SAS - Stat descriptives bivariées, graphique

Alexandre Colin, Leo Delecourt, Amina Lrhoula, Guillaume Martial

INTRODUCTION

- Ce support met en avant les statistiques bivariées et les graphiques qui leurs sont associés sur le logiciel SAS. Il s'agit d'une présentation très technique des fonctionnalités de SAS et la théorie sur les tests statistiques notamment ne sera pas abordée.

- Il est accompagné d'une fiche de TD ainsi que d'une vidéo tutoriel de correction des exercices.

- Nous verrons dans un premier temps les tableaux croisés (FREQ, MEANS, TABULATE) puis les graphiques ((G)PLOT, (G)CHART, SGPLOT) et enfin les tests statistiques (CORR, KHI², TTEST, REG, GLM).

I) Tableaux croisés

PROC FREQ

- Créé un tableau de contingence via l'option * entre deux variables. Sinon, l'option BY créé plusieurs tableaux (un par modalité de var2).

PROC FREQ DATA = nom_table;

TABLES var1 * var2 */options*;

RUN;

PROC FREQ DATA = nom_table;

BY var2;

TABLES var1 */options*;

RUN;

La procédure FREQ

Fréquence Pourcentage	Table de num_dept par categorie_de_l_etablissement		
	categorie_de_l_etablissement		
num_dept	Laboratoire d'Analyses	Laboratoire de Biologie Médicale	Total
75	53 6.44	133 16.16	186 22.60
77	7 0.85	59 7.17	66 8.02
78	26 3.16	89 10.81	115 13.97
91	20 2.43	64 7.78	84 10.21
92	35 4.25	90 10.94	125 15.19
93	21 2.55	59 7.17	80 9.72
94	20 2.43	72 8.75	92 11.18
95	19 2.31	56 6.80	75 9.11
Total	201 24.42	622 75.58	823 100.00

Les options norow (%lignes), nocol (%colonnes) ou nopercent sont utiles pour simplifier le tableau « classique ».

PROC MEANS

- Statistiques quantitatives (médiane, moyenne, …) d'une variable en fonction d'une autre (un tableau par modalité de var2) via l'option BY.

PROC MEANS DATA = nom_table */options*;

BY var2;

VAR var1;

RUN;

La procédure MEANS

categorie_de_l_etablissement=Laboratoire d'Analyses

Variable d'analyse : lat				
N	Moyenne	Ec-type	Minimum	Maximum
201	48.8359362	0.1205741	48.2651880	49.1277355

categorie_de_l_etablissement=Laboratoire de Biologie Médicale

Variable d'analyse : lat				
N	Moyenne	Ec-type	Minimum	Maximum
622	48.8388419	0.1195089	48.3722819	49.1550446

On peut spécifier les indicateurs souhaités en option, par exemple mean, median, min, max ou Q1 et Q3.

I) Tableaux croisés

4 PROC TABULATE

- Procédure pouvant croiser deux variables ou plus dans un tableau. Pour deux variables, donne exactement la même chose que FREQ sans les pourcentages lignes et colonnes.
- L'instruction CLASS prend toutes les variables, qu'on organise ensuite dans le TABLE. Différentes options d'affichage et de calcul d'aggrégat disponibles.

```
PROC TABULATE DATA = LABO;

CLASS num_dept categorie_de_l_etablissement cp_ville;

TABLES num_dept = "" * (cp_ville = "" ALL = "Total"), categorie_de_l_etablissement = "Catégories" ALL = "Total" * N ="" / BOX = "Départements/Villes";

RUN;
```

Syntaxe:

TABLES (lignes),

(colonnes) *

(cellules) / *options*;

Le symbole « = » sert ici à renommer des colonnes ou des lignes.

A la place du **N**, *on peut mettre toutes sortes de statistiques comme sum ou mean*

Départements/Villes		Catégories		Total
		Laboratoire d'Analyses	Laboratoire de Biologie Médicale	
		N	N	
75	75001 PARIS	2	1	3
	75002 PARIS	1	1	2
	75003 PARIS	1	1	2
	75004 PARIS	.	2	2
	75005 PARIS	2	3	5
	75006 PARIS	2	3	5
	75007 PARIS	1	3	4
	75008 PARIS	2	7	9
	75009 PARIS	2	7	9
	75010 PARIS	4	4	8

II) Graphiques

PROC GPLOT

- Permet de faire des nuages de points ou des courbes (nuages de points reliés).

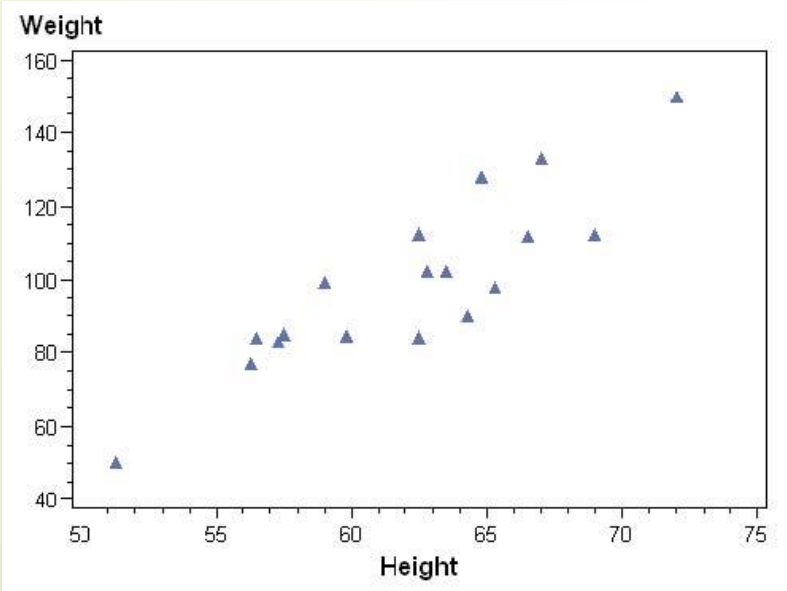
PROC GPLOT DATA = nom_table;

PLOT var1 * var2 / *options*;

RUN;

- L'instruction SYMBOL (à mettre avant la procédure) permet de modifier les points du graphique. JOIN permet une courbe, NONE un nuage de points.

SYMBOL i=JOIN/NONE v = triangle… c = blue;



PROC GCHART

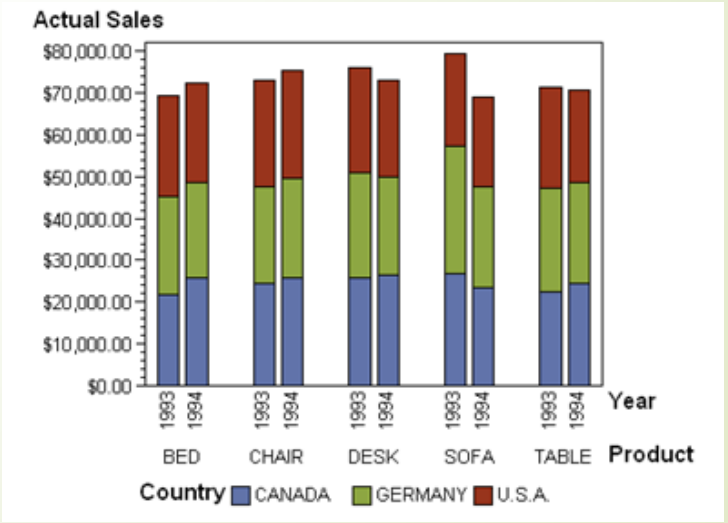
- Permet de faire des diagrammes en barres (cumulées ou non), avec des groupes et sous groupes d'analyse pour croiser des variables.
- Dans les options, on retrouve GROUP et SUBGROUP qui permettent de faire des analyses croisées.

PROC GCHART DATA = nom_table;

VBAR/HBAR var1 / GROUP = var2 *options*;

RUN;

On peut aussi choisir l'agrégation des données en mettant sumvar = statistique avec TYPE = mean, max, freq…



II) Graphiques

PROC BOXPLOT

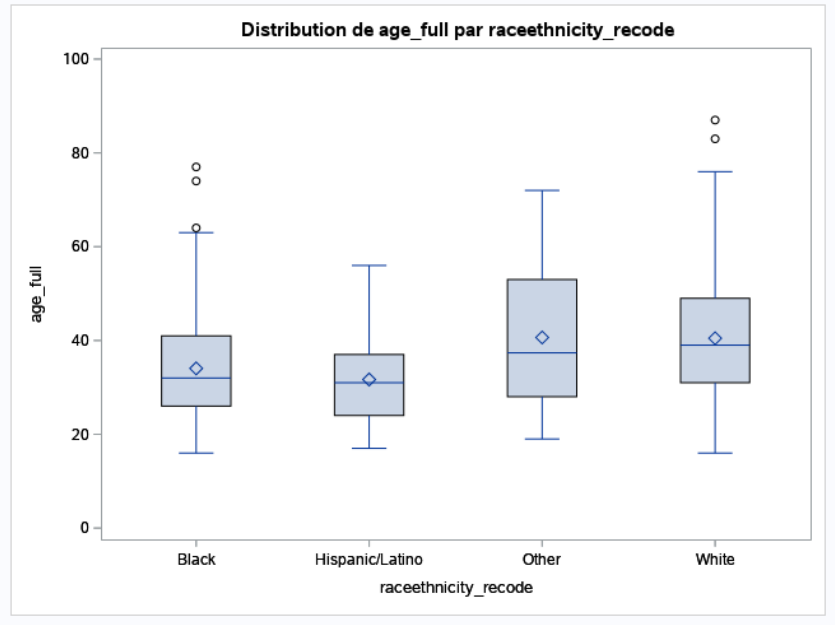
- Procédure pour créer des boîtes à moustaches.

PROC BOXPLOT DATA = nom_table;

plot var1 * var2 / *options*;

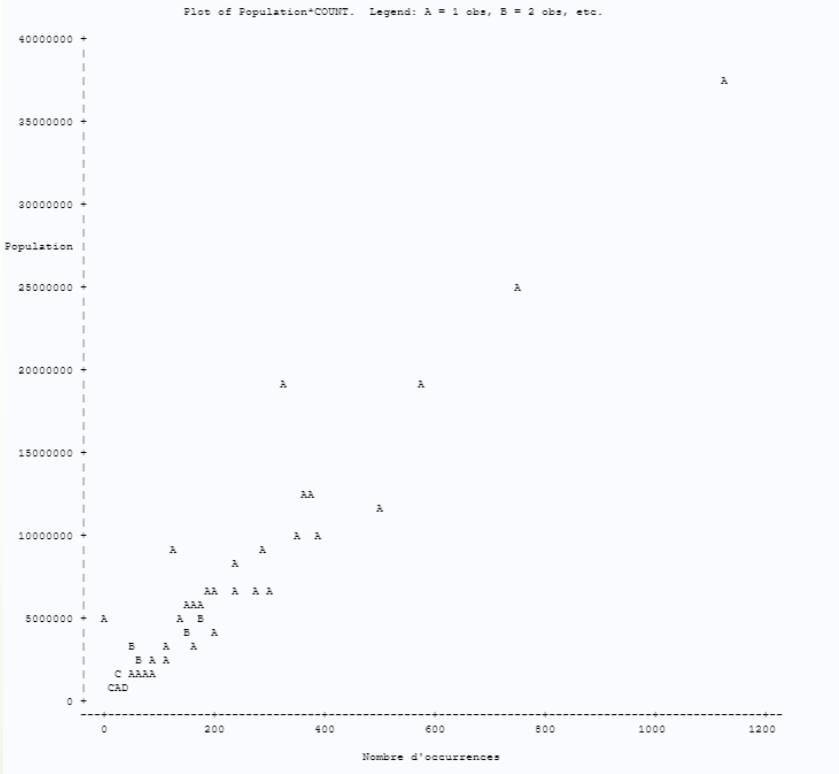
run;

Options pour modifier les axes, les formes des boîtes ou le type de boîte (boxtype, boîtes qui montrent les valeurs aberrantes ou non par exemple).



PROC CHART et PLOT

- Marchent de la même façon que plot et chart mais graphique beaucoup moins beau. Procédures très peu utilisées quand on a sgplot, gplot et gchart.



II) Graphiques

7 PROC SGPLOT

- Permet de créer la plupart des diagrammes utilisés pour les statistiques bivariées, dans l'ordre: nuage de points, courbe, boîtes à moustaches, diagramme en barres, histogramme.
- Il est possible de superposer différents graphiques dans une même procédure sgplot, en mettant par exemple une instruction scatter et une instruction series ou deux instructions series.

PROC sgplot DATA = nom_table;

scatter x=var1 y=var2 */options*;

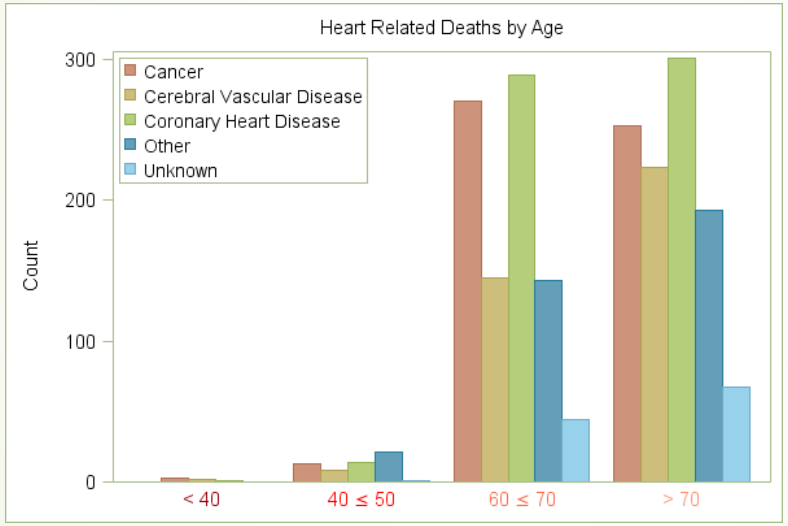
series x=var1 y=var2 */options*;

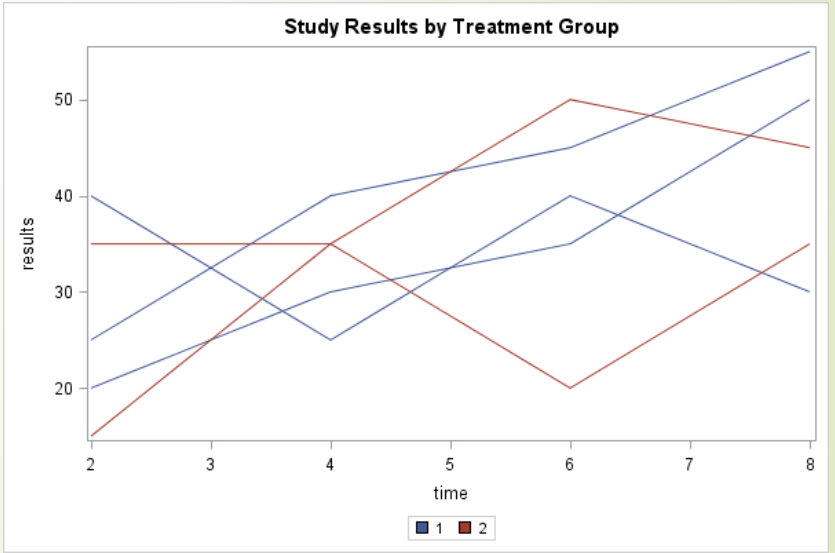
vbox/hbox var1 / category=var2 *options*;

vbar/hbar var1 / category=var2 *options*;

histogram var */options*;

RUN;





On peut gérer dans les options les axes avec x(ou y)axis label = …, min = …, max = … et même ajuster la transparence et la largeur des diagrammes (barwidth et transparancy).

III) Tests statistiques

PROC CORR

- Mesurer et tester la corrélation linéaire entre deux variables quantitatives.

PROC CORR DATA=nom_table *options*;

VAR var1;

WITH var2;

RUN;

Coefficients de corrélation de Pearson Proba > |r| sous H0: Rho=0 Nombre d'observations	
	COUNT
Population	0.95642 <.0001 50

Statistiques simples						
Variable	**N**	**Moyenne**	**Ec-type**	**Somme**	**Minimum**	**Maximum**
Population	52	6016451	6778520	312855438	562803	37325068
COUNT	50	186.86000	207.35664	9343	6.00000	1122

On peut choisir le type de coefficient calculé (spearman ou pearson par exemple), print les covariances avec cov, nomiss pour exclure les valeurs manquantes, etc…

PROC FREQ (Khi²)

- Test d'indépendance du Khi² entre deux variables qualitatives. C'est une option de la proc freq classique. L'option expected permet d'ajouter les valeurs théoriques au tableau de contingence classique.

PROC FREQ DATA=nom_table;

TABLE var1*var2 / chisq expected *options*;

RUN;

Mêmes options que la proc freq classique pour alléger le tableau de sortie.

Le tableau ci-dessous est affiché en plus d'un tableau de proc freq classique.

Statistique	DDL	Valeur	Prob
Khi-2	7	10.1229	0.1817
Test du rapport de vraisemblance	7	11.3584	0.1237
Khi-2 de Mantel-Haenszel	1	0.1243	0.7245
Coefficient Phi		0.1109	
Coefficient de contingence		0.1102	
V de Cramer		0.1109	

III) Tests statistiques

PROC TTEST

- Test de student de comparaison de moyennes (Hypothèse nulle d'égalité des moyennes).
- On teste les valeurs de var2 selon une variable binaire qualitative var1.

PROC TTEST DATA=nom_table *options*;

CLASS var1;

VAR var2;

RUN;

Sex	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
F		60.5889	56.7315	64.4463	5.0183	3.3897	9.6140
M		63.9100	60.3776	67.4424	4.9379	3.3965	9.0147
Diff (1-2)	Pooled	-3.3211	-8.1447	1.5025	4.9759	3.7339	7.4596
Diff (1-2)	Satterthwaite	-3.3211	-8.1551	1.5129			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	17	-1.45	0.1645
Satterthwaite	Unequal	16.727	-1.45	0.1652

Options pour modifier l'hypothèse nulle (H0), sides = 2/L/U pour choisir le sens de H1 (different, inférieur, supérieur), alpha pour modifier le risque d'erreur, plot pour obtenir des graphiques supplémentaires.

PROC REG

- Procédure de régression linéaire simple ou multiple. Permet d'avoir les coefficients d'une droite de régression.
- Var1 est la variable à expliquer, les autres sont les variables explicatives.

PROC REG DATA=nom_table;

MODEL var1 = var2;

RUN;

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Hips

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	25.78571	25.78571	5.32	0.0331
Error	18	87.16429	4.84246		
Corrected Total	19	112.95000			

Root MSE	2.20056	R-Square	0.2283
Dependent Mean	37.05000	Adj R-Sq	0.1854
Coeff Var	5.93943		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	20.31190	7.27020	2.79	0.0120	5.03778	35.58603
Chest	1	0.45238	0.19604	2.31	0.0331	0.04051	0.86425

III) Tests statistiques

10 PROC GLM

- Permet d'effectuer une ANOVA sous SAS. Option intégrée à la procédure pour effectuer le test d'égalité des variances (hovtest=levene), pour avoir des statistiques sur chaque population étudiée par le test (means var1), …
- On met la variable qualitative comme CLASS et les deux variables à utiliser dans model.

```
proc glm data=nom_table;
class var1;
model var2=var1;
means var1 / hovtest=levene options;
run;
```

Variable dépendante : age_full

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Modèle	3	6258.79125	2086.26375	13.46	<.0001
Erreur	463	71784.78974	155.04274		
Total sommes corrigées	466	78043.58099			

R-carré	Coef de var	Racine MSE	age_full Moyenne
0.080196	33.32234	12.45162	37.36717

Test de Levene pour l'homogénéité de la variance age_full
ANOVA des écarts au carré des moyennes de groupe

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
raceethnicity_recode	3	683373	227791	3.86	0.0095
Erreur	463	27329178	59026.3		

Niveau de raceethnicity_recode	N	age_full	
		Moyenne	Ec-type
Black	135	34.0444444	11.4408711
Hispanic/Latino	67	31.6771219	9.1757435
Other	29	40.6460118	14.9352902
White	236	40.4803694	13.4389833

BIBLIOGRAPHIE

- http://www.sasreference.com/ : Un blog avec beaucoup de documentation sur SAS.
- https://od-datamining.com/ : Un blog ayant tout une partie de documentation SAS utile et très accessible, avec des exemples concrets.
- https://support.sas.com/en/documentation.html : Documentation officielle de SAS, très complète mais pour des exemples, le second lien est souvent plus efficace.