IUP SID M1, 2005

T.P. SAS numéro 01

Après quelques généralités sur le logiciel SAS, l'objet de ce T.P. est de réaliser les premières manipulations avec ce logiciel. Avant cela, il faut mettre en place l'environnement UNIX nécessaire à l'utilisation de SAS dans des conditions commodes.

Généralités sur le logiciel SAS

Le logiciel SAS est de conception américaine : il est developpé et commercialisé par la société SAS-Institute, située à Cary, en Caroline du nord. Écrit en langage C, SAS (*Statistical Analysis System*) est, à l'origine, un logiciel de statistique polyvalent, c'est-à-dire susceptible de traiter pratiquement tous les domaines de la statistique. Il est assez ancien (ses débuts remontent aux années 1970) et est constamment enrichi de nouvelles méthodes. Par suite, il est très volumineux et souvent redondant : le même problème statistique peut être traité par différents modules du logiciel (avec souvent des présentations différentes !). Aujourd'hui, SAS est devenu un véritable système de gestion de l'information plutôt qu'un simple logiciel de statistique.

La "doc" (documentation) papier de SAS est "monstrueuse" : plusieurs centaines de volumes, certains dépassant les 1 000 pages... Bien entendu, elle n'existe qu'en anglais (on imagine le coût d'une traduction...). Toutefois, des cours polycopiés synthétiques existent en français. Par ailleurs, cette doc est aujourd'hui partiellement en ligne.

Les 6 volumes les plus utiles de la doc, pour un utilisateur de base de SAS, sont les suivants :
- *SAS language* : ce volume donne une description générale du logiciel, ainsi que des informations sur les commandes SAS, les sous-programmes standards, le principe général des *macros*...
- *SAS procedures guide* : donne une notice détaillée sur toutes les procédures statistiques de base contenues dans SAS ;
- *SAS/STAT user's guide, volumes 1 & 2* : donnent également une notice détaillée sur les procédures statistiques plus avancées ;
- *SAS/GRAPH software, volumes 1 & 2* : précisent la façon d'écrire des procédures SAS pour obtenir des graphiques élaborés.

Toutefois, la documentation la plus accessible dans le cadre de ces T.P., et à laquelle on se réfèrera en permanence dans toute la suite, est le cours polycopié suivant

SAS 8.2 sous UNIX (version de septembre 2001)

corédigé par J.M. Azaïs, P. Besse, H. Cardot, V. Couallier et A. Croquette (Laboratoire de Statistique et Probabilités, Université Paul Sabatier, Toulouse). Les renvois aux pages seront relatifs à ce cours polycopié. On peut se le procurer sur le site internet suivant :

http ://www.lsp.ups-tlse.fr/Besse/enseignement.html

Signalons, pour terminer, 3 ouvrages (en anglais) présentant, à différents niveaux, l'usage du logiciel SAS dans le traitement statistique de données.
- *How SAS works*, de P.A. Herzberg, Springer, 1990 (pour un niveau élémentaire).
- *Applied statistics and the SAS programming language*, de R.P. Cody et J.K. Smith, third edition, Prentice Hall, 1991 (pour un niveau plus avancé).
- *A handbook of statistical analyses using SAS*, de B.S. Everitt et G. Der, Chapman and Hall, 1996 (également pour un niveau plus avancé).

Les fenêtres principales de SAS

Se connecter au CICT via la machine ondine. Créer un répertoire spécifique dans lequel on se placera systématiquement pour lancer SAS et enregistrer les différents fichiers (on pourra par exemple l'appeler TPSAS).
Entrer dans SAS en faisant la commande `sas &` (la version actuelle est la version 8.2). Une fois connecté à SAS, plusieurs fenêtres s'ouvrent à l'écran (pages 11–12).
Les 3 plus utiles sont :

- *SAS : Program Editor* : c'est l'éditeur de texte de SAS, dans lequel on doit entrer tout programme à exécuter ; des rudiments sur sa manipulation se trouvent page 12 ;
  Il est également déconseillé d'éditer directement les programmes dans la fenêtre SAS : Program Editor*,*
  l'éditeur de texte SAS n'est pas très performant. Utilisez plutôt un éditeur de texte dont vous avez l'habitude, quitte à "draguer" avec la souris votre programme pour le transporter dans la fenêtre "Program Editor".
- *SAS : Log* : fenêtre dans laquelle s'affichent, au cours de l'exécution d'un programme, le programme lui-même, séquence par séquence (en noir) et les commentaires du système SAS sur ce programme (en bleu) ; le cas échéant, s'affichent également ici un message d'avertissement, lorsqu'un problème non fatal est détecté (précédé de *warning*, en vert) ou un message d'erreur, lorqu'une erreur fatale est détectée (précédé de *error*, en rouge) ; (noter que *to log* signifie, en anglais, enregistrer, noter sur un registre) ;
- *SAS : Output* : fenêtre dans laquelle s'affichent tous les résultats obtenus à l'issue d'un programme (lorqu'il a marché !).

Ces 3 fenêtres possèdent sensiblement les mêmes "menus déroulants" sur leur partie haute. Il est vivement conseillé de les disposer de façon commode à l'écran (en gardant également accessible la fenêtre UNIX).
Figurent également deux autres fenêtres que vous pouvez fermer : elles ne sont pas utiles lors d'un premier apprentissage.

On quitte SAS en se plaçant dans la fenêtre *program editor* et en utilisant le menu déroulant `File/Exit`.

Les 2 possibilités d'utilisation de SAS en mode interactif

Il existe diverses façons de faire du traitement de données avec SAS (pages 11-13). Les 2 façons de le faire en mode interactif sont indiquées ci-dessous.

- *Programmation SAS.* Cela consiste à écrire un programme SAS et à lancer son exécution par le menu déroulant `Run/Submit` qui apparait en tête de la fenêtre "Programm editor". C'est essentiellement de cette façon que nous procèderons dans le cadre de ces T.P. Un programme SAS est une succession de procédures, chacune réalisant un traitement statistique homogène ou un graphique.
- *SAS/INSIGHT* (menu déroulant `Solutions/Analysis/Interactive Data Analysis`). Permet un traitement interactif immédiat et puissant des données ; de nombreuses méthodes sont disponibles et on peut réaliser des graphiques très élaborés. SAS/INSIGHT sera abordé dans les dernières séances de T.P.

Notion de table SAS

Un fichier de données ne peut être reconnu, lu et traité par SAS que s'il est dans un format spécifique. De même, les fichiers produits en sortie d'une procédure SAS seront dans ce format spécifique. Nous appellerons *table SAS* (traduction officielle de *SAS data set*) un tel fichier mis dans un tel format (page 9).

Articulation d'un programme SAS

Un programme SAS comprend des entrées-sorties (lectures et écritures de données) et des enchaînements de procédures.

Gestion des données avec SAS

Il existe 2 possibilités pour lire un fichier de données dans un programme SAS. Tout d'abord, il est possible de lire directement les données, en les incluant dans le programme, au moyen de la commande `cards` . Cette façon de procéder, peu commode, n'est pas recommandée. Ensuite, on peut lire des données préalablement enregistrées dans un fichier ASCII (extérieur à SAS), au moyen de la commande `infile` . Dans un cas comme dans l'autre, une déclaration des variables est obligatoire, au moyen de la commande `input`. Enfin, noter que les données ne seront réellement utilisables qu'une fois transformées en table SAS, ce qui se fait par la commande `data`, suivie du nom que l'on souhaite donner à cette table ; la commande `data` doit être placée en début de séquence, avant même la lecture des données. Ainsi, une séquence de lecture des données se présente en général de la façon suivante :

```
data <nom de la table SAS>;
infile '<nom du fichier des données>';
input <liste des variables>;
run;
```

Noter que, dans la liste des variables, le séparateur est un blanc.

Par ailleurs, chaque procédure SAS réalisant un traitement statistique produit un certain nombre de résultats. Ces résultats sont soit affichés dans la fenêtre *output*, soit enregistrés dans une table SAS particulière. Dans ce dernier cas, la procédure `print` permet d'afficher le contenu de cette table SAS dans la fenêtre *output*. On peut ensuite archiver le contenu de la fenêtre *output* dans un fichier ASCII en utilisant le menu déroulant `File/Save as`.

Les procédures SAS

Un programme SAS est en fait un enchaînement de procédures, chacune réalisant un traitement homogène. Les procédures de base sont répertoriées dans le volume *SAS procedures guide*. En dehors de la procédure `print`, déjà citée, les principales procédures sont les suivantes (pages 33–36) :

- *means, univariate* : servent à la description élémentaire de variables quantitatives (nombre d'observations, minimum, maximum, moyenne, écart-type...) ; *univariate* est plus élaborée ;
- *freq* : sert à la description élémentaire de variables qualitatives (effectifs, fréquences... ; permet aussi de croiser 2 ou plusieurs variables, de déterminer des profils, de calculer des khi-deux...) ;
- *plot* : réalise le nuage de points relatif à 2 variables quantitatives ;
- *chart* : réalise différents graphiques pour une variable qualitative ;
- *sort* : range le fichier selon les valeurs croissantes ou décroissantes d'une variable quantitative spécifiée par `by` ;
- *rank* : calcule une variable "rang" pour chaque variable quantitative déclarée (déclaration obligatoire) ;
- *standard* : permet de déterminer les valeurs centrées et réduites associées à une variable quantitative donnée (nécessite les options *mean = 0* et *std = 1*) ;
- *corr* : permet de calculer la matrice des corrélations (ainsi que la matrice des variances-covariances) d'un ensemble de variables quantitatives.

Exemple : `proc univariate; run;`

Options, title, footnote et commentaires

Diverses commandes générales peuvent être rajoutées au début d'un programme SAS.

- *Options* : il est ici question des options générales, à mettre en début de programme ; on les déclare avec la commande `options` (le s est facultatif) ; citons : `pagesize`, qui spécifie le nombre de lignes dans une page de sortie (*output*) ; `linesize`, qui spécifie le nombre de caractères par ligne ; `nodate`, qui supprime l'impression de la date dans les sorties.
- *Title (titre)* : permet de placer un titre en haut de chaque page des sorties ; exemple : `title 'ceci est un titre';`

– *Footnote (pied-de-page)* : permet de placer un titre en bas de chaque page des sorties ; permet également une meilleure mise-en-page des sorties SAS sur une imprimante ; exemple :
`footnote 'ceci apparaitra en bas de page';`
– *Commentaires* : des commentaires peuvent être insérés n'importe où dans un programme SAS ; ils doivent être placés de la façon suivante :
`/* ceci est un commentaire */` (surtout commode au sein d'une ligne de commande)
`* ceci en est`
`un autre;`

Exercice 1

1. Connectez-vous au CICT via la machine ondine. Lancez le logiciel `sas&`. Vous voyez apparaître les fenêtres EDITOR, RESULTS, LOG et EXPLORE.

2. Dans la fenêtre "SAS : Program Editor" entrez le programme suivant (noter que la commande `input` précède ici la commande `cards`) :

```
data tp1; /*creation d'une table provisoire*/
input taille poids age sexe $;
cards;
174 65 35 m
169 56 28 f
166 48 30 f
181 80 27 m
168 53 26 f
176 76 34 m
190 77 32 m
159 70 31 f
162 60 25 f
164 51 22 f
160 73 21 f
;
run;
proc print;
run;
```

À cette occasion, utilisez les fonctionnalités de la fenêtre *programm editor* (voir page 12).

3. Sauvegardez le fichier `file> save as > progtp1.sas`

4. Exécutez le programme `run > submit` ou la touche F3.

On constate que le programme disparaît de la fenêtre SAS : Program Editor. Afin de le rappeler pour des modifications éventuelles, `run > recall last submit` ou la touche F4.

Pour sauvegarder après avoir modifié : `file > save`

5. Pour vérifier le contenu de la table créée :

```
tools > table editor
file > open > work > TP1 > open
```

On remarquera que ce menu peut servir non seulement à la visualisation des tables existantes mais aussi à la création de nouvelles tables.

6. Exécuter SAS/ASSIST qui permet de générer des exemples de programmes que l'on peut enregistrer et utiliser en adaptant à nos besoins.

```
solution > ASSIST
```

... répondre aux questions
```
graphics > pie chart
table > work > TP1
chart column > sexe
run >submit
```

Pour récupérer le programme créé, allez dans la fenêtre SAS : Editor et faîtes `recall last submit`.

Remarques

1. SAS ne différencie pas les minuscules et les majuscules.

2. Ne pas oublier les ; à la fin de chaque instruction.

3. Si après avoir exécuté le programme vous obtenez un résultat bizarre, vérifiez la fenêtre SAS :Log.

4. Le signe "dollar" dans la déclaration des variables indique que la variable précédant `$` est qualitative.

5. Dans un programme SAS, tout ce qu'on écrira entre /* et */ ne sera pas pris en compte pendant l'exécution du programme. Ceci sert à commenter différentes parties du programme.

6. Sauvegardez toujours votre programme avant de le soumettre.

Personnalisation

1. Création d'une librairie de travail.

Par défaut, les données entrées dans SAS sont enregistrées dans la librairie WORK qui est effacée à chaque fois que l'on quitte le logiciel. Pour conserver les données, il est recommandé de créer une librairie (un répertoire) qui sera conservée entre différentes sessions SAS. Pour cela, il suffit de créer un répertoire dans l'invite de commande de ondine (celle où vous avez tapé `sas&`). La commande est la suivante : `mkdir TPSAS` .

Pour prendre en compte ce changement, il faut alors modifier le programme `progtp1.sas` en le précédant de la ligne
```
libname TPSAS '?/TPSAS';
(? désigne le chemin qui mène au répertoire. Il est de la forme '/home/...')
```

Cette étape n'est réalisée qu'une fois au début de la session. Pour indiquer à SAS que l'on souhaite enregistrer les données dans cette librairie il faut précéder le nom des données par "TPSAS." (à chaque création de données). Ainsi la ligne `data...` du programme `progtp1.sas` deviendra :
```
data TPSAS.tp1 ;
```

Si, au cours d'une session, le préfixe "TPSAS." est omis, les données seront enregistrées dans la librarie par défaut WORK.

2. Il est souvent difficile de lire le rapport d'erreur qui apparaît dans la fenêtre LOG lorsque les rapports s'accumulent. Pour eviter ce problème on peut simplement faire
```
Edit > Clear All
```
dans la fenêtre LOG.

3. Enfin, vous aurez sans doute remarqué le nombre grandissant de fenêtres. Pour retrouver vos fenêtres, on peut choisir dans le menu
```
View > {Editor, Log, Output}
```

Exercice 2

L'objet de cet exercice est de se familiariser avec les procédures de base de SAS relatives aux variables quantitatives et de réaliser les premiers graphiques. Le support est constitué des chapitres 2 et 3 du cours polycopié.

Dans tous les exercices considérés, on réutilisera la table tp1.

Commencez par entrer les commandes :

```
options pagesize=64 linesize=78 nodate;
title;
footnote 'TP1';
```

1. Pour calculer les caractéristiques statistiques les plus élémentaires (moyenne, écart type, variance, min, max, ...) d'une variable, on peut utiliser la procédure `means`. On ne la testera pas car la procédure `univariate` traitée ci-après est plus générale. Pour l'illustrer, saisissez et exécutez le programme suivant :

```
proc univariate data=tp1 normal plot;
var taille;
run;
proc print;
run;
```

On remarque que les mots `normal` et `plot` qui suivent la procédure `univariate` sont des options. La première permet d'obtenir des tests de normalité, alors que la seconde dessine des graphiques.

2. La procédure `sort` permet de trier les données ; elle range par défaut les données quantitatives en ordre croissant et les données qualitatives en ordre alphabétique. Afin d'obtenir l'ordre inverse, il faut intercaler l'option `descending` après `by`. Pour tester cette procédure, saisissez et exécutez le programme suivant :

```
proc sort data=tp1;
by taille;
run;
proc print;
run;
```

3. La procédure `freq` permet de faire un tri à plat de chacune des variables.

```
proc freq data=tp1;
run;
proc print;
run;
```

4. La procédure `plot` permet de dessiner des graphiques en basse résolution de nuages de points en deux dimensions.

```
proc plot data=tp1;
plot taille*age='*';
run;
```

Dans le cas où on a plus de deux variables quantitatives, par exemple taille, poids et age, on peut demander dans un seule commande les graphiques des nuages de points taille*poids et poids*age. Cela se fait comme suit :

```
plot taille*age='*' poids*age='-';
```

et si l'on veut les superposer

```
plot taille*age='*' poids*age='-' /overlay;
```

5. La procédure `chart` permet de réaliser des graphiques relatifs à deux variables considérées. On utilisera la commande `vbar` pour obtenir un diagramme en colonnes, `hbar` pour obtenir un diagramme en barres horizontales et `pie` pour obtenir un diagramme en secteurs. Observer l'effet produit par les options `/type=percent`, `/type=cfreq` et `/type=cpercent` dans ces différents graphiques.

Les procédures `gplot` et `gchart` permettent d'obtenir le même genre de graphiques mais en haute résolution.

IUP SID M1, 2005

T.P. SAS numéro 02

1 Etude de la normalité d'un échantillon

Dans de nombreux exercices nous avons fait l'hypothèse agréable que l'échantillon suivait une distribution normale. En réalité, dans la pratique, on ne dispose pour le processus que des valeurs d'un échantillon et non de son modèle mathématique décrivant la variabilité. Donc la question naturelle qui se pose est : "Quel est le modèle de distribution le plus adapté aux données ?" Chaque processus de fabrication présente un modèle de variation particulier. Ce modèle peut être déterminé à l'aide de méthodes statistiques. Pour sélectionner la loi de distribution qui ajuste au mieux les données, on exploite l'information contenue dans l'échantillon en tenant compte de la caractéristique de qualité que l'on cherche à étudier.
La procédure de sélection est déclinée généralement en deux étapes. Tout d'abord, on exploite des méthodes statistiques empiriques afin de s'orienter vers une ou des familles de distributions les mieux adaptées. Ensuite, on emploie les tests de vérification d'hypothèses afin qu'une décision formelle soit prise.

1.1 Méthodes empiriques pour la normalité d'une distribution

Les méthodes empiriques ne se substituent pas aux tests formels pour la vérification de la normalité, mais elles fournissent des informations supplémentaires sur la loi de distribution de l'échantillon. On inclut dans les méthodes empiriques la forme de l'histogramme, la vérification des valeurs du coefficient d'asymétrie et d'aplatissement et le graphique de normalité.

• La forme de l'histogramme

La distribution normale est symétrique à une courbe de densité de probabilité en forme de cloche. L'histogramme des fréquences estime la courbe de densité de probabilité. Une forme symétrique de l'histogramme suggère de poser l'hypothèse de normalité tandis qu'une forme asymétrique suggère d'utiliser un modèle de distribution asymétrique.

• Les valeurs des coefficients d'asymétrie et d'aplatissement

Le *coefficient d'asymétrie de l'échantillon*, noté g_2, est une mesure de la symétrie de la répartition des valeurs d'une variable. Il est nul lorsque la distribution est symétrique ; il est négatif lorsque la distribution est asymétrique à gauche et positif lorsque la distribution est asymétrique à droite. Il est donné par :

$$g_2 = \frac{m_3}{S^3}$$

où m_3 est le moment d'ordre 3 de l'échantillon : $m_3 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3$.

Le *coefficient d'aplatissement de l'échantillon*, noté g_3, est comme son nom l'indique une mesure de l'aplatissement de la distribution d'une variable. Il est donné par :

$$g_3 = \frac{m_4}{S^4} - 3$$

où m_4 est le moment d'ordre 4 de l'échantillon : $m_4 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4$.

Les valeurs du coefficient d'asymétrie et d'aplatissement d'une variable normale sont 0 et 3 respectivement. On vérifiera sur l'échantillon que les valeurs empiriques de ces coefficients sont proches des valeurs théoriques.

• **Le graphe de normalité : droite de Henry**

Lorsque l'on trace la fonction de répartition d'une loi normale sur un papier à échelle fonctionnelle gaussienne pour l'axe des fréquences, la courbe de répartition est représentée par une droite. Pour un échantillon de taille n, on vérifiera l'adéquation de la loi normale aux données, en comparant simplement le polygone des fréquences cumulées à une droite, sur un papier gaussien. Le mode opératoire pour la construction de la droite de Henry est simple :

1. Classer les observations par ordre croissant :

$$x_1 \leq x_2 \leq ... \leq x_n$$

2. Estimer les fréquences cumulées correspondantes :

$$F(x_i) = \frac{i}{n+1}$$

3. Placer sur un papier d'échelle gaussienne les points :

$$(x_i, F(x_i))$$

4. Tracer si cela semble possible, la droite qui ajuste au mieux le nuage de points. Si l'ajustement paraît convenable, la normalité est acceptée.

   Différentes estimations des fréquences cumulées, qui sont pratiquement équivalentes, peuvent être utilisées. Le graphique de normalité nécessite moins d'observations que l'histogramme des fréquences. On obtient de bons résultats dès que $n > 10$. Il convient d'utiliser des échantillons de plus grande taille, qui donnent plus d'informations sur la nature de la distribution de la caractéristique étudiée.

1.2 Les tests d'ajustements

Plusieurs méthodes statistiques sont fondées sur l'hypothèse que les échantillons proviennent de populations normales. La loi de distribution normale est un modèle mathématique pour la distribution de la caractéristique de qualité étudiée dans la population. Pour un échantillon donné, il est donc nécessaire de décider formellement si cette supposition est légitime ou pas. Etant donné que l'on dispose uniquement des valeurs d'un échantillon de la population, on ne sera jamais sûr que l'hypothèse de normalité est justifiée ou non. Tout ce que l'on peut faire est de tester l'hypothèse de normalité et selon les résultats du test, décider de garder ou non cette hypothèse.
Quand on teste la normalité des données, on part de la supposition que l'échantillon provient d'une population normale. Ensuite on vérifie si les observations s'accordent ou non avec cette hypothèse. La vérification est effectuée à l'aide d'une statistique test appelée aussi variable de décision, qui est calculée à partir des valeurs observées. La statistique test compare la forme de la distribution de l'échantillon avec la distribution normale.
A l'issue de cette comparaison, une décision est prise dépendant de la valeur statistique test. Soit on rejette l'hypothèse que l'échantillon provient de populations normales, soit on l'accepte. Le non rejet de l'hypothèse de normalité conduit à utiliser la loi normale pour les analyses statistiques ultérieures. Les tests d'ajustement se regroupent en deux catégories ; les tests d'ajustement spécifiques et les tests généraux. Les tests d'ajustement spécifiques sont conçus spécialement pour l'ajustement d'une famille de distribution spécifique, par exemple la distribution normale, et ils ne peuvent être utilisés que pour vérifier si l'échantillon provient ou non de cette famille de distribution. Les tests généraux peuvent être utilisés pour l'ajustement de n'importe quelle loi de distribution connue $F_0(x)$.

1.2.1 Les test de normalité

Pour un échantillon $X_1, ..., X_n$, la question qui se pose dans la pratique est la suivante : "Peut-on considérer que l'échantillon provient d'une population normale ?" Quand on teste la normalité des données, on part de la supposition que l'échantillon provient d'une population normale. Ensuite on vérifie si les observations s'accordent ou non avec cette hypothèse. Formellement, on écrit le couple d'hypothèses suivant :

$$H_0 : X \text{ suit une loi normale } \mathcal{N}(\mu, \sigma)$$

$$H_0 : X \text{ ne suit une loi normale } \mathcal{N}(\mu, \sigma)$$

L'hypothèse H_0 est appelée *hypothèse nulle* et H_1 est appelée *hypothèse alternative*. Les tests spécifiques pour la normalité sont des procédures conçues pour trancher uniquement entre les hypothèses précédentes, au vu des résultats d'un échantillon. On distingue deux catégories de tests selon qu'on précise ou non l'hypothèse alternative par rapport à la loi normale. Lorsque l'hypothèse alternative est précisée, le test est directionnel. Par contre lorsque l'hypothèse alternative est laissée vague le test est omnibus. Les tests spécifiques de normalité les plus courants sont :

- le test omnibus de Shapiro-Wilks
- le test directionnel d'asymétrie
- le test directionnel d'aplatissement
- le test conjoint d'asymétrie et d'aplatissement (multidirectionnel)

• Test de normalité de Shapiro-Wilks
Le test de Shapiro-Wilks est un mécanisme qui permet de trancher entre deux hypothèses :

$$H_0 : X \text{ suit une loi normale } \mathcal{N}(\mu, \sigma)$$

$$H_0 : X \text{ ne suit pas une loi normale } \mathcal{N}(\mu, \sigma)$$

au vu des résultats d'un échantillon. Le test est fondé sur la variable de décision W, qui est le rapport de deux estimateurs liés à la variance de la population. La procédure du test est la suivante :

(a) Ordonner les observations par ordre croissant :

$$X_1 \leq X_2 \leq ... \leq X_n$$

(b) Calculer la somme des carrés des écarts des observations à la moyenne de l'échantillon, noté Z^2 :

$$Z^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

(c) Calculer les étendues partielles :

$$\begin{aligned} d_1 &= X_n - X_1 \\ d_2 &= X_{n-1} - X_2 \\ &... \\ d_i &= X_{n-i+1} - X_i \end{aligned}$$

Si la taille de l'échantillon est paire : $n = 2k$, on obtient k étendues partielles. Sinon, $n = 2k + 1$, la médiane n'est pas utilisée.

(d) A l'aide des tables de Shapiro-Wilks, calculer :

$$b = \sum_{i=1}^{k} a_i d_i$$

Les coefficients a_i sont tabulés selon i et n.

(e) A partir des observations, calculer la valeur expérimentale de la variable de décision $W_{obs} = b^2/Z^2$. Il est nécessaire de calculer W_{obs} avec au moins 3 décimales, pour distinguer correctement les seuils.

(f) Selon le risque de première espèce α et n, déterminer la valeur critique W_α.

(g) Décision : si $W_{obs} > W_\alpha$, on garde l'hypothèse de normalité ; si $W_{obs} < W_\alpha$ on rejette l'hypothèse de normalité. Le non rejet de l'hypothèse de normalité conduit à l'utilisation de la loi normale pour les analyses statistiques ultérieures.

• Seuil de significativité

Dans le test de Shapiro-Wilks, la statistique W compare la forme de la distribution de l'échantillon à la distribution normale. A l'issue de cette comparaison, dans les sorties de logiciel un nombre appelé *probabilité critique* ou *seuil de significativité* (*p-value* en anglais) est calculé. Le calcul du seuil de significativité exploite la distribution de la statistique de test W est donné par :

$$p - value = \mathbb{P}(W < W_{obs})$$

Une faible valeur de la probabilité critique est une preuve contre l'hypothèse nulle H_0. La décision est prise en comparant la valeur de p-value du test avec le risque de première espèce α :

$$\text{si p-value} > \alpha, \text{ on garde l'hypothèse } H_0$$

ou de façon équivalente :

$$\text{si p-value} < \alpha, \text{ on rejette l'hypothèse } H_0$$

1.2.2 Les tests d'ajustement généraux

Ils permettent de vérifier si l'échantillon provient ou non d'une population dont la loi de probabilité est $F_0(x)$, connue. Formellement, on écrit le couple d'hypothèses suivant :

$$H_0 : X \text{ suit la loi } F_0(x)$$

$$H_0 : X \text{ ne suit pas la loi } F_0(x)$$

Les tests généraux les plus courants sont :

– le test d'ajustement du chi-deux
– le test d'ajustement de Kolmogorov
– le test d'ajustement de Cramer-von Mises

Evidemment, les tests d'ajustement généraux peuvent être utilisés pour vérifier si l'échantillon provient d'une loi normale.

2 Application numérique

Pour illustrer la méthode du graphique de probabilité, un échantillon de 24 observations sur le processus de fabrication de pivots est utilisé. Le tableau suivant donne les valeurs des diamètres de chaque pivot :

24.9568	24.9545	24.9522	24.9416	24.9518	24.9492
24.9502	24.9472	24.9446	24.9581	24.9505	24.9436
24.9425	24.9496	24.9481	24.9456	24.9534	24.9502
24.9556	24.9506	24.9562	24.9547	24.9528	24.9458

On travaillera sous Matlab.

1. Réordonnez les relevés avec la procédure `sort`, calculez les fréquences cumulées et tracez la droite de Henry.

2. Tracez l'histogramme avec la procédure `hist`.

Peut-on supposer que la distribution des diamètres des pivots est normale ?

3. Faîtes le test de Shapiro-Wilks, sachant que pour $n = 24$, les coefficients sont donnés par :

a_1	0.4493	a_5	0.1807	a_9	0.0764
a_2	0.3098	a_6	0.1512	a_{10}	0.0539
a_3	0.2554	a_7	0.1245	a_{11}	0.0321
a_4	0.2145	a_8	0.0997	a_{12}	0.0107

et $W_{5\%} = 0.916$ et $W_{1\%} = 0.884$.

IUP SID M1, 2005

T.P. SAS numéro 03

L'objet de ce TP est de se familiariser avec la procédure `capability` de `sas` qui permet de faire l'étude de la capabilité pour un jeu de données.

Un produit chimique concentré est stocké dans des flacons. Les masses de 15 flacons ont été relevées (en kg) et sont répertoriées dans le tableau suivant :

0.986	1.005	1.018	0.9965	0.9475
1.0075	1.02	0.9975	0.9705	1.01
0.996	0.9770	0.9775	1.025	1.0175

1. Tracez l'histogramme avec la procédure `chart` avec l'option `vbar` et 4 classes.

Peut-on supposer que la distribution des masses de produit chimique est normale ?

2. Faîtes le test de Shapiro-Wilks, sachant que pour $n = 15$, les coefficients sont donnés par :

a_1	0.515	a_5	0.1353
a_2	0.3306	a_6	0.088
a_3	0.2495	a_7	0.0433
a_4	0.1878		

et $W_5\%$=0.881 et $W_1\%$=0.835.

3. Les tolérances sont $T_s = 2$ et $T_i = 0.5$. Calculer les C_p et C_{pk}.
Que concluez-vous ?
Rajoutez les spécifications sur l'histogramme.

4. On change les spécifications : 1.15 et 0.9. Calculer les nouveaux C_p et C_{pk}.
Est-ce mieux ? Quel est le taux de non conformités associé au C_p ? Proposer un intervalle de confiance pour le C_p. Peut-on assurer avec un niveau de 95 % que le $C_p > 1.33$?

Le client exige un $C_p = 2$ et un $C_{pk} = 1.67$. Ces exigences sont-elles satisfaites ? Sur quel(s) paramètre(s) de la distribution faudrait-il agir pour satisfaire les exigences de ce client ?

Utiliser la procédure `capability` pour retrouver ces résultats en tapant dans la fenêtre "SAS : Program Editor" le programme suivant :

```
proc capability data=tp3;
var poids;
spec usl=1.15 lsl=0.9;
run;
proc print;
run;
```

5. Un nouveau client exige des tolérances de 0.88 et 1.11. Quel est le taux de non conformités associé au C_p si le processus reste identique. Ce client exige également un $C_p = 2$ et un $C_{pk} = 1.67$. Ces exigences sont-elles satisfaites ? Sur quel(s) paramètre(s) de la distribution faudrait-il agir pour satisfaire les exigences de ce client ?

D'après l'ensemble de ces résultats quel client est le plus facile à satisfaire ?

IUP SID M1, 2005

T.P. SAS numéro 04

L'objet de ce TP est de se familiariser avec la procédure `shewhart` de `sas` qui permet de tracer des cartes de contrôle d'un jeu de données.

Exercice 1

Une fabrique de papier utilise des cartes de contrôle pour le suivi des imperfections sur du papier fini. Le papier est stocké sous forme de rouleaux. Le nombre d'imperfections est relevé pendant 20 jours. Les résultats sont consignés dans le tableau suivant :

Jour	Nb de rouleaux produits	Nombre total d'imperfections	Jour	Nb de rouleaux produits	Nombre total d'imperfections
1	18	12	11	18	18
2	18	14	12	18	14
3	24	20	13	18	9
4	22	18	14	20	10
5	22	15	15	20	14
6	22	12	16	20	13
7	20	11	17	24	16
8	20	15	18	24	18
9	20	12	19	22	20
10	20	10	20	21	17

1. Enregistrer la table ci-dessus sous le nom tp41 en tapant les commandes :

```
data tp41;
input jour imperfections production;
cards;
1 12 18
2 14 18
3 20 24
4 18 22
...
18 18 24
19 20 22
20 17 21
;
run;
proc print;
run;
```

2. Construire une carte de contrôle pour les non-conformités en utilisant l'option `uchart` de la procédure `shewhart`. Pour cela, taper dans la fenêtre "SAS : Program Editor" le programme suivant :

```
proc shewhart data=tp41;
uchart imperfections*jour/subgroupn=production;
run;
proc print;
run;
```

Commenter.

Exercice 2

On cherche à étudier la stabilité d'un processus d'usinage de rotors fabriqués au rythme de 1000 rotors par semaine à l'aide des cartes $(\bar{X}, S)$. Pour commencer 15 échantillons de taille $n = 5$ sont prélevés. La fréquence de prélèvement est de 5 toutes les deux heures. Pour chaque échantillon, on a calculé la moyenne, son étendue et son écart-type. Le tableau suivant contient les mesures de ces 15 échantillons, ainsi que les moyennes $\bar{X}$, les étendues R et les écart-type s.

i	X_1	X_2	X_3	X_4	X_5	$\bar{X}_i$	R_i	s_i
1	135.0014	135.0000	135.0006	134.9992	135.0010	135.0000	0.0022	0.000865
2	134.9988	135.0011	134.9973	135.0009	134.9971	134.9990	0.0040	0.001907
3	135.0009	134.9989	135.0003	134.9983	135.0002	135.0000	0.0026	0.001078
4	134.9996	134.9995	134.9992	134.9990	134.9994	134.9990	0.0006	0.000241
5	134.9996	135.0007	135.0000	134.9995	134.9997	135.0000	0.0012	0.000485
6	134.9982	134.9992	135.0018	134.9987	135.0008	135.0000	0.0036	0.001509
7	134.9993	135.0003	135.0013	134.9991	135.0013	135.0000	0.0022	0.001053
8	134.9996	135.0005	135.0001	135.0001	134.9986	135.0000	0.0019	0.000733
9	134.9996	134.9999	135.0029	135.0008	134.9982	135.0000	0.0047	0.001737
10	134.9993	135.0005	135.0019	134.9994	135.0020	135.0010	0.0027	0.001303
11	135.0016	135.0003	135.0002	134.9989	135.0018	135.0010	0.0029	0.001180
12	135.0003	135.0002	134.9988	134.9993	134.9981	134.9990	0.0022	0.000934
13	135.0003	135.0006	135.0011	134.9990	135.0005	135.0000	0.0021	0.000784
14	134.9996	134.9990	135.0003	134.9998	134.9999	135.0000	0.0013	0.000476
15	134.9993	134.9990	134.9983	134.9995	135.0005	134.9990	0.0022	0.000801

On a enregistré la table ci-dessus sous le nom tp42, en appelant la variable diamètre : `diametre` et le numéro de l'échantillon `lot`.

1. Executer la commande envoyée par mail :

```
data tp42;
input lot @;
do i=1 to 5;
input diametre @;
output;
end;
drop i;
datalines;
1  135.0014  135.0000  135.0006  134.9992  135.0010
2  134.9988  135.0011  134.9973  135.0009  134.9971
3  135.0009  134.9989  135.0003  134.9983  135.0002
4  134.9996  134.9995  134.9992  134.9990  134.9994
5  134.9996  135.0007  135.0000  134.9995  134.9997
6  134.9982  134.9992  135.0018  134.9987  135.0008
7  134.9993  135.0003  135.0013  134.9991  135.0013
8  134.9996  135.0005  135.0001  135.0001  134.9986
9  134.9996  134.9999  135.0029  135.0008  134.9982
10 134.9993  135.0005  135.0019  134.9994  135.0020
11 135.0016  135.0003  135.0002  134.9989  135.0018
12 135.0003  135.0002  134.9988  134.9993  134.9981
13 135.0003  135.0006  135.0011  134.9990  135.0005
14 134.9996  134.9990  135.0003  134.9998  134.9999
```

```
15 134.9993  134.9990  134.9983  134.9995  135.0005
;
run;
proc print;
run;
```

2. Construire une carte de contrôle $(\bar{X}, s)$ en utilisant l'option `xschart` de la procédure `shewhart`. Pour cela, taper dans la fenêtre "SAS : Program Editor" le programme suivant :

```
proc shewhart data=tp42;
xschart diametre*lot;
run;
proc print;
run;
```

Commenter.

Pour garder les limites de contrôle qui sont convenables, on rajoute l'option :

```
xschart diametre*lot/outlimits=limits;
```

On peut rajouter les options suivantes :

```
xschart diametre*lot/outhistory=history outtable=table;
```

Outlimits = on a l'essentiel pour prolonger la carte de contrôle. On va l'utiliser pour prolonger la carte de contrôle à d'autres valeurs. On a une ligne avec

$$\text{lsl(x)},\ \text{ucl(x)},\ \bar{\bar{X}},\ \bar{s}$$

History = il s'agit de l'historique résumé. On a 15 lignes avec $\bar{x}_i$, s_i et n_i.

Table = il s'agit de tout ce qui permet de tracer à nouveau les cartes.

3. Effectuer une série de tests en tapant la commande :

```
proc shewhart data=tp42 limits=limits;
xschart diametre*lot/tests=1 2 3 4 5 6 7 8;
run;
proc print;
run;
```

Règles de decision

Carte de Shewhart seule :

si un point sort des limites (sur la carte des moyennes ou sur la carte des écart-type)
- procédé hors-contrôle
- réglage à faire

Carte de Shewhart avec les tests supplémentaires :

- si un point sort des limites
- si au moins un des tests est positif
  - procédé hors-contrôle
  - réglage à faire

Puis, pour montrer les différentes zones, taper

```
proc shewhart data=tp42 limits=limits;
xschart diametre*lot/tests=1 2 3 4 5 6 7 8;
run;
proc print;
run;
```

4. Après l'étude précédente, on fait à nouveau 5 prélèvements dont les résultats sont consignés dans le tableau suivant :

16	134.9987	134.9987	135.0000	135.0012	135.0011	134.9997	0.0025	0.001226
17	135.0015	134.9998	135.0005	135.0007	134.9989	134.9994	0.0026	0.000983
18	134.9995	135.0004	135.0005	135.0004	134.9996	135.0000	0.0010	0.000661
19	134.9985	135.0003	134.9989	135.0021	134.9998	135.0009	0.0036	0.000784
20	135.0006	134.9982	135.0002	135.0001	135.0023	135.0003	0.0041	0.001353

Enregistrer les nouvelles mesures avec la commande : `data tp42suite;`

```
input lot @;
do i=1 to 5;
input diametre @;
output;
end;
drop i;
datalines;
16 134.9987 134.9987 135.0000 135.0012 135.0011
17 135.0015 134.9998 135.0005 135.0007 134.9989
18 134.9995 135.0004 135.0005 135.0004 134.9996
19 134.9985 135.0003 134.9989 135.0021 134.9998
20 135.0006 134.9982 135.0002 135.0001 135.0023
;
run;
proc print;
run;
```

Tracer la carte de contrôle avec les limites précédentes en tapant la commande :

```
proc shewhart data=tp42suite limits=limits;
xschart diametre*lot;
run;
proc print;
run;
```

Commenter.