IMIS : Master 1
Université Paris Est – Marne la Vallée

TP 5 : Procédures de Test

1. Test d'indépendance et de normalité sur des données simulées

On veut obtenir une table contenant 50 réalisations de deux variables indépendantes X_1 et X_2 de même loi de distribution normale centrée réduite. On utilise la méthode de Box-Muller à cette fin.

Méthode de Box-Muller (rappel) : On considère U_1 et U_2 deux variables aléatoires suivants la loi uniforme sur $[0,1]$ et on définit V_1 et V_2 par $V_i = 2U_i - 1$ ainsi que

$$S = V_1^2 + V_2^2 \,.$$

Pour $S \leq 1$, on définit enfin X_1 et X_2 par

$$X_i = V_i \sqrt{-2\frac{\log S}{S}} \,.$$

Les variables aléatoires X_1 et X_2 ainsi définies sont indépendantes et sont de même loi de distribution normale centrée réduite.

Créer une table contenant toutes ces variables en une seule étape data. Utiliser l'nstruction ranuni (k), qui permet de simuler une variable aléatoire de loi uniforme sur $[0,1]$ (prendre $k = 0$ pour la variable U_1 et $k = 1$ pour la variable U_2). L'instruction retain permet de mémoriser des informations au fil des observations (par exemple un compteur). Cette instruction empêche SAS de réinitialiser une variable avant de passer au traitement de l'individu suivant dans la boucle implicite. L'exemple suivant crée une variable compteur, nommée numero, attribuant un numéro d'ordre correspondant au numéro de ligne dans la table.

```
DATA TableCree;
        SET TableSource;
        RETAIN Cpt 0; /*la variable 'Cpt' est initialisée a 0*/
        Cpt +1; /*Incrémente la variable 'Cpt' de 1 à chaque passage*/
RUN;
```

Dressez un histogramme pour chacune des deux variables X_i (instruction 'HISTOGRAM' de la procédure 'UNIVARIATE' ; ou bien l'instruction 'VBAR' de la procédure 'GCHART'). A l'aide de la procédure 'UNIVARIATE', vérifier la normalité des données. Les deux variables sont-elles corrélées (utiliser la procédure 'CORR') ?

2. Tests sur des données réelles : revenus de deux villes

Télécharger les fichiers 'VilleA.txt' et 'VilleB.txt' se trouvant à l'adresse
http ://perso-math.univ-mlv.fr/users/hebiri.mohamed/ (rubrique 'Enseignements')

Enregistrer les dans votre répertoire de travail. Ces fichiers contiennent des données sur les revenus des résidents de deux villes. Nous allons comparer la richesse de ces deux populations.

1. On va commencer par mener une étude sur les revenus pour chacune des deux villes.
   Importer les données dans deux tables SAS séparées en ajoutant pour chacune d'elles (à la variable existante que vous appellerez 'Revenus') une variable nommée 'Groupe' (=A pour la ville A et =B pour la ville B) permettant de distinguer ces deux villes. Construire par la suite et à partir des deux tables précédemment définies, une unique table SAS intitulée 'RevenusAB'.
   Testez la normalité de chacun des deux groupes de données. Tester l'hypothèse $\mu_0 = 1500$ pour chacun des deux groupes (utiliser l'option `< mu0 = 1500 >`). Ajouter l'option 'CIBASIC' à la procédure 'UNIVARIATE' pour obtenir des intervalles de confiance pour la moyenne et la variance de chacun des deux groupes.

2. On va à présent comparer les revenus des deux villes.
   a) Tracez les boxplots des revenus pour chacune des deux villes. Interpréter les résultats.
   b) Avec la procédure 'TTEST' (les variables sont gaussiennes), étudier une éventuelle différence entre les deux groupes. Montrer que l'on peut considérer que les variances des deux groupes sont égales mais pas les moyennes. (*Lorsque les variables ne sont pas gaussiennes on peut utiliser des tests non paramétriques avec la procédure* NPAR1WAY).

3. D'après ce qui a été fait précédemment, on a constaté que les moyennes des revenus des deux villes est différentes alors que leur variance semble être les mêmes. On veut donc construire un estimateur de cette variance commune.
   Pour cette situation où les espérances des deux groupes ne sont pas les mêmes, la procédure TTEST ne fournit pas d'estimation de la variance supposée commune. On se propose donc de construire un intervalle de confiance pour la variance commune.
   On considère l'estimateur sans biais de la variance :
   $$\mathcal{S}^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}.$$
   On note
   $$SC_{tot} := SC_1 + SC_2 := \sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2.$$
   L'intervalle de confiance bilatéral pour σ de niveau de confiance $1 - \alpha$ est
   $$\left[\frac{\sqrt{SC_{tot}}}{\sqrt{G^{-1}(\alpha/2)}} ; \frac{\sqrt{SC_{tot}}}{\sqrt{G^{-1}(1-\alpha/2)}} \right]$$
   où $G(t) = P(Z > t)$ avec $Z \sim \chi^2(n_1 + n_2 - 2)$.

a) SAS ne fournit pas cet intervalle. Lorsque l'on doit mener des calculs sur des sorties SAS, on commence par créer une table contenant ces sorties. L'instruction pour créer une table en sortie dépend de la procédure, mais on utilisera surtout les instructions 'ODS OUTPUT' et 'OUTPUT', de syntaxe respectives :

```
ODS OUTPUT NomSAS (Nom donné par SAS) = *** (Nom donné par l'utilisateur)
OUTPUT OUT = *** (Nom donné par l'utilisateur)
```

Utilisez les instructions 'ODS' pour créer la table nommée 'CALCUL' en sortie de la procédure 'TTEST' :

```
PROC TTEST data=RevenusAB ;
        CLASS Groupe ;
        VAR Revenus ;
        ODS OUTPUT statistics=CALCUL ;
RUN ;
```

Afficher la table 'CALCUL' dans la fenêtre des sorties.

b) A l'aide de la table 'CALCUL' et des options 'KEEP', 'FIRSTOBS', 'OBS' (première et dernière observations lues), *etc.*, de l'instruction 'SET', créer une table, nommée 'CALCUL2 ne contenant que les deux premières lignes des variables 'N' (numéro de l'observation) et 'StdDev' (écart type ; attention, cette variable est toujours divisée par $N-1$). Ajouter pendant cette étape 'DATA une variable nommée SC (de composantes SC_1 et SC_2, définies comme ci-dessus).
Construiser et afficher par la suite, à l'aide d'une nouvelle table l'intervalle de confiance de niveau 95% pour la variance. Utiliser l'instruction 'CINV(α,q)' qui permet d'inverser en α la fonction de répartition d'un Khi2 à q degrés de liberté. La valeur obtenue par cette instruction est donc le quantile de niveau α de la loi du Khi2 à q degrés de liberté.