Fiche 2 - TISD - Master Pro
Variables quantitatives - Moyenne et Variance empiriques

Tran Viet Chi, chi.tran@math.univ-lille1.fr, bureau 316 (bâtiment M3).

1 Une introduction au logiciel SAS

1.1 Acquisition de données - Librairies

1.1.1 Saisie de données

1. Dans la fenêtre "SAS : Program Editor" entrer le programme suivant :

```
data tp;/*création d'une table*/
input numero taille poids sexe $ sexecode;
/*le dollar indique que la variable le précédant est qualitative*/
cards;
1 174 65 m 1
2 169 56 f 2
3 166 48 f 2
4 181 80 m 1
5 168 53 f 2
6 176 76 m 1
7 190 77 m 1
8 159 70 f 2
9 162 60 f 2
10 164 51 f 2
11 160 73 f 2
run;
```

2. Pour afficher les données entrer : `proc print; run;`

3. Pour exécuter le programme faites `Run>Submit` ou `F3`.

4. La table a été crée dans un répertoire de travail (librairie) appelé `work`. Rechercher-la avec l'explorateur de SAS et l'ouvrir.

5. Le nom des variables ne peut pas contenir d'espace et ont 8 caractères maximum. Pour avoir des noms plus précis, on utilise des labels :

```
proc print data=tp label;
label taille="taille de l'élève" poids="poids de l'élève";
run;
```

Le `label` de la première ligne indique que la procédure doit remplacer certains noms de variables par leurs labels si ceux-ci sont précisés. Le second `label` indique les labels à utiliser. Il est possible de préciser les labels lors de l'étape `data` (ces labels sont alors permanents) ou de la même façon dans d'autres procédures.

6. Dans la table `tp`, la variable `sexecode` donne le sexe de l'individu avec le code `1=masculin` et `2=féminin`. Pour faire apparaître ces mots à la place de 1 et 2, nous utilisons une `PROC FORMAT` :

```
proc format;
value sexecodage 1='masculin' 2='feminin';
run;
proc print data=tp;
```

```
format sexecode sexecodage.;
run;
```

7. Nous pouvons faire de même pour afficher 'masculin' et 'feminin' à la place de `m` et `f` pour la variable `sexe`. Comme il s'agit d'un caractère, nous utilisons cette fois la syntaxe suivante :

```
proc format;
value $ bsexecodage m='masculin' f='feminin';
run;
proc print data=tp;
format sexe $bsexecodage.;
run;
```

8. On souhaite placer nos table dans un autre répertoire *chemin* (par exemple `C:/Docs/TD`). On peut créer une nouvelle librairie avec la commande :
`libname malib '`*chemin*`';`
Si nous reprenons la question **1** en remplaçant `tp` par `malib.tp`, nous créons la table dans le répertoire `malib`.

1.1.2 Importation de données

1. Importer les données du fichier `ozone.xls` en entrant les commandes suivantes (remplacer `malib` et `chemin` par ce qu'il faut) :

```
proc import out= malib.ozone
datafile= "D:\Documents and Settings\Chi\Mes documents\TISD\Donnees\ozone2.xls"
DBMS=EXCEL REPLACE;
SHEET="tp1$";
GETNAMES=YES;
MIXED=YES;
SCANTEXT=YES;
USEDATE=YES;
SCANTIME=YES;
run;
```

2. Refaire la manipulation en utilisant `File>Import Data`.

1.2 Manipulation des données

1. Entrer `proc contents data=malib.ozone; run;`

2. Opérations élémentaires sur les variables numériques :

```
data malib.nombres;
input x y;
cards;
5 5
2 -3
4.5 10
3.2 1
2 0
run;

data malib.calcul;
set malib.nombres;
a=x+y;    b=x-y;    c=x*y;   d=x**y;   e=min(x,y);   f=max(x,y);
g=x/y;    h=abs(y);   i=exp(x);     j=int(x);    k=log(y);
l=log10(x);   m=sign(y);   n=sqrt(x);
run;
```

3. On peut faire des boucles avec l'instruction `do..to..by..end` :

```
data malib.compt;
do i=1 to 100 by 1;
x=rand('binomial', 0.4, 20);
y=1+x;
z=x;
x=x-1;
output malib.compt;
end;
run;
```

Ouvrir la table, remarquer qu'il y a un ordre dans les opérations. Remplacer la loi binômiale par d'autres lois (voir l'aide sur `rand` pour obtenir les lois disponibles).

4. Pour générer deux suites aléatoires où $X_1, \cdots X_n$ sont des variables aléatoires *iid* uniformes sur $[0, 1]$ et $Y_0 = 0$, $Y_k = Y_{k-1} + X_{k-1}$, on procède de la façon suivante :

```
data malib.marchealeat;
retain x y (1 0);
do i=1 to 10 by 1;
y=y+x;
x=rand('uniform');
output malib.marchealeat;
end;
run;
```

5. On reprend la table `malib.tp` créée à la Section 1.1.1. Pour trier cette table suivant les tailles des individus, entrer :

```
proc sort data=malib.tp;
by taille;
run;
```

6. Trier la table `malib.tp` par sexe et utiliser la `proc print` avec la commande `by sexe` pour avoir la liste des garçons et la liste des filles.

7. Pour garder des variables avec `keep`,

```
data malib.tp2;
set malib.tp;
keep numero taille poids sexecode;
run;
```

8. Pour supprimer des variables avec `drop`,

```
data malib.tp3;
set malib.tp;
drop taille poids;
run;
```

9. Pour reconstituer la table `tp` à partir de `tp2` et `tp3` :

```
data malib.tp4;
merge malib.tp2 malib.tp3;
by numero;
run;
```

La ligne `by numero;` assure que les tableaux sont bien regroupés en joignant les lignes correspondant à une même valeur de la variable `numero` qui nous sert ici d'identifiant. Il est nécessaire que les deux tableaux soient triés préalablement suivant `numero`.

10. On veut ajouter les trois informations suivantes (incomplètes) à la table `malib.tp` :

```
data malib.tab1;
input numero sexecode;
cards;
12 1
13 1
14 2
run;
```

Pour cela, faire :

```
data malib.tp5;
set malib.tp malib.tab1;
run;
```

11. Taper :

```
data malib.tp5;
set malib.tp;
where sexecode=1;
run;
```

qui revient au même que :

```
data malib.tp6;
set malib.tp;
if sexe='f' then delete;
run;
```

12. On peut séparer une table en deux :

```
data malib.garcons malib.filles;
set malib.tp;
if sexecode=1 then output malib.garcons;
if sexecode=2 then output malib.filles;
run;
```

qui revient au même que :

```
data malib.garcons2 malib.filles2;
set malib.tp;
select(sexecode);
when(1) output malib.garcons2;
otherwise output malib.filles2;
end;
run;
```

2 Etude d'une variable quantitative avec le logiciel SAS

Exercice 1 (Maximum d'ozone)
Nous nous intéressons à la série des pics d'ozone qui se trouve dans la table `ozone.xls` (variable `maxO3` correspondant au maximum d'ozone pour chaque jour de la table).

1. Importer cette base sous SAS. Combien y a-t-il d'observations ? Créer une variable `t` qui correspond au numéro de l'observation en utilisant une étape `data` avec la commande `t=_n_`.

2. Tracer l'évolution au cours du temps de la variable `maxO3` à l'aide de la `PROC GPLOT`.

```
proc gplot data=malib.ozone;
plot maxO3*t;
symbol i=join;
run;
```

3. Tracer l'histogramme de la variable `maxO3` avec la `PROC GCHART` :

```
proc gchart data=malib.ozone;
vbar maxO3;
run;
```

4. La `PROC MEANS` permet d'obtenir les statistiques les plus courantes.

```
proc means data=malib.ozone;
var maxO3;
run;
```

5. La `PROC UNIVARIATE` permet d'obtenir un nombre plus important de statistiques. Quelle est la moyenne de des observations `maxO3` ? leur variance ? leur écart-type ? le minimum ? le maximum ? l'étendue ? le coefficient de variation ? le coefficient d'asymétrie ? le coefficient d'aplatissement ? Donner la médiane, les centiles, les valeurs extrêmes.

6. En utilisant l'option `plot` dans la `PROC UNIVARIATE`, obtenir les diagrammes suivants : *steam and leaf*, *boxplot*. Obtenir un histogramme et superposer les densités de lois théoriques qui vous semblent pertinentes pour modéliser la distribution de `maxO3`. Faire un QQ-plot avec la loi qui vous semble la plus adaptée.

7. Faire de même une analyse statistique de la variable `T12` (température à midi).

8. Tracer le nuage des points d'abscisses `T12` et d'ordonnées `maxO3` avec la `PROC GPLOT`. Que cela vous inspire-t-il ? Utiliser la `PROC CORR` pour préciser cela.

9. Résumer les informations obtenues à la question **4** en quelques phrases destinées à un non-statisticien.

Exercice 2 (Des hommes et des femmes)
Les données du fichier `salaires.xls` fournissent le salaire, le sexe, la catégorie socio-professionnelle (CSP) et le nombre de jours d'absence pour chaque salarié d'une entreprise.

Partie A Nombre de jours d'absence

1. Importer ce fichier sous SAS.

2. En utilisant la `PROC GCHART` avec l'option `by` ou `class`, faites un histogramme par sexe pour la variable nombre de jours d'absence.

3. Nous nous intéressons au nombre de jours d'absence. Obtenir en utilisant la `PROC MEANS` avec l'option `by` ou `class`, obtenir les moyennes et écart-type pour les hommes, pour les femmes et pour l'ensemble. Combien y a-t-il d'hommes, de femmes ? Récupérer ces statistiques dans une table de sortie avec l'option `output`. Comparer les moyennes par sexe. Que cela vous inspire-t-il ?

3. A partir des données précédentes, calculer la variance inter et la variance intra. Conclusion ?

Partie B Salaires

1. Donner les moyennes et les écart-type des salaires par sexe. Tracer un histogramme des salaires par sexe.
2. Faire une décomposition de la variance pour étudier les disparités de salaires entre les hommes et les femmes. Conclusion ?

3. Donner la somme de tous les salaires et le nombre total de salariés.

4. On s'intéresse maintenant aux inégalités de répartition des salaires (pour l'ensemble des salariés). Tracer la courbe de Lorentz et calculer l'indice de Gini. Pour cela :
- classer les salariés par salaires croissants.
- calculer une variable correspondant aux salaires cumulés.
- associer à chaque individu la part P_2 que représente la masse salariale de l'ensemble des personnes gagnant moins que lui par rapport à la masse salariale totale.
- créer une variable P_1 associant à chaque individu la proportion de salariés gagnant moins que lui.
- tracer la courbe de Lorentz : P_2 en fonction de P_1. Superposer la première bissectrice. Commenter.
- calculer l'indice de Gini qui est égal à 2 fois l'aire entre la courbe de Lorentz et la première bissectrice. Pour cela, on approchera l'aire en décomposant la surface sous la courbe en trapèzes. Commenter le résultat.

Partie C Données groupées

1. Créer un découpage des salaires suivant les quartiles :
- demander à SAS les quartiles de la variable salaire.
- on crée une variable auxiliaire *groupe* qui vaut 1 si l'individu gagne moins que 11450, 2 s'il gagne entre 11 450 et 15 650, 3 s'il gagne entre 15 650 et 39 150 et 4 sinon. A votre avis, d'où viennent ces valeurs ?

– appeler la `PROC MEANS` en utilisant l'option `by gpesal`. Récupérer grâce à l'option `output` une table `malib.sortie` dans laquelle les observations sont les différents groupes de salaires et avec pour chacun de ces groupes l'effectif, la moyenne et la variance des salaires dans ce groupe.

2. On suppose qu'on ne dispose maintenant plus que des données résumées de `malib.sortie`. Retrouver la moyenne et la variance.

3. Supposons qu'on ne dispose plus que du nombre d'employés par classe. Donner une approximation de la moyenne et de la variance.

3 Partie avec R : Moyenne et variance empiriques

Exercice 3 (Simulations)
Nous nous proposons d'illustrer par des simulations quelques fait simples sur la moyenne et la variance empiriques.

1. Simuler un échantillon de $n = 1000$ variables aléatoires *iid* $X_1, \cdots X_n$ suivant une loi normale $\mathcal{N}(m = 2, \sigma^2 = 121)$.
1.1. Tracer l'histogramme des observations et superposer la densité approchée. Dessiner un *boxplot* des observations.
1.2. Calculer la moyenne empirique et la variance empirique corrigée. Quelles sont l'espérance et la variance de la loi exponentielle étudiée ? Rappeler la loi des grands nombres et expliquer les proximités observées.

2. Réaliser le programme suivant :
- simuler $N = 300$ échantillons de $n = 1000$ variables aléatoires *iid* $X_1, \cdots X_n$ suivant la loi $\mathcal{N}(m = 2, \sigma^2 = 121)$.
- pour l'échantillon $i \in [\![1, N]\!]$, calculer la moyenne empirique $\bar{X}_n^{(i)}$, l'écart-type corrigé $s_n^{(i)}$, la statistique $\zeta_n^{(i)} = \sqrt{n}(\bar{X}_n^{(i)} - 2)/\sigma$ et la statistique $\xi_n^{(i)} = \sqrt{n}(\bar{X}_n^{(i)} - 2)/s_n^{(i)}$. Nous avons donc N réalisations *iid* de ces variables aléatoires correspondant à des tirages d'échantillons différents. Ces valeurs seront conservées dans un tableau de 4 colonnes et $N = 300$ lignes.

2.2. Tracer l'histogramme des $\bar{X}_n^{(1)}, \cdots \bar{X}_n^{(N)}$ et lui superposer sa densité approchée. Faire de même pour les $s_n^{(1)}, \cdots s_n^{(N)}$, pour les $\zeta_n^{(1)}, \cdots \zeta_n^{(N)}$ et pour les $\xi_n^{(1)}, \cdots \xi_n^{(N)}$.
2.3. Vérifier par les *QQ-plots* adéquats que :
- la loi de la moyenne empirique est $\mathcal{N}(m, \sigma^2/n)$,
- la loi de la variance empirique corrigée est $\Gamma((n-1)/2, 1/2)$, encore appelée $\chi^2(n-1)$ (citer une façon d'obtenir ce résultat),
- la loi des ζ_n est une loi normale $\mathcal{N}(0, 1)$ (d'où vient ceci ?),
- la loi des ξ_n est une loi de Student à $n - 1$ degrés de liberté (faire aussi le *QQ-plot* avec la loi normale $\mathcal{N}(0, 1)$.