Laboratoire de Statistique et Probabilités — U.M.R. 5583 (C.N.R.S. - U.P.S. - INSAT)

Compléments sur le T.P. SAS numéro 02

Quelques précisions sur la procédure univariate

En utilisation standard, la procédure `univariate` de SAS produit une page de résultats pour chaque variable quantitative considérée. En rajoutant l'option `plot`, on obtient en plus divers graphiques pour chaque variable.

L'objet de ce document est de donner quelques précisions sur la signification statistique des sorties de la procédure `univariate`.

Le tableau "Moments"

Il comporte systématiquement 6 lignes, chacune donnant la valeur de deux indicateurs relatifs à la variable analysée. Nous précisons ci-dessous la signification de ces indicateurs.

1. **N** : représente le nombre d'observations sur lesquelles ont été calculés les différents indicateurs statistiques ; c'est en général la taille du fichier (autrement dit de l'échantillon considéré), mais çà peut être le nombre de données présentes, s'il y a des données manquantes.
   Sum Weights (= somme des poids) : total des poids des observations présentes, si une variable poids a été déclarée ; sinon, est égal à **N**.
2. **Mean** : moyenne arithmétique des observations présentes.
   Sum Observations : somme de ces observations.
3. **Std Deviation** (= Standard Deviation = écart-type) : écart-type des observations.
   Attention : dans toutes les procédures SAS, variances et écarts-types sont calculés, par défaut, avec $N-1$ en dénominateur (optique statistique inférentielle) ; pour avoir les valeurs calculées avec N au dénominateur (optique statistique descriptive), il faut rajouter, dans chaque procédure concernée, l'option `vardef = n`.
   Variance : variance des observations.
4. **Skewness** (= coefficient d'asymétrie) : c'est le coefficient noté γ_1 (voir ci-dessous).
   Kurtosis (= coefficient d'aplatissement) : c'est le coefficient noté γ_2.
   Rappel : pour une suite de N observations notées $x_1, \ldots, x_N$, notons μ_r le moment empirique centré d'ordre r : $\mu_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^r$ ($\overline{x}$ désigne la moyenne arithmétique empirique) ; on appelle alors coefficient d'asymétrie (skewness) la quantité $\gamma_1 = \dfrac{\mu_3}{\mu_2^{3/2}}$ et coefficient d'aplatissement (kurtosis) la quantité $\gamma_2 = \dfrac{\mu_4}{\mu_2^2} - 3$.
5. **Uncorrected SS** (= Uncorrected Sum of Squares) : c'est la somme des carrés des observations (c'est-à-dire $\sum_{i=1}^{N}(x_i)^2$).
   Corrected SS (= Corrected Sum of Squares) : somme des carrés des écarts à la moyenne (c'est-à-dire $\sum_{i=1}^{N}(x_i - \overline{x})^2$).
6. **Coeff Variation** : c'est le coefficient de variation, rapport de l'écart-type à la moyenne. Il n'est défini que si la moyenne est non nulle, est du même signe que la moyenne, n'est nul que pour les séries constantes et s'exprime sans unité. Ce coefficient n'a réellement de sens que pour les séries à valeurs positives. Pour des raisons peu claires, SAS multiplie systématiquement (dans toutes les procédures où il intervient) ce coefficient par 100.
   Std Error Mean (= Standard Error of the Mean) : représente l'erreur-type de la moyenne, c'est-à-dire l'écart-type de la statistique $\overline{X}_N$, estimateur de la moyenne de la loi dont sont

issues les observations (cas d'un échantillon i.i.d.). Cette erreur-type vaut $\hat{\sigma}/\sqrt{n}$ où $\hat{\sigma}$ est la valeur donnée dans **Std Deviation** (estimateur de l'écart-type de la loi).

Le tableau "Basic Statistical Measures"

Il se décompose en deux colonnes. Celle de gauche fournit la valeur de 3 indicateurs de tendance centrale (Location) : la moyenne (arithmétique), la médiane et le mode. La colonne de droite fournit la valeur de 4 indicateurs de dispersion (Variability) : l'écart-type, la variance, l'étendue (range, écart entre la plus grande et la plus petite des observations) et l'intervalle interquartiles.

Le tableau "Tests for Location : Mu0=0"

Les résultats fournis par ce tableau sont relatifs à des tests statistiques et sont détaillées dans la feuille de T.P. numéro 11.

Le tableau "Quantiles"

On notera tout d'abord que `Definition` est un paramètre entier pouvant prendre les valeurs de 1 à 5, sa valeur par défaut étant 5. Ce paramètre régit la façon dont est déterminé précisément un quantile lorsqu'il est situé entre 2 observations consécutives de la série étudiée.

Le tableau "Quantiles" fournit les quantiles respectifs d'ordre 1 (maximum de la série), 0,99, 0,95, 0,90 (neuvième décile), 0,75 (troisième quartile), 0,50 (médiane), 0,25 (premier quartile) 0,10 (premier décile), 0,05, 0,01 et 0 (minimum de la série).

Le tableau "Extremes Observations"

Fournit d'abord, dans sa partie gauche, les 5 plus petites valeurs observées de la série (**Lowest**) avec, en face, le numéro des observations correspondantes (**Obs**). Fournit ensuite, dans sa partie droite, les 5 plus grandes valeurs observées (**Higest**) avec, de la même manière, le numéro des observations correspondantes (**Obs**).

Les diagrammes "Steam and Leaf" et "Boxplot"

Ces diagrammes ne sont fournis que si l'on rajoute l'option `plot` dans la procédure `univariate`.

Le principe du diagramme "Steam and Leaf" (tige et feuille) a été exposé dans le cours de Statistique Descriptive (volume 1, chapitre 2). Selon les cas, dans la partie "dizaine", chaque valeur peut être dédoublée en une première classe pour les "unités" 0 1 2 3 4 et une seconde classe pour les "unités" 5 6 7 8 9. En face de chaque classe est fourni son effectif.

Dans la même sortie est dessiné le "Boxplot" (diagramme en boîte) correspondant (voir le même chapitre du cours polycopié).

Le diagramme "Normal Probability Plot"

Comme les précédents, ce diagramme n'est fourni que si l'on rajoute l'option `plot`. Il représente ce que l'on appelle habituellement la droite de Henry.

Le principe de la droite de Henry est le suivant : après avoir ordonné, par ordre croissant, les observations de la série considérée, on représente chacune d'elle par le point ayant pour abscisse le quantile théorique, d'ordre i/n, d'une loi normale réduite et pour ordonnée le quantile empirique associé à cette observation, autrement dit i/n. Si les données considérées proviennent d'une loi normale, et seulement dans ce cas là, on peut montrer que les points représentés sont sensiblement alignés.

La droite de Henry est donc une façon simple, bien que sommaire, d'étudier l'ajustement de la loi normale à un ensemble d'observations.

On se reportera à la feuille de T.P. numéro 10 pour plus de détails.