SAS

Travaux Dirigés

II. La Statistique Univariée et Bivariée avec SAS

Références Bibliographiques

- *Introduction à SAS*, E. Duguet, Chapitres 8, 9, 10, Economica.

Pour toute question, commentaire ou suggestion, contacter : salim.lardjane@univ-ubs.fr

1) Lire les données …/duguet/euro15.sas7bdat dans une table SAS nommée tab et les visualiser. Cette base de donnée contient l'évolution de la population (variable *pop*) et du revenu par habitant (variable *rgdpl*) entre 1950 et 1992 pour 15 pays de l'UE.

2) Supprimer les observations pour lesquelles la population ou le revenu par habitant sont manquants de la table tab.

3) Le code utilisé est le suivant.

```
data tab;
       set dir.euro15;
       pop = 1000 * pop;
       if rgdpl ~=. and pop ~=.;
run;
proc print;
run;
```

Rédiger un programme SAS permettant de déterminer l'évolution du revenu moyen par habitant calculé sur l'ensemble des pays de la base, entre 1950 et 1992. A cet effet, on utilisera la **proc means** de SAS.

Noter que ce calcul implique deux choses : *i)* il faut pondérer le calcul du revenu moyen en faisant intervenir la taille des populations ; *ii)* il faut utiliser l'année comme variable de classement, car les calculs doivent être effectués séparément pour chaque année.

Concernant le point i), on a plus précisément, en notant *Y* le revenu moyen par habitant pour l'ensemble des pays,

$$Y = \frac{\sum_i POP_i \times RGDPL_i}{\sum_i POP_i}$$

4) Le code utilisé est le suivant.

```
proc sort data = tab;
       by year;
run;
proc means data = tab mean;
       by year;
       var rgdpl;
       freq pop;
run;
```

Récupérer l'évolution du revenu moyen par habitant dans une table SAS à l'aide de la **proc means**. Visualiser le résultat à l'aide de la **proc print**. Représenter graphiquement l'évolution du revenu moyen par habitant entre 1950 et 1992 à l'aide de la **proc sgplot**.

5) Le code utilisé est le suivant.

```
proc means noprint data = tab mean;
       by year;
       var rgdpl;
       freq pop;
       output out = tab2 mean =;
run;
proc print data = tab2;
run;
proc sgplot data = tab2;
       xaxis values = (1950 to 1992 by 3);
       yaxis values = (3000 to 14000 by 1000);
       series x = year y = rgdpl;
run;
```

On souhaite à présent comparer le revenu par habitant entre la France et le Royaume-Uni sur la période 1950-1992.

On néglige le fait qu'il s'agit de séries temporelles et on suppose que les hypothèses du test de Student de comparaison des moyennes sont vérifiées.

A l'aide de la **proc ttest**, tester l'égalité du revenu moyen par habitant sur la période 1950-1992 entre la France et le Royaume-Uni. Au vu des sorties SAS obtenues, les hypothèses du test de Student vous semblent-elles vérifiées ?

6) Le code utilisé est le suivant.

```
data compa;
       set dir.euro15;
       if country in ('FRANCE','U.K.');
run;
proc ttest data = compa;
       var rgdpl;
       class country;
run;
```

En fait, le résultat obtenu résulte d'un phénomène de compensation entre deux périodes. Rédiger un programme SAS permettant d'effectuer le test séparément sur les périodes 1950-1968 et 1969-1992.

7) Le code utilisé est le suivant.

```
data compa;
       set dir.euro15;
       if country in ('FRANCE','U.K.');
       per = (year >= 1969);
run;
proc sort;
       by per;
run;
proc ttest data = compa;
       by per;
       var rgdpl;
       class country;
run;
```

Représenter graphiquement l'évolution du revenu par habitant pour la France et le Royaume-Uni sur la période 1950-1992. Le graphique obtenu est-il cohérent avec les résultats précédents ?

8) Le code utilisé est le suivant.

```
proc sgplot data = compa;
       xaxis values = (1950 to 1992 by 3);
       series y = rgdpl x = year / group = country;
run;
```

Lire les données …/duguet/univ1.sas7bdat dans une table SAS nommée tab.

Celle-ci fournit la population (variable *pop*) et le PIB réel par habitant (variable *rgdpl*) pour l'année 1990 pour l'ensemble des pays de la planète.

Etudier les propriétés statistiques de la distribution du PIB réel par habitant à l'aide de la **proc univariate** de SAS. Tester en particulier l'hypothèse de normalité de la distribution et obtenir des représentations semi-graphiques de celle-ci.

9) Le code utilisé est le suivant.

```
proc univariate normal plot data = tab;
       var rgdpl;
       id country;
       histogram / normal kernel;
run;
```

On constate globalement que la distribution est bimodale. Le premier mode correspond au pays les plus riches et comporte peu de points et le second correspond au pays les moins riches et comporte beaucoup plus d'observations.

Dans la suite, on va s'attacher à illustrer graphiquement cette inégalité. Pour cela, nous allons représenter l'évolution dans le temps du premier et troisième quartiles et de la médiane de la distribution du PIB réel par habitant entre 1950 et 1992.

10) Utiliser les données …/duguet/univ2.sas7bdat pour obtenir l'évolution des premier et troisième quartiles et de la médiane de la distribution du PIB réel par habitant des différents pays entre 1950 et 1992. Sauvegarder le résultat dans une table SAS nommée tab2 et la visualiser.

11) Le code utilisé est le suivant.

```
proc univariate noprint data = dir.univ2;
       by year;
       var rgdpl;
       output out = tab2 q1 = q1 median = me q3 = q3;
run;
proc print data = tab2;
run;
```

Représenter sur un même graphique l'évolution des premier et troisième quartiles et de la médiane de la distribution étudiée, pour la période 1950-1992.

12) Le code utilisé est le suivant.

```
proc sgplot data = tab2;
       xaxis values = (1950 to 1992 by 3);
       yaxis values = (0 to 10000 by 1000);
       series y = me x = year;
       series y = q3 x = year;
       series y = q1 x = year;
run;
```

On veut à présent examiner à l'aide de divers tests non-paramétriques si la distribution des revenus réels par habitants entre les pays a changé ou non entre 1980 et 1990.

13) Utiliser les données …/duguet/npar.sas7bdat et la **proc npar1way** (avec l'option **edf**) de SAS pour répondre à la question. La variable d'intérêt est *rgdpl*.

14) Le code utilisé est le suivant.

```
proc npar1way edf data = dir.npar;
       class year;
       var rgdpl;
run;
```

On s'intéresse à présent à la relation entre consommation, revenu, et population en France sur la période 1949-1986.

Lire les données …/duguet/corr.sas7bdat dans une table SAS nommée tab et les visualiser.

A l'aide de la **proc corr** de SAS, calculer les coefficients de corrélation de Pearson et de Spearman entre les logarithmes de la consommation (variable *dlc*), du revenu (variable *dlr*) et de la population (variable *dlp*) en France sur la période considérée. Sauvegarder les résultats obtenus dans la table SAS tab2 et les visualiser à l'aide de la **proc print**.

15) Le code utilisé est le suivant.

```
data tab;
       set dir.corr;
run;
proc corr data = tab pearson spearman outp = tab2;
       var dlr dlc dlp;
run;
proc print data = tab2;
run;
```

A l'aide de la **proc corr**, calculer le coefficient de corrélation, à **population constante**, entre consommation et revenu (coefficient de corrélation partiel, obtenu à l'aide de l'instruction **partial**).

16) Le code utilisé est le suivant.

```
proc corr data = tab pearson;
       var dlr dlc;
       partial dlp;
run;
```

Que constate-t-on ?

17) Afin de pousser plus loin notre étude, on souhaite régresser la consommation en volume (variable *cvol*) sur le revenu disponible brut des ménages (variable *rdbmvol*) pour la période 1949-1986. Les données sont dans la table …/duguet/nconso.sas7bdat.

Utiliser la **proc reg** de SAS pour effectuer cette régression. Afficher les prévisions (option **p** de l'instruction **model**) et les résidus obtenus (option **r**).

18) Le code utilisé est le suivant.

```
data tab;
       set dir.nconso;
       if an >= 1949;
run;
proc print;
run;
proc reg data = tab;
       model cvol = rdbmvol / p r;
run;
```

On souhaite à présent représenter la variable expliquée et sa prévision sur un même graphique. Modifier le programme précédent de façon à sauvegarder toutes les variables du tableau tab, la prévision et les résidus dans un tableau SAS nommé tab2 (instruction **output**).

19) Le code utilisé est le suivant.

```
proc reg noprint data = tab;
       model cvol = rdbmvol;
       output out = tab2 p = prevision r = residu;
run;
```

Représenter sur un même graphique la variable expliquée et sa prévision pour la période 1949-1986.

20) Le code utilisé est le suivant.

```
proc sgplot data = tab2;
       xaxis values = (1949 to 1986);
       series y = cvol x = an;
       series y = prevision x = an;
run;
```

21) Tester la normalité des résidus à l'aide de la **proc univariate** de SAS.

22) Le code utilisé est le suivant.

```
proc univariate normal data = tab2;
       var residu;
run;
```

Représenter sur un même graphique un histogramme et un estimateur par noyaux de la densité de la distribution des résidus standardisés (centrés et réduits) , ainsi que la densité de la loi normale associée.

23) Le code utilisé est le suivant.

```
proc stdize data = tab2 out = tab3;
       var residu;
run;
proc print data = tab3;
run;
proc sgplot data= tab3;
  histogram residu;
  density residu;
  density residu / type=kernel;
run;
```

On va à présent s'intéresser à un type différent de problèmes : on dispose d'une base de donnée indiquant la destination des jeunes 9 mois après la fin de leurs études. Les données comportent les informations suivantes.

- Le niveau d'études (variable *niveau*) qui comporte quatre modalités : niveaux I à III (au moins Bac+2), niveau IV (sortie des classes terminales du 2nd cycle, ex. baccalauréat), niveau V (sortie des classes terminales de 2nd cycle court (ex. CAP et BEP) ou abandon du 2nd cycle long avant la classe terminale), niveaux Vbis et VI (pas de qualification professionnelle).
- La destination (variable *destinat*) comporte six modalités pour les hommes et cinq pour les femmes : emploi, chômage, apprentissage, stage, service national (hommes), inactivité.
- La sous-population (variable *spop*) : homme ou femme
- Le nombre de personnes correspondant à chaque croisement des trois variables précédentes est stocké dans la variable *effectif*.

Les données sont dans la table …/duguet/destin.sas7bdat. Les lire dans une table SAS nommée tab et les visualiser an utilisant les labels (étiquettes).

24) Le code utilisé est le suivant.

```
data tab;
       set dir.destin;
run;
proc print label data = tab;
       title "Données Initiales";
run;
```

Obtenir la répartition des trois variables d'intérêt à l'aide de la **proc freq** de SAS. Noter qu'il est nécessaire de faire intervenir une pondération par la variable *effectif* (instruction **weight**) pour que les calculs ne soient pas faussés.

25) Le code utilisé est le suivant.

```
proc freq data = tab;
       weight effectif;
       tables spop niveau destinat / nocum;
       title "Tableaux à une seule entrée";
run;
```

En utilisant la proc freq de SAS, comparer les niveaux d'étude des hommes et des femmes à l'aide d'un tableau à double entrée.

26) Le code utilisé est le suivant.

```
proc freq data = tab;
       weight effectif;
       tables spop*niveau;
       title "Tableau à double entrée";
run;
```

Modifier le code précédent de façon à n'afficher que des pourcentages par rapport au total dans les cellules du tableau obtenu (options **nofreq**, **nocol**, **norow**).

27) Le code utilisé est le suivant.

```
proc freq data = tab;
       weight effectif;
       tables spop*niveau / nofreq nocol norow;
       title "Tableau à double entrée";
run;
```

modifier le programme précédent afin d'obtenir les résultats de l'analyse sous forme de liste (option **list**).

28) Le code utilisé est le suivant.

```
proc freq data = tab;
       weight effectif;
       tables spop*niveau / list;
       title "Tableau à double entrée";
run;
```

On souhaite à présent examiner les croisements entre le niveau d'étude et la destination à la fin des études par sous-population (femmes et hommes séparément). Utiliser pour ce faire la **proc freq** de SAS.

29) Le code utilisé est le suivant.

```
proc freq data = tab;
       weight effectif;
       tables spop*destinat*niveau;
run;
```

Réaliser le même traitement à l'aide de l'instruction **by**. Que constate-t-on ?

30) Le code utilisé est le suivant.

```
proc sort data = tab;
       by spop;
run;
proc freq data = tab;
       weight effectif;
       tables destinat*niveau;
       by spop;
run;
```

On souhaite à présent tester l'indépendance entre le niveau d'étude et la destination à l'aide d'un test du Khi-2.

Supprimer de la table les observations relatives au service national, les données devant concerner à la fois les femmes et les hommes pour que le test puisse être effectué.

31) Le code utilisé est le suivant.

```
data tab2;
       set tab;
       if destinat = "SERVICE NATIONAL" then delete;
run;
```

Réaliser un test du Khi2 de l'indépendance entre niveau d'étude et destination à l'aide de la **proc freq** de SAS (option **chisq**). Afficher les effectifs estimés sous l'hypothèse d'indépendance (option **expected**) et l'écart entre cet effectif estimé et l'effectif réel (option **deviation**).

32) Le code utilisé est le suivant.

```
proc freq data = tab2;
       weight effectif;
       tables niveau*destinat / chisq expected deviation;
run;
```

Sauvegarder dans des tables SAS (option **out** de l'instruction **tables**) le tableau à une entrée concernant la variable niveau et les tableaux croisés entre niveau d'étude et destination par sous population, obtenus à l'aide de la **proc freq** de SAS. Noter qu'on ne peut sauvegarder qu'un tableau par instruction **tables**.

33) Le code utilisé est le suivant.

```
proc freq noprint data = tab2;
       weight effectif;
       tables niveau / out = tab3;
       tables spop*destinat*niveau / out = tab4;
```

```
      title "Tableaux en sortie";
run;
proc print data = tab3;
      title2 "Les niveaux d''étude";
run;
proc print data = tab4;
      title2 "Le tableau de départ hors service national";
run;
```