IMIS : Master 1
Université Paris Est – Marne la Vallée

TP 6 : Quelques procédures statistiques

Analyse de données et d'influence

Télécharger le fichiers 'univ.sas7bdat' se trouvant à l'adresse
http ://perso-math.univ-mlv.fr/users/hebiri.mohamed/ (rubrique 'Enseignements')

Enregistrer le dans votre répertoire de travail.
La table 'univ' détaille pour 60 étudiants les résultats à un examen de littérature, les résultats à un examen de mathématiques, le nombre d'heures de travail personnel et le nombre d'heures passées devant la télévision la semaine précédant les deux examens. Les étudiants proviennent de 3 classes différentes (les 20 premiers dans la table sont dans la classe A, les 20 suivants dans la classe B et les 20 derniers dans la classe C).

1. Faire une procédure '`CONTENTS`' pour savoir comment sont nommées les variables. Créez une une variable '`Score`' pour la somme des notes aux deux examens, ainsi que la variable '`groupe`' (elle pourra prendre les modalités suivantes : **groupe 1** pour les 20 premiers étudiants, **groupe 2** pour les 20 suivants, et **groupe 3** pour les 20 derniers). Pour ce faire, voici le début de la procédure :

```
DATA TableCree ;
        FORMAT groupe : $1. ;
        SET TableSource ;
        Score = ? + ? ;
        i + 1 ; /* Creation d un numero d observation */
        IF i <= 20 THEN DO ;...
```

   Rajoutez une instruction pour supprimer le numéro i dans la table.
   Analysez la variable score à l'aide de la procédure '`UNIVARIATE`', déterminer les statistiques élémentaires de l'échantillon complet et discutez la normalité des données. La distribution de la variable '`Score`' est-elle symétrique ? Evaluez graphiquement la normalité des données et ajustez une densité normale sur celui-ci.

2. Dressez le graphique du nombre d'heures passées devant la télévision en fonction du nombres d'heures passées à faire des exercices. Représentez sur un même graphe le Score en fonction du nombre d'heures passées à faire des exercices, et le score en fonction du nombre d'heures de TV.

3. Analysez la corrélation entre les variables avec la procédure '`CORR`'.

4. Analysez la variable score pour chacun des 3 groupes.

5. On veut savoir si la classe a une influence sur le score des étudiants. Utilisez la procédure `BOXPLOT` permet de comparer graphiquement les distributions de plusieurs groupes :

```
PROC BOXPLOT Data = *** ;
        PLOT Var*VarClasse ;
RUN ;
QUIT ;
```

6. Comparez les distributions en utilisant des procédures ' `TTEST`' et une procédure '`ANOVA`'.

```
PROC ANOVA Data = *** ;
        CLASS groupe ;
        MODEL score = groupe ;
RUN ;
QUIT ;
```

Vous pouvez également ajouter à la procédure '`ANOVA`', l'instruction ' `MEANS`' avec les options ' `/ scheffe hovtest`' pour effectuer des comparaisons entre les moyennes, et un test d'homogénéité de la variance de la variable '`Score`' selon les différentes modalités de la variable '`Groupe`'.

Voici quelques quantités qui apparaissent dans la procédure '`ANOVA`' et que vous êtes souvent amenés à rencontrer :

Model DF	p (nombre de variables explicatives),
Error DF	$n-p$,
Model Sum of Squares	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = SSR$,
Error Sum of Squares	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = SSE$,
Corected Total Sum of Squares	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = TSS$,
Mean Square	Sum of Squares/DF,
F-value	c'est la valeur de la statistique F pour tester l'hypothèse de nulleté de tous les coefficients sauf l'intercept. C'est le rapport entre Model Mean Square et Error Mean Square
Pr > F	c'est la probabilité d'avoir observé une valeur $> F$ sous l'hypothèse que tous les coefficients sont nuls. Cette quantité permet de vérifier si la variable explicative influence significativement la variable à expliquer ou pas. Si la valeur de `Pr > F` est plus petite que 0.05, on peut affirmer que la prédiction de la variable à expliquer basée sur la variable explicative est fiable. En revanche, si cette valeur est > 0.05, on dira que la variable explicative ne fournit pas une prévision fiable de la variable à expliquer,
Root MSE	c'est l'estimateur $\hat{\sigma} = \|y - X\hat{\theta}\|^2/(n-p)$ de l'écart-type σ des erreurs ξ_i, c'est également la racine carrée de Mean Square Error,
Dependent Mean	c'est la moyenne empirique $\overline{Y}$ de la variable à expliquer Y,
Coeff Var	c'est le coefficient de variation. C'est une quantité sans unité de mesure qui caractérise la dispersion des données. Par définition, il est égal à Root MSE/$\overline{Y}$,
R-Square	c'est le coefficient de détermination : la part da la variance totale de Y expliquée par la variable (ou les variables) explicative(s).