

SAS

Cours pour l'école doctorale SHS

Le système SAS est un système complet d'analyse et de représentation des données. Il est utilisé dans de très nombreuses entreprises et constitue sur le marché un standard reconnu.

C'est un programme très complet et il existe à l'intérieur de SAS de très nombreuses procédures d'analyse économétriques ou statistiques des données. De façon à ce que le cours soit le plus ouvert possible, nous ne nous concentrerons pas sur ces procédures mais sur les procédures de représentation des données.

Si dans votre travail de thèse, vous utilisez des données et que vous devez offrir une représentation de vos données au moyen de tableaux (de fréquence, tableaux croisés, rapports...), ou de graphiques (histogrammes, courbes, cartes, en deux ou trois dimensions), SAS peut vous être utile en réalisant pour vous et de façon relativement rapide des sorties de très bonne facture (que vous auriez de grandes difficultés à réaliser sous Excel).

Certains doctorants ont suivis le cours d'introduction à SAS que je dispense en master 1 économétrie et statistique appliquée. Une bonne partie de ce cours est consacré à la construction des tables (bases de données) et à leur manipulation. Je tenterai dans ce cours dédié aux étudiants de l'école doctorale de limiter au maximum le temps passé à ces questions. Pour cela, la phase de construction de la table de données sera essentiellement effectuée sous Excel et nous consacrerons une première partie du cours à la préparation des données sous Excel et à leur importation à l'intérieur de SAS.

Afin de vous indiquer les possibilités de ce logiciel (et ce que vous serez normalement capable de produire après ce cours), nous allons utiliser deux jeux de données :

- des données d'enquête (simulées) : au moyen de ces données, on connaît la réponse à deux questions ainsi que le revenu de 217 personnes.

- les données des résultats des élections présidentielles 2002 qui peuvent être téléchargées à partir de cette page :

http://www.interieur.gouv.fr/sections/a_votre_service/elections/resultats/presidentielle/presidentielle-2002

Le second jeu de données nous sera utile afin de présenter les possibilités de production de cartes par SAS.

A partir des données de la première base, vous pouvez souhaitez construire des tableau de fréquences :

Q1				
Q1	Fréquence	Pourcentage	Fréquence cumulée	Pctage. cumulé
0	19	8.88	19	8.88
1	34	15.89	53	24.77
2	52	24.30	105	49.07
3	44	20.56	149	69.63
4	48	22.43	197	92.06
5	17	7.94	214	100.00

Bien entendu, les réponses codées 0,1,2.. sont généralement à associer à de « vraies » réponses. Vous pouvez sans difficultés remplacer ces réponses 0,1.. par le texte de votre choix :

Q1				
Q1	Fréquence	Pourcentage	Fréquence cumulée	Pctage. cumulé
très peu satisfait	19	8.88	19	8.88
peu satisfait	34	15.89	53	24.77
bof bof	52	24.30	105	49.07
satisfait	44	20.56	149	69.63
très satisfait	48	22.43	197	92.06
exceptionnel !	17	7.94	214	100.00

Vous pouvez produire des tableaux croisés :

Table de Q1 par Q2				
Q1(réponse à la question 1)	Q2(réponse à la question 2)			
Fréquence Pourcentage Pctage en ligne Pctage en col.	0	1	2	Total
très peu satisfait	1 1.20 14.29 6.67	4 4.82 57.14 11.76	2 2.41 28.57 5.88	7 8.43
peu satisfait	3 3.61 21.43 20.00	5 6.02 35.71 14.71	6 7.23 42.86 17.65	14 16.87
bof bof	3 3.61 10.71 20.00	10 12.05 35.71 29.41	15 18.07 53.57 44.12	28 33.73
satisfait	4 4.82 30.77 26.67	6 7.23 46.15 17.65	3 3.61 23.08 8.82	13 15.66
très satisfait	3 3.61 17.65 20.00	8 9.64 47.06 23.53	6 7.23 35.29 17.65	17 20.48
exceptionnel !	1 1.20 25.00 6.67	1 1.20 25.00 2.94	2 2.41 50.00 5.88	4 4.82
Total	15 18.07	34 40.96	34 40.96	83 100.00

des statistiques simples en fonction des réponses à une question (par exemple) :

Variable d'analyse : Revenu						
Q1	N Obs	N	Moyenne	Ecart-type	Minimum	Maximum
très peu satisfait	19	19	25302.05	25520.59	3530.00	97231.00
peu satisfait	34	34	26064.00	20217.53	955.0000000	72221.00
bof bof	52	52	26348.06	23841.19	81.0000000	88584.00
satisfait	44	44	27377.05	19334.48	1404.00	70543.00
très satisfait	48	48	21601.75	21506.76	169.0000000	90113.00
exceptionnel !	17	17	27649.59	24042.38	90.0000000	84848.00

ou encore :

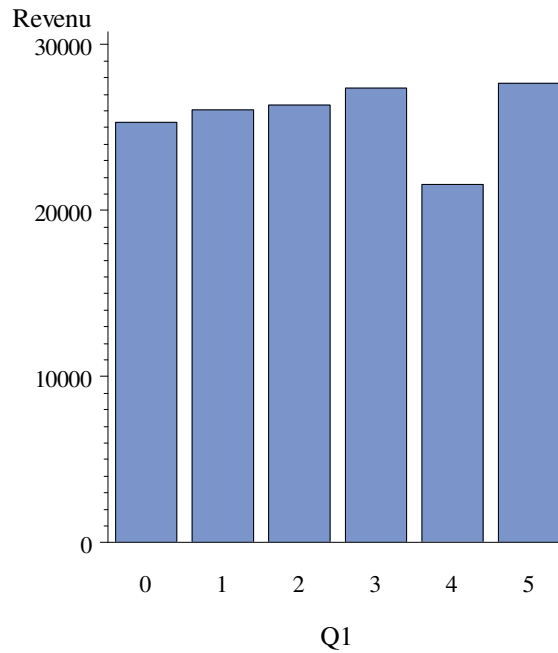
	Revenu		
	revenu moyen	Min	nombre de répondants
Q1			
très peu satisfait	25302.05	3530.00	19
peu satisfait	26064.00	955.00	34
bof bof	26348.06	81.00	52
Satisfait	27377.05	1404.00	44
très satisfait	21601.75	169.00	48
exceptionnel !	27649.59	90.00	17
toutes modalités de réponse confondues	25460.43	81.00	214

Ou enfin,

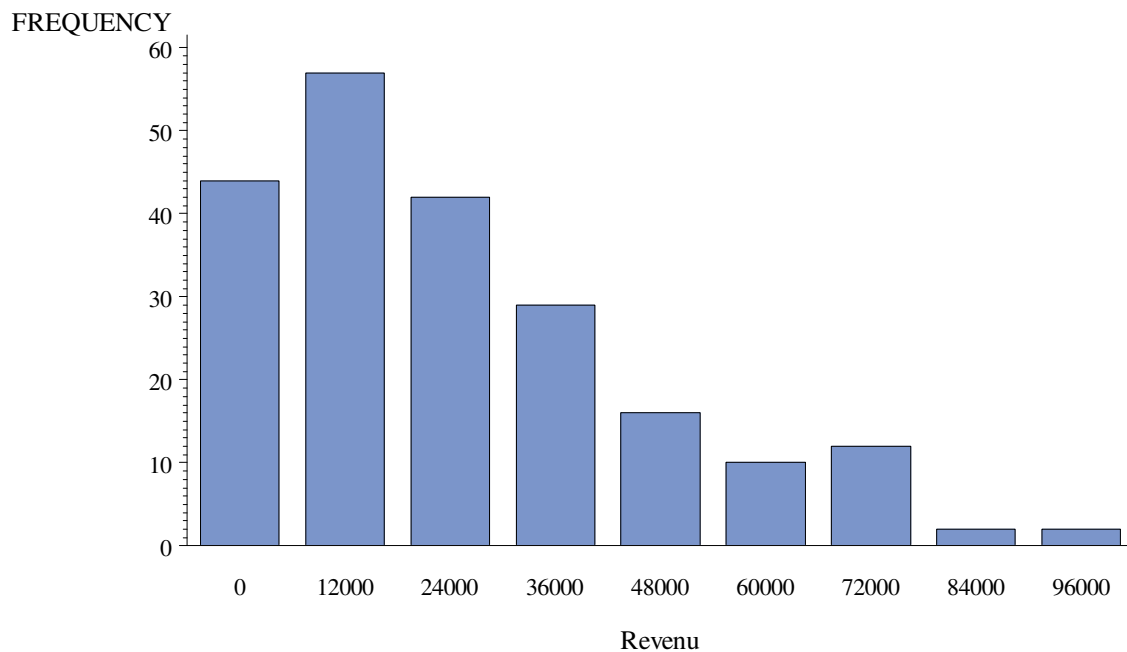
	Q2						toutes modalités de réponse confondues	
	0		1		2			
	Revenu		Revenu		Revenu		Revenu	
	revenu moyen	nombre de répondants	revenu moyen	nombre de répondants	revenu moyen	nombre de répondants	revenu moyen	nombre de répondants
Q1								
très peu satisfait	23014.00	1	16954.75	4	7596.50	2	15146.57	7
peu satisfait	7525.67	3	28062.20	5	26344.50	6	22925.36	14
bof bof	46405.00	3	25908.20	10	30533.07	15	30581.89	28
satisfait	14882.25	4	20542.17	6	25534.00	3	19952.62	13
très satisfait	26499.00	3	23572.00	8	19230.00	6	22556.06	17
exceptionnel !	8590.00	1	12751.00	1	31294.50	2	20982.50	4
toutes modalités de réponse confondues	22161.47	15	23288.00	34	26053.74	34	24217.36	83

Au moyen des procédures graphiques de SAS, vous avez la possibilité de créer des histogrammes :

Revenu moyen en fonction des réponses à Q1

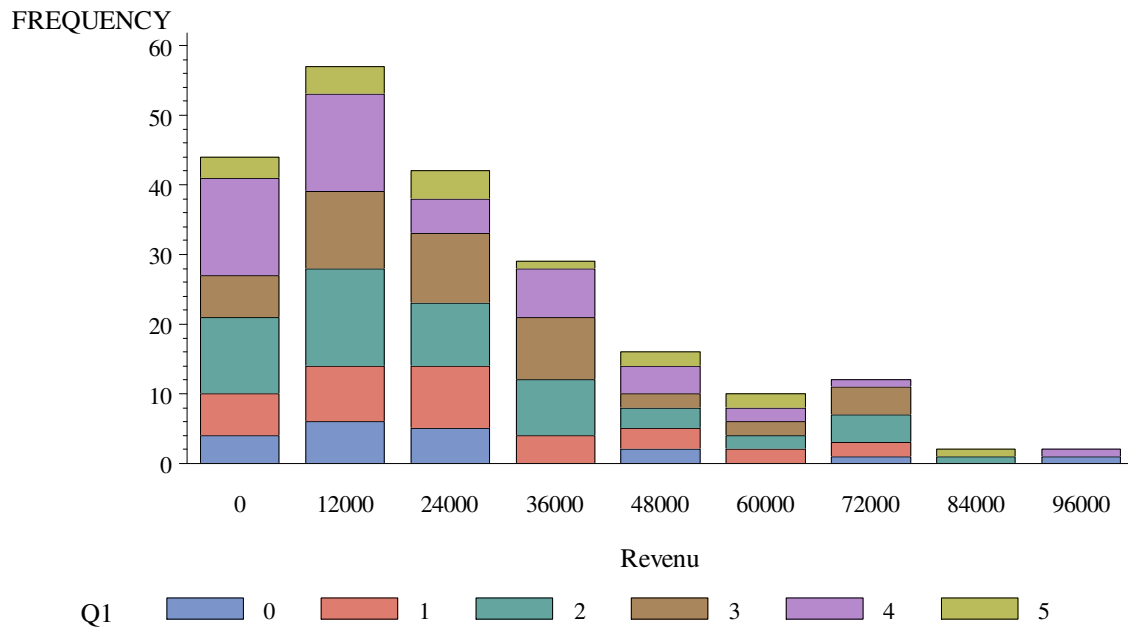


Histogramme de la distribution des revenus

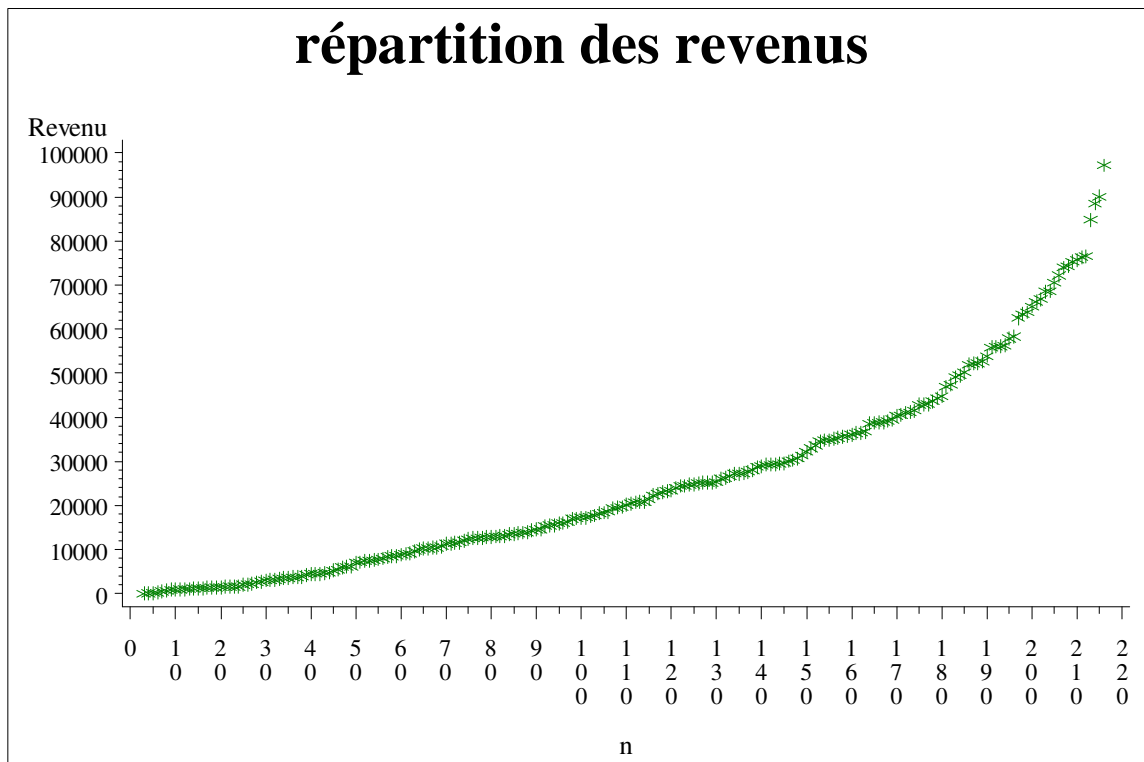


Histogramme de la distribution des revenus

répartition des réponses à Q1 au sein de chaque barre

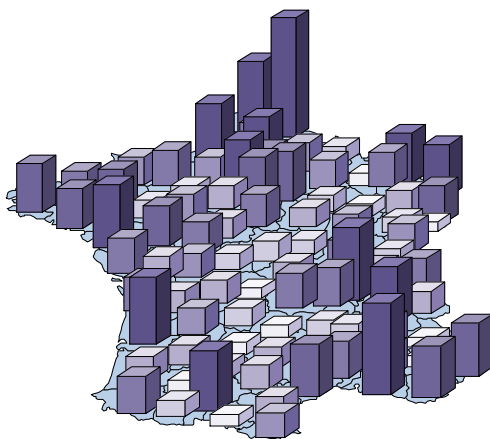









des nuages de points ou des courbes :



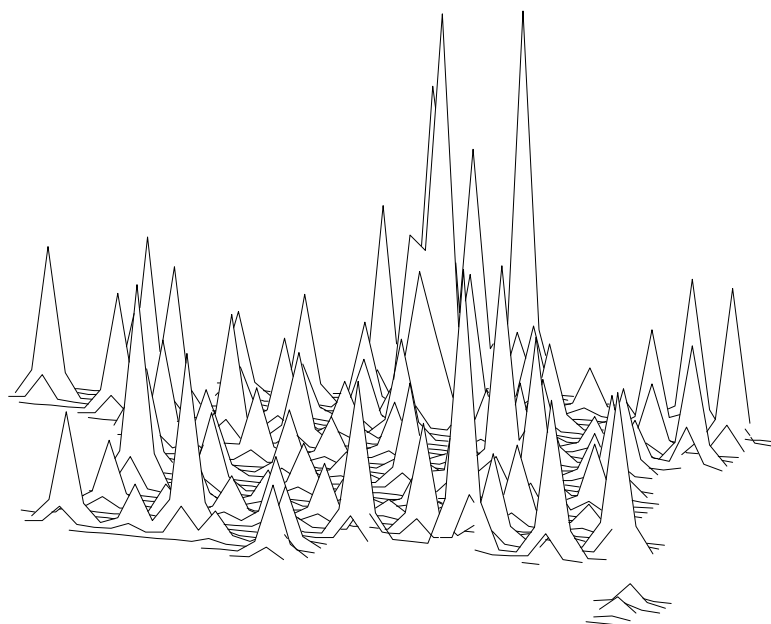
Enfin, si vous êtes intéressés par la production de cartes, c'est tout à fait possible :

Suffrages exprimés - election présidentielle 2002



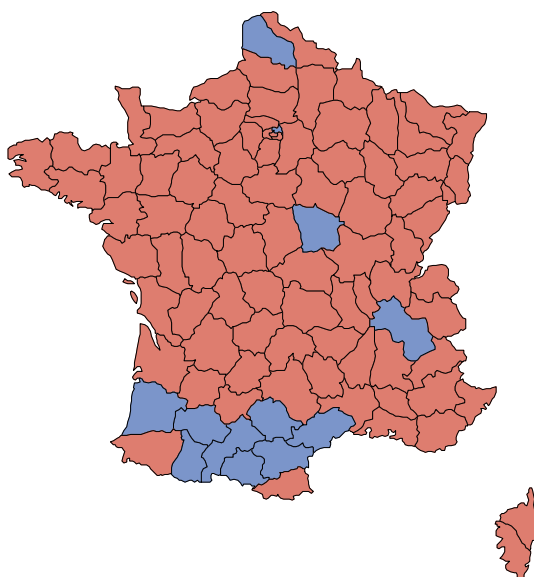
Votants		45517 - 104683		119618 - 167668		167817 - 207163
		208397 - 280875		281692 - 352329		352993 - 488202
		491117 - 1224216				



Suffrages exprimés - election présidentielle 2002



Election présidentielle 2002

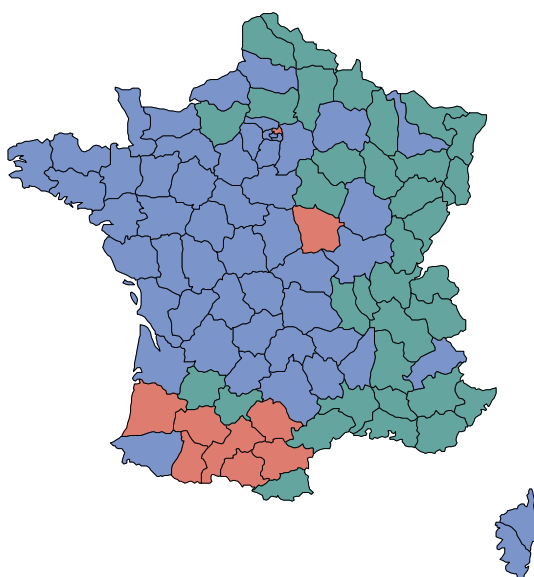
qui de Chirac ou de Jospin à fait le meilleur score ?



tep  Jospin  Chirac

Election présidentielle 2002

meilleurs scores par département pour qui ?



tt  Chirac  Jospin  Le Pen

Les résultats que vous produisez au moyen de SAS sont aisément exportables. Ainsi les résultats présentés dans ce descriptif ont été directement produit en RTF pour ensuite être intégrés dans mon traitement de texte.

Vous avez aussi la possibilité de créer des fichiers PDF, HTML (ou LaTeX). Les résultats présentés ici peuvent être téléchargés dans un format PDF ici :

<http://www.univ-orleans.fr/deg/masters/ESA/SR/ecoldoct/resultat.pdf>

Dans un format HTML ici :

<http://www.univ-orleans.fr/deg/masters/ESA/SR/ecoldoct/frame.html>

Les questions à vous poser et ce que vous devez savoir :

1- Est-ce que les données que je souhaite présenter dans mon travail de thèse peuvent être aisément présentées au moyen d'Excel ?

Si votre réponse est oui, vous n'avez peut-être pas besoin de suivre ce cours

2- Quels sont les pré requis ?

Aucun si vous savez allumer un ordinateur et si vous savez ce que vous voulez produire au moyen de vos données. Ce cours ne peut cependant pas se substituer à un cours de statistique descriptive

3- Est-ce que l'environnement SAS est simple à comprendre ?

Attention, il convient de « programmer » : il ne s'agit pas de « cliquer sur des boutons ». Vous devez écrire vos programmes. La syntaxe de la programmation n'est cependant pas difficile. Ainsi pour produire la dernière carte présentée plus haut, le programme est :

```
title "Election présidentielle 2002";
title2 "meilleurs scores par département pour qui ?";
proc gmap data=election
          map=maps.france ;
id id;
choro tt / discrete ;
format tt teppp.;
run;
```

Chaque instruction est « aisément » compréhensible :

TITLE : permet de définir les titres

PROC GMAP : est la procédure qui crée la carte

DATA=election : il faut utiliser le fichier ELECTION

MAP=maps.france : il faut utiliser le fond de carte suivant
(carte de France au niveau département)

ID id : au sein du fichier de données, les départements sont saisis par la variable ID

CHORO tt / discrete : type de graphique produit (TT est la variable représentée)

FORMAT ... permet de créer la légende

RUN = exécute le programme

4- Est-ce qu'une fois ce cours terminé, je pourrai me débrouiller sous SAS sans aucune aide ?

Comprenez ce cours comme une « introduction », vous aurez peut être par la suite à approfondir vos connaissances soit en consultant la documentation SAS soit en lisant un manuel d'introduction, par exemple le mien :

<http://www.pearson.fr/livre/?GCOI=27440100978750>

Vous devriez normalement pouvoir tout de même produire vos premiers tableaux de résultats et vos premiers graphiques à la suite de ce cours.

5- Vous avez des données à exploiter et à présenter dans votre thèse ma thèse mais vous ne savez pas si SAS va réaliser ce que vous souhaitez.

Contactez moi, décrivez moi vos données et dites moi ce que vous souhaitez en faire : sebastien.ringuede@univ-orleans.fr

6- SAS est un logiciel libre ?

Pas vraiment... voyez si votre laboratoire dispose de licences recherche. Si ce n'est pas le cas, demandez à votre directeur de thèse ou à votre directeur de laboratoire de me contacter.

Durée du cours proposé : 3 jours - 7 heures / jours

Date : à préciser avec les participants

Lieu : UFR Droit Eco Gestion