

I Statistiques courantes

On reprend le programme initial de la fiche précédente :

```
data table1;
options nodate;
title 'Saisie de données';
input id$ age genre$ salaire anciennete etatcivil$;
cards;
AD4632 42 M 42000 10 M
HF4523 39 F 38000 12 C
SD4510 47 F 35000 14 M
DE4512 24 M 28000 2 M
S05601 43 M 38000 15 M
CK4539 62 M 45000 28 C
VK3461 27 M 44000 4 C
S04218 55 F 30000 2 M
ST5324 32 F 28000 4 M
D06529 24 F 24000 2 C
GI4319 42 M 40000 8 M
EZ4531 34 M 42000 10 M
run;
proc print;
run;
```

La procédure `proc means` permet d'obtenir les statistiques les plus courantes en précisant les mots-clef en options (elle les calcule et les affiche), par exemple

- `n` : la taille de l'échantillon
- `mean` : la moyenne arithmétique
- `std` : l'écart-type
- `var` : la variance
- `min` : la valeur minimale de l'échantillon
- `max` : la valeur maximale de l'échantillon
- `range` : l'étendue
- `sum` : la somme

Comme avec la procédure `proc print`, il est possible de n'obtenir des statistiques que pour certaines variables à l'aide de la commande `var`. De même, à condition d'avoir trié les variables au préalable, il est possible de demander des statistiques triées par groupe à l'aide de la commande `by`.

Remarquons que la commande `class`, qui ne demande pas de tri préalable, renvoie quasiment le même résultat que la commande `by`, mais en présentant les valeurs différemment.

Par exemple, pour afficher la moyenne et l'écart-type des salaires et des anciennetés pour les femmes d'un côté et pour les hommes de l'autre,

```
(...)  
proc sort;  
by genre;  
run;  
proc means mean std;  
var salaire anciennete;  
by genre;  
run;
```

Remarquons qu'il n'y a pas besoin de procédure `proc print`.

Exercice 1 Modifiez votre programme en demandant l'affichage de la moyenne et de l'écart-type des âges, salaires et anciennetés pour les femmes d'un côté et les hommes de l'autre. Affichez également la somme des salaires. Comparez l'affichage obtenu en utilisant les commandes `by` ou `class`. Testez également la procédure `means` sans options : quelles statistiques affiche-t-elle par défaut ?

La procédure `proc univariate`, utilisée sans options, permet d'obtenir encore plus de statistiques (valeurs extrêmes, médiane, centiles, distributions de fréquences, diagrammes, ...) et d'effectuer des tests statistiques.

L'option `plot` permet de dessiner un diagramme en feuilles, en boîte, et une courbe de probabilités normales cumulées, et l'option `freq` permet d'obtenir la liste des valeurs de la variable, la fréquence absolue, le pourcentage de ces valeurs et le pourcentage cumulé de ces valeurs.

Exercice 2 Testez cette procédure sur votre programme, avec les différentes options possibles.

II Corrélation linéaire

La procédure `proc corr` permet d'obtenir les coefficients de corrélation linéaire, éventuellement à l'aide des commandes `by` et `var`.

La procédure renvoie d'abord des statistiques de base pour chaque variable : taille de l'échantillon (`n`), moyenne arithmétique (`Mean`), écart-type (`Std Dev`), somme (`Sum`), valeur minimale (`Minimum`), valeur maximale (`Maximum`) puis les valeurs des coefficients de corrélation linéaire (valeurs du haut).

On peut éliminer les statistiques de base pour se restreindre aux coefficients de corrélation linéaire en ajoutant les options `noprob` et `nosimple` et demander l'affichage des covariances à l'aide de l'option `cov`.

Exercice 3 Quelle est la valeur du coefficient de corrélation linéaire entre l'ancienneté et le salaire pour les femmes ? Entre l'âge et le salaire ? Même question pour les hommes.

La commande `with` permet d'obtenir des coefficients de corrélation linéaire pour certains couples de variables seulement. Ainsi, l'instruction suivante donne les coefficients de corrélation linéaire entre les variables `a` et `c` et entre les variables `b` et `c` :

```
proc corr;
var a b;
with c;
run;
```

Exercice 4 Utilisez `with` afin d'obtenir les résultats de l'exercice 3 à l'aide d'un seul appel à `proc corr`.

III Distribution de fréquences

La procédure `proc freq` permet d'afficher les fréquences et fréquences cumulées précisées à l'aide de la commande `tables` :

```
proc freq;
tables anciennete etatcivil;
run;
```

Il est possible de préciser l'ordre souhaité en option :

```
proc freq;
tables anciennete;
run;
proc freq order=data;
tables anciennete;
run;
proc freq order=freq;
tables anciennete;
run;
```

Exercice 5 Faites afficher les fréquences concernant l'ancienneté des employés de l'entreprise. Quelle est la différence entre les trois écritures ?

Exercice 6 Faites afficher les fréquences concernant les salaires des employés en séparant les résultats pour les hommes et les femmes.

IV Graphiques

La procédure `proc chart` permet d'afficher 5 types de graphiques différents : histogramme (`vbar`), diagramme à barres horizontales (`hbar`), diagramme en blocs (`block`), diagramme à secteurs circulaires (`pie`) et diagramme en étoile (`star`). Nous n'en décrivons que 3 ici.

1 Histogrammes

C'est la commande `vbar` qui permet d'obtenir des histogrammes. Elle s'utilise de la façon suivante :

```
proc chart;  
vbar variable;  
run;
```

et l'on peut rajouter des options ainsi :

```
proc chart;  
vbar variable / options;  
run;
```

Exercice 7 Dans le tableau suivant, chaque ligne est composée d'un numéro d'élève, de son genre, de son groupe et de ses trois notes :

01	M	01	7	9	12
02	F	03	12	14	11
03	M	02	13	17	12
04	M	02	15	16	11
05	F	01	8	11	12
06	M	03	11	14	11
07	F	01	12	13	11
08	M	01	6	10	11
09	F	02	10	11	11
10	F	03	5	7	11
11	F	03	13	8	11
12	M	02	19	18	10
13	F	01	17	16	11
14	M	02	14	18	11
15	M	01	4	9	11
16	F	03	12	14	11
17	M	02	13	15	13
18	F	02	14	9	11
19	F	01	2	3	11
20	M	01	10	7	11
21	F	01	14	15	10
22	M	02	20	18	11
23	M	02	12	15	11

Écrivez le programme SAS permettant de lire et faire afficher ces données, puis faites afficher l'histogramme de la variable correspondant à la première note.

Par défaut, l'échelle de l'axe vertical représente les fréquences absolues. Il est possible de demander l'échelle en pourcentages (`type = percent`), en fréquences absolues cumulées (`type = cfreq`), en pourcentages cumulés (`type = cpercent`). Si les données sont quantitatives, on peut aussi changer la valeur des points milieux avec une option du type `midpoints = 2 6 10 14 18`.

Exercice 8 Dans le programme précédent, faites afficher l'histogramme, mais cette fois-ci en fréquences cumulées et avec les entiers pairs comme points milieux.

Si l'on souhaite séparer dans chaque barre de l'histogramme les résultats en fonction d'une autre variable, on utilise l'option `subgroup = autrevariable`.

Exercice 9 Faites afficher l'histogramme des notes, en séparant pour chaque les filles et les garçons.

Enfin, si vous souhaitez faire afficher un histogramme par modalité d'une autre variable, vous pouvez utiliser l'option `group = autrevariable`.

Exercice 10 Faites afficher trois histogramme des notes, un par groupe. Séparez ensuite dans les histogrammes obtenus les filles et les garçons dans chaque barre.

Un autre type possible est `type = mean sumvar = autrevariable`, qui est un peu surprenant au premier abord. L'échelle devient alors la moyenne de la variable `autrevariable`. C'est très pratique pour faire afficher par exemple la moyenne d'une variable par groupe. On utilise alors :

```
proc chart;  
vbar groupe / type=mean sumvar=variable;  
run;
```

Notez que la variable n'est plus précisée juste après `vbar`, mais bien juste après `sumvar`.

Exercice 11 Comment modifier votre programme pour faire afficher un histogramme à deux colonnes, la première donnant la moyenne des filles et la seconde celle des garçons ? Comment faire afficher trois tels histogrammes, chacun correspondant à un groupe ?

2 Diagrammes circulaires

La commande `pie` permet d'afficher un diagramme circulaire. Elle prend les mêmes options que la commande `vbar` (à l'intérieur d'une procédure `proc chart`).

Exercice 12 En vous inspirant des programmes précédents sur les histogrammes, affichez un diagramme circulaire représentant la moyenne de la première note par groupe.

Exercice 13 Les ventes mensuelles d'une entreprise se répartissent de la façon suivante :

Mois	J	F	M	A	M	J	J	A	S	O	N	D
Ventes	75	67	55	45	46	47	87	34	74	56	46	77

Écrivez le programme SAS permettant de lire et faire afficher ces données, puis représentez-les sous la forme d'un diagramme circulaire.

3 Diagrammes en étoile

La commande `star variable / sumvar autrevariable` permet d'obtenir un diagramme en étoile de la variable `autrevariable` en fonction de la variable `variable`. Un tel diagramme est très similaire à un histogramme, sauf qu'il ressemble à une toile d'araignée.

Exercice 14 Représentez le diagramme en étoile des ventes de l'exercice 13.

4 Avec la procédure gplot

La procédure `gplot` (successeur de `proc plot`) permet d'afficher graphiquement des données numériques. Par exemple, pour obtenir un nuage de points (avec `variable1` en abscisses et `variable2` en ordonnées) :

```
plot variable2*variable1;  
run; quit;
```

Exercice 15 Reprenez les données de l'exercice 7 et faites afficher la deuxième note en fonction de la première.

On peut choisir l'aspect des points à l'aide de l'instruction

```
symbol1 v=marqueur i=graphique c=couleur; /* ou value, interpol, color */
```

où `marqueur` peut prendre (entre autres) les valeurs `plus`, `x`, `star`, `square`, `diamond`, `triangle`, `point`, `dot`, `circle`, `none`, `graphique` peut valoir `none` (nuage de points), `join` (points reliés dans l'ordre de la table), `spline` (courbe lissée) ou `rl` (avec une droite de régression linéaire) et enfin `couleur` représente le nom d'une couleur (en anglais). Notez qu'on peut définir `symbol1`, `symbol2`, ..., `symbol99`.

Exercice 16 Testez ces options sur le graphique de l'exercice précédent.

Afin de faire afficher plusieurs variables en ordonnées en fonction de la même variable en abscisses,

```
plot (variable2 variable3)*variable1 / legend overlay;
```

Exercice 17 Représentez les deux nuages de points des deuxièmes et troisièmes notes en fonction de la première. Le premier nuage sera symbolisé par des triangles bleus et le deuxième par des carrés rouges.

Pour préciser les axes, on peut les définir au préalable :

```
axis1 label=("Nom de l'axe des abscisses" justify=right);  
axis2 label=("Nom de l'axe des ordonnées") order=(0 to 1000 by 100);
```

Le `label` permet de donner un nom à l'axe et le `order` de faire figurer l'échelle sur l'axe. Pour utiliser les axes ainsi définis, on précise alors l'appel au graphique avec :

```
plot variable2*variable1 / haxis=axis1 vaxis=axis2;
```

On peut aussi ajouter une grille (quadrillage) avec `grid`.

Attention, SAS Studio ne dispose pas de la procédure `proc gplot` : on peut utiliser `proc sgplot`. Voici les équivalents des principales options vues ci-dessus :

```
proc gplot data=table3;  
/* proc sgplot data=table3; */  
  
axis1 label=("Nom de l'axe des abscisses");  
axis2 label=("Nom de l'axe des ordonnées") order=(0 to 1000 by 100);  
/* xaxis label="Nom de l'axe des abscisses"; */  
/* yaxis label="Nom de l'axe des ordonnées" values=(0 to 1000 by 100); */  
  
symbol1 value=star;  
plot variable2*variable1;  
/* scatter x=variable1 y=variable2 / markerattrs=(symbol=star); */
```

Exercice 18 Quelles sont les erreurs dans ce programme ? Comment le corriger ? (N'hésitez pas à le tester pour le corriger au fur et à mesure.)

```
data;  
input age genre docteur statut;  
18 M Lebrun 01  
18 F Riviere 01  
19 F Lebrun 02  
...  
15 F Lebeau 03  
proc chart by genre;  
    vbar statut;  
    vbar age;  
proc gplot;  
    docteur by statut;
```

Exercice 19 On reprend les données suivantes, déjà utilisées dans la première fiche (identifiant, âge, genre, salaire, ancienneté et état-civil d'une part, et identifiant, département d'origine et nombre d'enfants d'autre part).

```
AD4632 42 M 42000 10 M  
HF4523 39 F 38000 12 C  
SD4510 47 F 35000 14 M  
DE4512 24 M 28000 2 M  
S05601 43 M 38000 15 M  
CK4539 62 M 45000 28 C  
VK3461 27 M 44000 4 C  
S04218 55 F 30000 2 M  
ST5324 32 F 28000 4 M  
D06529 24 F 24000 2 C  
GI4319 42 M 40000 8 M  
EZ4531 34 M 42000 10 M}  
  
AD4632 77 3  
HF4523 86 2  
SD4510 45 1  
DE4512 73 0
```

S05601 62 1
 CK4539 13 4
 VK3461 83 1
 S04218 49 2
 ST5324 29 2
 D06529 14 0
 GI4319 75 3
 EZ4531 92 1

1. Commencez par fusionner (horizontalement) les deux tables afin d'obtenir un unique tableau pour les employés comportant leur identifiant, leur âge, leur genre, leur salaire, leur ancienneté, leur état-civil, le nombre de leurs enfants et leur département d'origine.
2. Faites afficher graphiquement le salaire en fonction de l'ancienneté. Testez l'effet des différentes options présentées ci-dessus. Adaptez en particulier les échelles pour les axes.
3. Faites afficher le nombre d'enfants en fonction de l'âge.

Exercice 20

1. Créez une table SAS avec les données suivantes :

Nb enfants	0	1	2	3	4	5	6
Effectif	235	183	285	139	88	67	3

2. Ajoutez une nouvelle variable représentant les fréquences relatives (il y a 1000 couples interrogés).
3. Créez une nouvelle table à partir de la première (avec `data / set`) triée par effectif croissant.
4. On peut créer une nouvelle variable qui se calcule de proche en proche en utilisant `retain`. Par exemple, pour les effectifs cumulés :

```
retain EffectifCumule 0;
```

```
EffectifCumule = EffectifCumule + Effectif;
```

Le plus simple pour attribuer un nom à cette nouvelle colonne est alors d'utiliser `attrib` :

```
attrib EffectifCumule label="Effectifs cumulés";
```

Ajoutez deux colonnes à votre nouvelle table, une pour les effectifs cumulés, l'autre pour les fréquences cumulées.

5. Tracez alors la courbe des fréquences cumulées en adaptant les axes. Testez aussi `interpol=step`.

V Exercices

Exercice 21 Une machine automatique fabrique des pièces métalliques de différentes largeurs, qui peut être réglée à l'aide d'un ajustement. Les pièces fabriquées par la machine sont utilisées dans un autre atelier et depuis quelques temps, le contremaître de cet atelier se plaint du manque d'uniformité dans la largeur des pièces, causant ainsi un taux de rejet élevé du produit fini. Selon les normes, ces pièces doivent avoir une largeur de 12 mm plus ou moins 0,5 mm. On décide donc d'analyser 60 pièces consécutives fabriquées par la machine. Les résultats sont regroupés dans le tableau suivant :

10.1 10.1 11.2 12.1 10.3 13.1 10.0 13.0 10.1 10.3 10.0 11.0 11.3 13.3 13.2
 13.7 12.8 12.3 14.0 10.7 10.7 13.0 14.4 10.4 14.1 12.2 10.0 10.2 10.5 10.3
 14.0 12.6 10.2 11.4 12.0 11.2 13.8 10.5 10.1 10.7 13.3 10.9 10.8 14.8 14.0
 13.5 14.4 12.1 10.5 10.9 10.6 10.5 13.3 11.7 14.1 15.2 12.3 14.7 10.9 10.1

1. Écrire le programme SAS qui permet de lire et afficher ces données.
2. Afficher les fréquences et fréquences cumulées de ces largeurs.
3. Tracer l'histogramme correspondant.
4. Afficher la moyenne et l'écart-type de ces largeurs. Donner ces valeurs.

Exercice 22 Vingt familles ont répondu à une enquête : la première question leur demandant si le couple était propriétaire, la deuxième le nombre d'enfants, la troisième le revenu et enfin la dernière l'épargne mensuelle. Les résultats sont regroupés dans le tableau suivant :

1	oui	3	40200	1700
2	oui	1	39800	3500
3	oui	3	32200	900
4	oui	1	31800	1800
5	oui	1	40900	2900
6	non	4	23800	300
7	non	5	32000	160
8	non	4	30500	80
9	non	2	31900	1500
10	non	1	29900	1900
11	oui	4	32000	480
12	non	3	20800	860
13	oui	1	27300	680
14	non	2	32000	820
15	non	2	27000	2040
16	oui	2	24000	450
17	oui	3	40200	510
18	oui	2	21900	1170
19	oui	3	25300	1700
20	oui	3	32600	700

1. Quel est le revenu moyen et son écart-type suivant si la famille est propriétaire ou non ?
2. Même question suivant le nombre d'enfants.
3. Quel est le coefficient de corrélation entre le revenu et l'épargne ?
4. Représenter le revenu sous forme d'histogramme, en séparant chaque barre en deux (les propriétaires et les autres).
5. Tracer un histogramme représentant le revenu moyen des familles : faites figurer en abscisses le nombre d'enfants, et pour chaque nombre séparer les valeurs pour les hommes et les femmes.

VI Exercices supplémentaires

Exercice 23 Un questionnaire canadien a servi à recueillir de l'information concernant l'intérêt, pour la population d'une grande ville, d'obtenir une équipe professionnelle de la ligue nationale de hockey (LNH). Les résultats, sur un échantillon de 60 personnes, sont regroupés dans le tableau suivant :

4	M	42	M	4	5	3
4	F	47	M	6	6	4
4	M	42	C	3	2	2
5	M	58	M	1	2	6
4	F	48	M	3	6	3
4	F	44	M	4	6	4
4	F	55	M	3	1	2
1	F	20	C	2	9	2
2	M	32	M	1	1	0
3	M	23	M	1	2	0
5	M	24	M	1	5	9
2	M	29	M	3	7	0
4	F	42	M	3	6	2
3	F	39	M	1	6	0
5	F	40	M	1	5	7
3	F	31	M	1	6	1
5	M	62	M	1	5	6
3	M	36	M	2	7	1
4	F	49	M	3	4	2
3	M	36	C	1	2	0
4	M	38	M	5	1	3
4	F	49	M	2	6	1
3	M	42	M	2	8	1
1	F	38	M	1	1	0
1	M	34	M	5	3	1
3	M	32	M	1	6	0

4	F	44	M	2	1	2
2	M	26	C	2	2	2
4	F	39	C	3	6	2
1	F	21	M	1	7	0
5	M	47	M	5	1	4
1	F	23	M	4	2	3
4	M	39	M	8	6	5
1	M	23	C	2	7	0
3	F	46	M	1	2	1
5	M	37	M	2	8	8
5	F	52	M	1	2	6
2	M	32	M	3	3	0
3	M	24	C	4	7	1
2	F	31	M	2	2	0
1	F	20	C	1	3	1
5	M	48	C	4	5	3
2	F	23	C	1	7	1
2	F	22	C	2	3	0
1	M	22	C	2	3	1
2	F	24	C	3	1	2
5	F	54	M	6	5	4
2	M	24	C	4	2	0
5	M	59	M	1	5	7
4	M	42	M	5	6	4
3	F	37	C	1	2	0
1	M	30	C	1	2	0
4	F	23	M	4	2	3
2	M	41	M	4	2	3
4	M	47	M	7	1	5
2	F	27	C	1	7	0
3	F	38	M	1	6	0
1	F	28	C	1	3	0
3	F	30	M	2	2	1
2	M	26	C	1	2	1

Les questions étaient les suivantes : intérêt (1 très peu intéressé, 2 peu intéressé, 3 indifférent, 4 intéressé, 5 très intéressé), genre, âge, état-civil, nombre de fois où la personne irait encourager leur équipe, nombre d'années depuis lesquelles la personne suit les activités de la LNH, nombre de fois où la personne est allée encourager l'équipe de la ville voisine au cours de la dernière saison.

1. Créer le programme permettant de lire et imprimer ces données.
2. Modifier ce programme pour identifier plus clairement les colonnes (au lieu d'imprimer le nom des variables).
3. Modifier ce programme pour écrire les résultats sous une forme plus lisible (par exemple remplacer 1 par "très peu intéressé" pour l'intérêt).
4. Modifier ce programme pour faire afficher l'âge, l'intérêt, l'état-civil et le nombre de fois où la personne irait encourager son équipe, en ordonnant les données par intérêt décroissant.
5. Créer une variable « catégorie d'âge » (moins de 30 ans, entre 30 et 40, entre 40 et 50, entre 50 et 60, plus de 60) et afficher la même chose qu'à la question précédente mais en séparant par catégorie.
6. Donner les statistiques suivantes :
 - l'âge moyen des hommes et des femmes qui ont répondu au sondage ;
 - le pourcentage de personnes mariées et célibataires parmi les personnes ;
 - le nombre de fois, en moyenne, que les hommes vont encourager leur équipe ; même question pour les hommes mariés ; les hommes célibataires ; les femmes ;
7. Représenter sur un histogramme l'intérêt des personnes : en abscisse l'intérêt (avec deux barres pour chaque valeur en séparant femmes et hommes), et en ordonnée le nombre de personnes concernées.
8. Donner les coefficients de corrélation linéaire entre les différentes variables quantitatives et commenter.
9. Donner les pourcentages des personnes pour chaque type d'intérêt.
10. Donner l'âge moyen des personnes pour chaque type d'intérêt.
11. Représenter l'intérêt à l'aide d'un diagramme à secteurs circulaires.

Exercice 24 On reprend les données de l'exercice 23 de la première fiche. Elles regroupent les salaires d'un groupe

de personnes dans une entreprise (année d'entrée, identifiant, âge, genre, salaire, état-civil) :

```
2009 ad09 35 F 34000 M
2008 sd08 34 M 33000 C
2007 hs07 32 M 37000 M
2007 xc07 28 F 36000 M
2004 fd04 39 F 42000 M
2008 fs08 27 F 36000 C
2010 qs10 25 M 34000 C
2011 jl11 24 F 33000 C
2011 am11 23 F 32000 M
2011 cm11 25 M 32000 C
2005 er05 37 M 43000 M
2006 rt06 35 M 39000 M
2006 lm06 34 F 38000 C
2007 cd07 34 M 35000 C
```

1. Écrire le programme SAS qui permet de lire et afficher ces données.
2. (*) Afficher les valeurs des coefficients de corrélation linéaire entre le salaire et l'année d'entrée pour les femmes, puis entre le salaire et l'âge (et uniquement entre ces variables). Même question pour les hommes. Donner les valeurs de ces coefficients.
3. Faire afficher graphiquement le nuage de points du salaire en fonction de l'âge.

Exercice 25

Les données suivantes (exercice 24 de la première fiche) représentent les notes à trois contrôles d'une classe d'élèves de sixième (groupe de l'élève, genre, note 1, note 2, note 3, âge).

```
01 M 7 9 8 12
02 F 12 14 11 11
01 M 13 17 14 12
01 M 15 16 17 11
01 F 8 11 10 12
02 M 11 14 12 11
03 F 12 13 14 11
03 M 6 10 9 11
02 F 10 11 12 11
02 F 5 7 8 11
01 F 13 8 14 11
02 M 19 18 14 10
03 F 17 16 15 11
02 M 14 18 17 11
02 M 4 9 12 11
03 F 12 14 10 11
01 M 13 15 16 13
02 F 14 9 12 11
02 F 2 3 7 11
01 M 10 7 11 11
01 F 14 15 12 10
02 M 20 18 17 11
03 M 12 15 18 11
```

1. Écrire le programme SAS qui permet de lire et afficher ces données (la table sera nommée **eleves**).
2. (*) Afficher la moyenne, l'écart-type, la valeur minimale et la valeur maximale de la classe pour chaque note. Donner ces valeurs pour la deuxième note.
3. (*) Expliquer comment modifier le programme pour qu'il n'affiche ces valeurs que pour la deuxième note.
4. Afficher l'histogramme correspondant à la première note.
5. Modifier votre programme pour séparer dans chaque barre les filles et les garçons.
6. Représenter graphiquement (avec le graphique de votre choix) la moyenne générale par groupe.