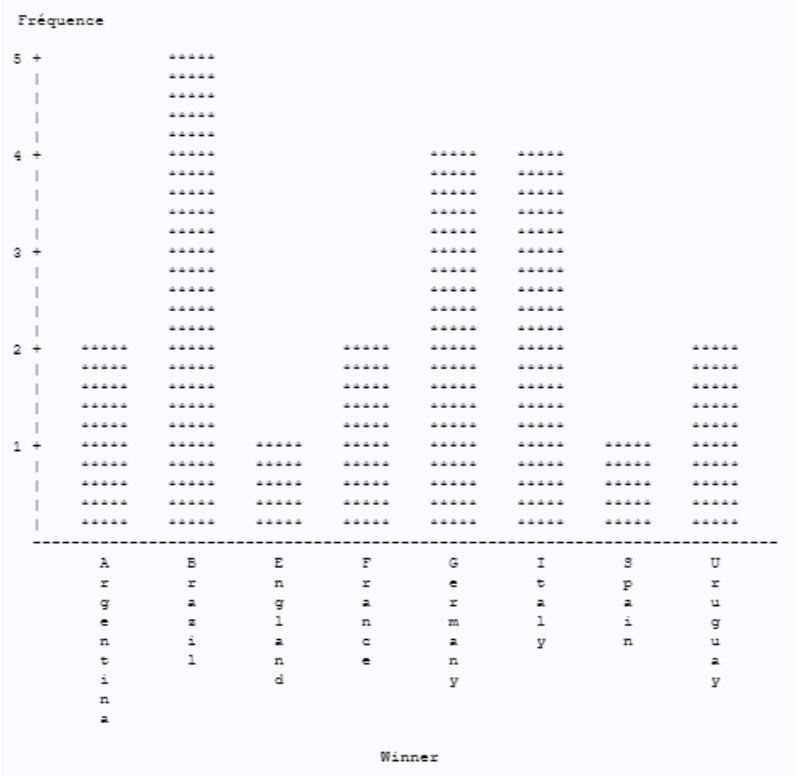
Les procédures graphiques

Proc chart

- Cette procédure permet la création de graphiques en bar, en mosaïque, en camembert ou en radar. Cependant cette procédure affiche une « sortie textuelle » des graphiques.

proc chart data=*fichierchoisie*;
vbar *variablechoisie* ; → histogramme
block *variablechoisie* ; → mosaïque
pie *variablechoisie* ; → camembert
star *variablechoisie* ; → étoile

TITLE '*Titrevoulu*';
run ;



Les procédures graphiques

Proc sgplot

- Cette procédure est la procédure la plus complète pour créer des graphiques.

Les instructions et options de cette procédure permettent de choisir le type de graphique que l'on veut afficher :

PROC SGPLOT DATA = *vosdonnées*;
V/HBAR *variablechoisie* / *options*; → graphique en bar (H pour horizontal et V pour vertical)
HISTOGRAM *variablechoisie* / *options*; → histogramme
DENSITY *variablechoisie* / *options*; → densité
V/HBOX *variablechoisie* / *options*; → boxplot (H pour horizontal et V pour vertical)

TITLE '*Titrevoulu*';
RUN;

Il existe également des options pour personnaliser l'affichage des graphiques : apparence (transparency, width…), axes, labels (legendlabel, ...) etc.

Les options sur les VBAR : stat=freq/mean/sum : spécifie la statistique pour l'axe vertical, groupdisplay=stack/cluster : spécifie comment afficher les bars groupées
Les options sur les HISTOGRAM : freq=numeric-variable : spécifie une variable pour les fréquences de chaque observation
Les options sur les DENSITY : freq=numeric-variable ou weight = numeric-variable : spécifie une variable qui contient des valeurs à utiliser comme poids
Les options sur les VBOX : nomean/nomedian pour choisir ce que l'on veut afficher ou non

Autres procédures

Proc format

```
PROC FORMAT ;
    value moda
    low- ... ="Modalité 1"
    ...
    ... -high="Modalité n" ;
RUN ;
```

Ici la proc format permet de recoder une variable quantitative en variable qualitative en créant des intervalles. On définit dans un premier temps les plages de données puis le nom de la modalité.

Proc stdize

```
proc stdize data=vosdonnées method=choixdelaméthode out=vosnouvellesdonnées
    var variablechoisie;
run;
```

Ici la proc stdize permet de centrer réduire une variable quantitative. On peut définir la méthode de standardisation grâce à l'option method, pour centrer réduire il faut utiliser std.

Proc rank

```
PROC RANK <option(s)>;
    BY <DESCENDING> variable-1
    <...<DESCENDING> variable-n>
    <NOTSORTED>;
    VAR data-set-variables(s);
    RANKS new-variables(s);
```

La proc rank permet d'attribuer pour chaque individu un rang. Cela peut servir ensuite à trier les données selon le rang ou encore faire des tests statistiques tel que Wilcoxon.

Merci de votre attention !

Avez-vous des questions ?