T.P. SAS numéro 11

Cette feuille de T.P., un peu particulière, est consacrée aux tests statistiques les plus courants, qu'ils soient paramétriques ou non paramétriques. Les problèmes envisagés sont ceux à un échantillon, à deux échantillons et d'ajustement à une loi donnée. Dans chaque cas considéré, on rappelle d'abord le principe général du test et ses conditions de validité ; on précise ensuite comment le mettre en œuvre avec SAS ; on renvoie enfin à des fichiers (en libre accès) permettant d'illustrer les notions traitées.

1 Problèmes à un échantillon

On considère un échantillon $(X_1, \ldots, X_i, \ldots, X_n)$ de n variables aléatoires réelles (v.a.r.) indépendantes et identiquement distribuées (i.i.d.) selon une loi de probabilité qui, dans toute la suite, sera supposée **continue**.

On s'intéresse à un paramètre de tendance centrale de cette loi, noté θ (en fait, il s'agira soit de la moyenne arithmétique soit de la médiane) et on souhaite, avec un niveau α fixé ($\alpha \in]0,1[$), tester l'hypothèse nulle $\{H_0 : \theta = \theta_0\}$. La seule alternative considérée dans SAS est $\{H_1 : \theta \neq \theta_0\}$, mais on verra, dans la remarque 2, que l'on peut aussi envisager des alternatives unilatérales.

Selon l'importance des hypothèses faites sur la loi de probabilité considérée, on rencontre divers tests statistiques. Nous donnons ci-dessous quelques indications sur le principe et la mise en œuvre des trois principaux tests fournis par SAS. Deux exemples, permettant d'illustrer cette mise en œuvre, sont proposés dans le dernier point de ce paragraphe.

1.1 Le test de Student

La loi de probabilité est ici supposée normale $\mathcal{N}(\mu, \sigma^2)$, le paramètre σ^2 étant inconnu. Le paramètre sur lequel porte le test est $\theta = \mu$, moyenne de la loi normale (c'en est aussi la médiane, puisque la loi est symétrique). L'hypothèse nulle est $\{H_0 : \mu = \mu_0\}$ et l'alternative $\{H_1 : \mu \neq \mu_0\}$.

De façon classique, la statistique de test est

$$T_n = \sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n} \ , \ \text{avec} \ \ \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \ \ \text{et} \ \ S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Sous H_0, T_n est distribuée selon une loi de Student à $n-1$ degrés de liberté (d.d.l.).

On notera que chaque v.a.r. $Y_i = X_i - \mu_0$ est $\mathcal{N}(\nu = \mu - \mu_0, \sigma^2)$ et donc, sous H_0, $\mathcal{N}(0, \sigma^2)$. La statistique de test peut ainsi se réécrire sous la forme

$$T_n = \sqrt{n}\frac{\overline{Y}_n}{S_n},$$

le test consistant alors à tester $\{H_0 : \nu = 0\}$ contre $\{H_1 : \nu \neq 0\}$.

Cette dernière optique est celle de SAS dans la procédure `univariate`, de sorte que, si l'on dispose d'un fichier de données contenant les observations x_i des v.a.r. X_i ($i = 1, \ldots, n$), on doit d'abord, pour effectuer le test considéré, calculer dans SAS les quantités $y_i = x_i - \mu_0$, puis utiliser la procédure `univariate` sur la variable Y pour obtenir la valeur de la statistique de test T_n.

En utilisation standard, cette procédure fournit en sortie, dans la première ligne du tableau `Tests for Location: Mu0=0`, la valeur t_n de T_n ainsi que la *p-value* associée. On notera que cette *p-value* représente la probabilité

$$P[\ |T_n| > |t_n|\ ] = P[\ T_n < -\ |t_n|\ ] + P[\ T_n > +\ |t_n|\ ]$$

correspondant bien à un test bilatéral : on rejette H_0 si et seulement si (ssi) cette *p-value* est strictement inférieure à α.

On peut encore obtenir ces 2 quantités (t_n et la *p-value* associée) dans une table SAS en rajoutant, dans la procédure `univariate`, la commande `output = <nom_de_la_table>` suivie des options `t = ...` et `probt = ...`

Remarque 1 Test sur la variance. *SAS ne fournit pas directement la statistique du test de l'hypothèse nulle* $\{H_0 : \sigma^2 = \sigma_0^2\}$. *Cette statistique s'écrit*

$$K_n = \frac{(n-1)S_n^2}{\sigma_0} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{\sigma_0}$$

et, sous H_0, *elle est distribuée selon une loi de khi-deux à* $n-1$ *d.d.l. En fait, elle se calcule très facilement avec SAS, tout comme la* p-value *associée.*

Remarque 2 Cas d'une alternative unilatérale. *Si l'on veut faire un test unilatéral (c'est-à-dire considérer comme hypothèse alternative soit* $\{H_1' : \mu < \mu_0\}$, *soit* $\{H_1'' : \mu > \mu_0\}$*), il suffit de comparer la* p-value *obtenue à* 2α *(rejet ssi la* p-value *est strictement inférieure à* 2α*). Cette remarque, ainsi d'ailleurs que la suivante, reste valable dans les deux autres tests présentés aux points suivants.*

Remarque 3 Cas de 2 échantillons appariés. *Il est fréquent, en statistique, que l'on considère 2 échantillons appariés* $(X_1, \ldots, X_n)$ *et* $(Y_1, \ldots, Y_n)$*; les* X_i *sont i.i.d., ainsi que les* Y_i*,* X_i *et* Y_i *représentant 2 mesures distinctes sur le même individu i (ou la même mesure à 2 instants différents, par exemple avant et après un certain traitement). On est alors amené à tester l'égalité d'un indicateur de tendance centrale sur les 2 séries de mesures. On se ramène pour cela au cas d'un unique échantillon en déterminant la série des différences* $Z_i = X_i - Y_i$ *et en testant la nullité de l'indicateur de tendance centrale considéré sur la série des* Z_i*. Pour le test de Student, on testera donc la nullité de la moyenne de la série des* Z_i*, à condition de pouvoir supposer que cette série est normale. La deuxième illustration proposée en 1.4 correspond au cas de deux échantillons appariés.*

1.2 Le test de Wilcoxon

En plus de la continuité de la loi de probabilité des v.a.r. X_i, hypothèse nécessaire pour tous les tests considérés ici, on ne suppose plus maintenant que la symétrie de cette loi (l'hypothèse de normalité est abandonnée).

Le test décrit ci-dessous s'appelle encore "test des rangs signés de Wilcoxon", ou *Wilcoxon signed-rank test*. Il s'agit d'un test non paramétrique (ou, plus exactement, sans hypothèse sur les distributions — *distribution-free*).

On pose $\theta = M$, médiane de la distribution des X_i (il s'agit encore de la moyenne, puisque la distribution est symétrique). L'hypothèse nulle est $\{H_0 : M = M_0\}$ et l'alternative $\{H_0 : M \neq M_0\}$. La statistique du test est W_n^+, somme des rangs affectés aux valeurs positives de $X_i - M_0$ dans le rangement par ordre croissant de la série des $|X_i - M_0|$, série dans laquelle on a enlevé les éventuels 0 (n désigne ainsi le nombre d'écarts $|X_i - M_0|$ non nuls).

Pour réaliser ce test avec SAS, on doit d'abord, là encore, faire calculer les quantités $y_i = x_i - M_0$, puis, après avoir supprimé les éventuels 0, utiliser la procédure `univariate` sur la variable Y. SAS calcule en fait la statistique

$$V_n = |W_n^+ - \frac{n(n+1)}{4}| \ ,$$

la quantité $\dfrac{n(n+1)}{4}$ représentant le rang moyen des $|X_i - M_0|$. Il est alors équivalent d'utiliser W_n^+ ou V_n pour faire le test, à condition de décaler d'autant les valeurs critiques.

Toujours dans la sortie standard de la procédure `univariate`, on trouve la valeur de la statistique V_n et la *p-value* associée dans la troisième ligne du tableau `Tests for Location: Mu0=0`. Pour obtenir ces 2 quantités dans une table SAS, il faut maintenant utiliser la commande `output` avec les options `signrank = ...` et `probs = ...`

Remarque 4 Traitement des ex-æquo (*ties*). *Dans toutes les statistiques de rangs, on attribue aux valeurs ex-æquo la moyenne des rangs qu'elles occuperaient si elles ne l'étaient pas. SAS procède de cette façon là.*

Remarque 5 À propos du calcul des *p-values*. *La* p-value *fournie par SAS dans ce cas est la valeur exacte si* $n \leq 20$. *Dans les autres cas* ($n > 20$), *la* p-value *est obtenue par une approximation de la statistique de test au moyen de la loi de Student (moins classique que l'approximation usuelle de* W_n^+ *par la loi normale).*

1.3 Le test des signes

Ce test est le seul qui soit réalisable si l'on ne fait aucune autre hypothèse que la continuité sur la loi des X_i. On le trouve sous le nom de *ordinary sign test* dans la littérature statistique anglo-américaine.

La moyenne et la médiane ne sont plus ici identiques et le test porte sur la médiane. Hypothèses nulle et alternative sont les mêmes que dans le test de Wilcoxon.

La statistique de test est S_n, nombre d'observations X_i strictement supérieures à M_0 (là encore, les valeurs éventuellement égales à M_0 sont supprimées, n désignant le nombre d'observations différentes de M_0 dans l'échantillon). Après avoir encore une fois calculé les quantités $y_i = x_i - M_0$, la procédure `univariate` de SAS appliquée à Y détermine la statistique

$$M_n = S_n - \frac{n}{2}.$$

Les tests basés sur S_n et sur M_n sont encore équivalents, à condition de décaler les valeurs critiques. La valeur de M_n et la *p-value* associée sont données dans la deuxième ligne du tableau `Tests for Location: Mu0=0`. Avec la commande `output`, les options correspondantes sont `msign = ...` et `probm = ...` Les *p-values* fournies avec ce test sont les valeurs exactes.

1.4 Illustrations

Les fichiers suivants

```
~baccini/tpsas/exolim/data/mesure.txt
~baccini/tpsas/exolim/data/mesure.don
~baccini/tpsas/exolim/exo11_1a.sas
```

proposent respectivement, pour un petit exemple fictif univarié, la description, les données et le programme SAS permettant d'illustrer les trois tests décrits ci-dessus. On peut les copier et mettre en œuvre le programme SAS.

Même chose avec les fichiers

```
~baccini/tpsas/exolim/data/course.txt
~baccini/tpsas/exolim/data/course.don
~baccini/tpsas/exolim/exo11_1b.sas
```

pour un exemple fictif de 2 séries appariées.

Dans ces deux exemples, on notera que les *p-values* des tests de Student et de Wilcoxon sont très proches, mais que celles du test des signes sont nettement plus grandes. Du fait de la faiblesse d'une part des hypothèses sur la loi des observations, d'autre part de l'information prise en compte par la statistique de test, le test des signes est peu puissant : il a tendance à ne rejeter que très rarement l'hypothèse nulle.

2 Problèmes à deux échantillons

Dans cette partie, on considère 2 échantillons indépendants entre eux, chacun étant constitué de v.a.r. i.i.d. Nous noterons n et m les tailles de ces échantillons, notés respectivement $(X_1, \ldots, X_i, \ldots, X_n)$ et $(Y_1, \ldots, Y_j, \ldots, Y_m)$.

La question que l'on se pose dans ce cas, et à laquelle on va essayer de répondre au moyen de tests statistiques, est la suivante : les 2 échantillons considérés sont-ils, ou non, issus de la même population ?

Là encore, selon l'importance des hypothèses faites sur les lois de probabilités dont sont issus les 2 échantillons, nous trouverons divers tests statistiques. Si l'hypothèse de normalité peut être

faite sur ces 2 lois de probabilités, on dispose des tests classiques de Fisher et de Student ; leur mise en œuvre dans SAS se fait au moyen de la procédure `ttest`. Dans le cas contraire, on a recours au test non paramétrique de Mann-Whitney dont la mise en œuvre se fait au moyen de la procédure `npar1way` de SAS.

Nous donnons ci-dessous quelques précisions sur ces tests et sur leur mise en œuvre dans SAS, ainsi que 2 exemples d'application.

2.1 Le test de Fisher

Notons $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$ les lois de probabilité normales ayant respectivement généré les 2 échantillons considérés ; μ_1, σ_1^2, μ_2 et σ_2^2 sont des paramètres inconnus. Le test de Fisher consiste à tester l'hypothèse nulle $\{H_0 : \sigma_1^2 = \sigma_2^2\}$ contre l'alternative $\{H_1 : \sigma_1^2 \neq \sigma_2^2\}$, avec un niveau α fixé.

La statistique de test est $F = \sup(F_1, F_2)$, où $F_1 = \dfrac{S_1^2}{S_2^2}$ et $F_2 = \dfrac{1}{F_1}$, avec :

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \ , \ \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \ S_2^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2 \ , \ \overline{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j.$$

On compare F à la valeur critique d'une loi de Fisher à $\nu_n - 1$ et $\nu_d - 1$ d.d.l., où ν_n est la taille de l'échantillon intervenant au numérateur de F (n ou m) et ν_d celle de l'échantillon intervenant au dénominateur de F (m ou n).

Pour la mise en œuvre dans SAS, il faut tout d'abord noter que les observations issues des 2 échantillons doivent être rassemblées dans le même fichier et dans la même colonne (un seul nom de variable doit donc être utilisé pour l'ensemble des observations). Ensuite, il est nécessaire de faire figurer, dans le fichier des données, une autre variable indiquant le numéro de l'échantillon d'origine. Dans la procédure `ttest`, cette dernière variable est alors déclarée avec la commande `class`, la variable contenant l'ensemble des observations étant déclarée avec la commande `var`. La sortie standard de cette procédure fournit, dans la ligne `Equality of Variances`, la valeur de F, les degrés de liberté correspondants et la *p-value* associée, ce qui permet de conclure quant à l'égalité des variances.

Remarque 6 Archivage des sorties. *On notera qu'il n'est pas possible de récupérer les sorties de la procédure* `ttest` *dans une table SAS.*

Remarque 7 Test unilatéraux. *On peut faire un test de Fisher unilatéral, à condition de procéder comme indiqué dans la remarque 2. Il en va de même avec les tests de Student décrits ci-dessous.*

2.2 Le test de Student

Ce test n'est envisageable que dans la mesure où le test de Fisher n'a pas rejeté l'hypothèse d'égalité des variances (les lois étant toujours supposées normales). L'hypothèse nulle est alors $\{H_0 : \mu_1 = \mu_2\}$, l'alternative étant $\{H_1 : \mu_1 \neq \mu_2\}$, le niveau du test valant toujours α.

La statistique de test est

$$T_\nu = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{n+m}{nm}}} \text{ , avec } \ S^2 = \frac{1}{n+m-2}[(n-1)S_1^2 + (m-1)S_2^2].$$

On compare $|T_\nu|$ à la valeur critique d'une loi de Student à $\nu = (n + m - 2)$ d.d.l.

La procédure `ttest` de SAS fournit dans sa sortie standard, à la ligne `Equal` du tableau `T-Tests`, la valeur de T_ν, celle de ν et la *p-value* associée à $|T_\nu|$.

2.3 Approximations du test de Student

En faisant toujours l'hypothèse de normalité des distributions, si le test de Fisher a conduit à rejeter l'égalité des variances, le Test de Student, décrit ci-dessus, n'est plus valable. De plus, on sait qu'il n'existe pas, dans ce cas, de test exact pour tester l'égalité des moyennes (ce problème, célèbre en statistique, porte le nom de problème de Behrens-Fisher). On doit alors se contenter d'un test approché (asymptotique). Nous décrivons ci-dessous 2 approximations possibles parmi les plus courantes.

Approximation normale

Posons $W_1 = \frac{S_1^2}{n}$ et $W_2 = \frac{S_2^2}{m}$; on prend maintenant, comme statistique de test, la v.a.r. :

$$T^* = \frac{\overline{X} - \overline{Y}}{\sqrt{W_1 + W_2}}.$$

Pour des valeurs assez grandes de n et de m (disons lorsque chacune vaut au moins 30) et sous l'hypothèse d'égalité des moyennes, T^* est approximativement distribuée suivant une loi normale réduite. Le test consiste alors à comparer la valeur de $|T^*|$ à la valeur critique d'une loi normale réduite. Cette approximation est bien connue, mais la suivante est, en général, meilleure.

Approximation de Satterthwaite

Pour des valeurs relativement petites de n et m (inférieures à celles nécessitées pour l'approximation normale), on peut montrer que T^* est approximativement distribuée suivant une loi de Student dont le degré de liberté, ν^*, est donné par la formule suivante :

$$\nu^* = \mathrm{AE}\left(\frac{(W_1 + W_2)^2}{\frac{W_1^2}{n-1} + \frac{W_2^2}{m-1}} \right),$$

où AE(.) désigne l'approximation entière. Cette approximation est connue sous le nom d'approximation de Satterthwaite.

Dans la procédure `ttest`, SAS calcule systématiquement la valeur de la statistique T^* (quelles que soient les tailles des 2 échantillons) et la *p-value* associée à $|T^*|$ pour une loi de Student à ν^* d.d.l. Ces valeurs sont fournies, dans la sortie standard, à la ligne `Unequal` du tableau `T-Tests`.

Remarque 8 Degrés de liberté décimaux. *On notera que SAS ne fait pas l'approximation entière de ν^* car il peut réaliser des calculs sur des lois de Student dont le d.d.l. est décimal. En fait, on peut très bien définir une généralisation de la loi de Student (ou de khi-deux, ou de Fisher) en considérant des d.d.l. décimaux, la difficulté étant de faire des calculs avec une telle loi ; SAS peut les faire grâce aux méthodes numériques utilisées pour calculer les intégrales intervenant dans ces calculs.*

2.4 Le test de Mann-Whitney

Dans le cas où l'on ne peut pas faire l'hypothèse de normalité sur les distributions des observations des 2 échantillons, on peut utiliser le test non paramétrique de Mann-Whitney, que l'on trouve parfois, dans la littérature statistique anglo-américaine, sous le nom de *Wilcoxon rank-sum test*.

Les 2 échantillons $(X_1, \ldots, X_i, \ldots, X_n)$ et $(Y_1, \ldots, Y_j, \ldots, Y_m)$ sont toujours considérés indépendants entre eux, chacun étant i.i.d., les distributions dont ils sont issus étant seulement supposées **continues**. L'hypothèse nulle est l'égalité de ces distributions, que l'on traduit généralement par $\{H_0 : F_X = F_Y\}$, F_X et F_Y désignant les fonctions de répartition respectives de ces distributions. La seule hypothèse alternative qui sera considérée ici est $\{H_1 : F_X \neq F_Y\}$ (les autres alternatives envisageables sont assez complexes).

Le principe du test de Mann-Whitney consiste alors à ranger par ordre croissant l'ensemble des observations mélangées des 2 échantillons, à leur affecter un rang et à calculer séparément les sommes des rangs des observations provenant de chacun des échantillons, notées Q_X et Q_Y. Désignant par Q_X la somme des rangs des observations de l'échantillon de plus petite taille (autrement dit, supposant $n \leq m$), la statistique du test de Mann-Whitney s'écrit alors :

$$U = Q_X - \frac{n(n+1)}{2}.$$

On peut vérifier que, sous H_0, l'espérance mathématique de U est $\frac{nm}{2}$ et l'on est donc amené à rejeter l'hypothèse nulle H_0 ssi $|U - \frac{nm}{2}|$ est trop grande.

En fait, compte tenu des tables statistiques dont on dispose dans la pratique, on calcule plutôt la statistique U^* définie comme suit :

$$U^* = \inf\,(U, U'),\ \text{avec :}\ U' = Q_Y - \frac{m(m+1)}{2}.$$

On rejette alors H_0 ssi U^* est trop petite (inférieure ou égale à la valeur critique).

On peut également faire, lorsque n et m sont suffisamment grands, une approximation normale de la statistique U. Sous H_0, on peut vérifier :

$$\mathbb{E}(U) = \frac{nm}{2}\ ;\ Var(U) = \tau^2 = \frac{nm(n+m+1)}{12}.$$

Par conséquent, la v.a.r. $\dfrac{U - \frac{nm}{2}}{\tau}$ est centrée et réduite et l'on peut montrer qu'elle est asymptotiquement normale (précisement, lorsque l'on a : $n \longrightarrow +\infty$, $m \longrightarrow +\infty$, $\dfrac{n}{m} \longrightarrow a$, $0 < a < +\infty$). L'approximation normale du test de Mann-Whitney consiste ainsi à déterminer la statistique

$$Z = \frac{|U - \frac{nm}{2}| - \frac{1}{2}}{\tau}$$

(la quantité $-\frac{1}{2}$ constitue ce que l'on appelle la correction de continuité ; elle améliore l'approximation) et à rejeter l'hypothèse nulle ssi la valeur calculée de Z dépasse la valeur critique de la loi normale réduite. Cette approximation est valable même pour de faibles valeurs de n et m (de l'ordre de 10). Comme on dispose de tables allant jusqu'à des tailles de 16, on l'utilisera pour $n \geq 8$ et $m \geq 17$.

La procédure `npar1way` de SAS réalise le test de Mann-Whitney en faisant systématiquement cette approximation normale (autrement dit, elle ne calcule que Z). Pour des valeurs assez petites de n et m, on préfèrera donc utiliser les tables. Dans la procédure `npar1way`, on doit rajouter l'option `wilcoxon`, puis les commandes `class` et `var`, comme dans la procédure `ttest`.

On notera, lorsqu'il y a des ex-æquo dans l'un ou l'autre des échantillons, que SAS effectue une correction mineure de τ par rapport à la formule donnée plus haut.

Remarque 9 Archivage des sorties. *Comme avec* `ttest`*, il n'est pas possible, avec la procédure* `npar1way`*, d'archiver les résultats dans une table SAS.*

2.5 Illustrations

Les fichiers suivants

```
~baccini/tpsas/exolim/data/habil.txt
~baccini/tpsas/exolim/data/habil.don
~baccini/tpsas/exolim/exo11_2a.sas
```

proposent, pour 2 échantillons indépendants de tailles respectives 8 et 10, la description, les données et le programme SAS permettant d'illustrer les tests de Fisher, de Student (test exact, car l'hypothèse d'égalité des variances n'est pas rejetée) et de Mann-Whitney. On peut les copier et mettre en œuvre le programme SAS.

Même chose avec les fichiers

```
~baccini/tpsas/exolim/data/note2.txt
~baccini/tpsas/exolim/data/note2.don
~baccini/tpsas/exolim/exo11_2b.sas
```

pour un exemple fictif de 2 séries indépendantes de tailles 10 et 15. Dans ce cas, l'approximation de Satterthwaite du test de Student est recommandée.

2.6 Remarque sur le cas de k échantillons

Une généralisation naturelle de ce qui précède est l'étude de l'homogénéité de k échantillons, $k \geq 3$.

Si l'on fait l'hypothèse que les k échantillons sont indépendants, chacun étant constitué des observations de lois i.i.d., normales et de même variance, l'étude de l'homogénéité des échantillons revient à la comparaison de leurs moyennes. Pour ce faire, on réalise le test de l'hypothèse nulle de l'égalité de ces moyennes. C'est ce que l'on appelle, en statistique, l'analyse de variance à 1 facteur. On met cette méthode en œuvre dans SAS en utilisant soit la procédure `anova` (*analysis of variance*), soit la procédure `glm` (*general linear model*).

Si l'on ne fait aucune hypothèse sur la distribution des observations, tout en les supposant i.i.d., on peut alors faire le test non paramétrique de Kruskal-Wallis. Ce dernier se met en œuvre dans SAS en utilisant la procédure `npar1way`.

Nous ne développons pas davantage le problème de k échantillons.

3 Problèmes d'ajustement

Le problème considéré ici est celui du contrôle d'une hypothèse du type : les observations réalisées proviennent d'une distribution donnée (le plus souvent, il s'agit de la loi normale).

Il existe divers moyens empiriques permettant de contrôler, au moins en première approximation, une telle hypothèse sur une série d'observations : connaissance a priori du phénomène étudié, représentations graphiques, droite de Henry...

Mais, de façon plus précise, il existe aussi des tests, dits d'ajustement, qui permettent de tester l'hypothèse nulle selon laquelle les données proviennent d'une loi particulière : tests du khi-deux, de Kolmogorov-Smirnov, de Cramér-von Mises...

Le plus courant de ces tests est celui de Kolmogorov-Smirnov. On peut le mettre facilement en œuvre dans SAS et il est détaillé ci-dessous.

3.1 Principe du test de Kolmogorov-Smirnov

On considère à nouveau un échantillon unique $(X_1, \ldots, X_i, \ldots, X_n)$ de n v.a.r. i.i.d. selon une loi de probabilité supposée **continue**, dont on notera F la fonction de répartition. On considère par ailleurs la fonction de répartition F_0 d'une loi de probabilité particulière. On souhaite tester, avec un niveau α fixé, l'hypothèse nulle $\{H_0 : F = F_0\}$ contre l'alternative $\{H_1 : F \neq F_0\}$ (encore une fois, les autres alternatives, plus complexes, ne seront pas considérées ici).

La statistique du test de Kolmogorov-Smirnov est définie par

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|,$$

où F_n désigne la fonction de répartition empirique de l'échantillon $(X_1, \ldots, X_i, \ldots, X_n)$.

En tout point x de $\mathbb{R}$, cette dernière vaut $F_n(x) = \dfrac{n(x)}{n}$, où $n(x)$ représente le nombre d'observations de l'échantillon inférieures ou égales à x. Par ailleurs, la fonction de répartition F_0 considérée sous l'hypothèse nulle doit être complètement spécifiée, ce qui signifie que, si elle dépend de paramètres inconnus, ceux-ci doivent être estimés (pour des raisons de convergence, on doit alors utiliser la technique du maximum de vraisemblance).

Si l'on note $x_{(i)}, i = 1, \ldots, n$, les valeurs observées de l'échantillon rangées par ordre croissant, n_i le nombre d'observations inférieures ou égales à $x_{(i)}$ et n_{i-1} le nombre d'observations strictement inférieures à $x_{(i)}$, on sait que, pour déterminer D_n, il suffit, dans un premier temps, de déterminer les écarts $|\dfrac{n_{i-1}}{n} - F_0(x_{(i)})|$ et $|\dfrac{n_i}{n} - F_0(x_{(i)})|$ en chaque valeur $x_{(i)}$, puis de retenir le plus grand des 2, noté S_i. On obtient alors :

$$D_n = \sup_{1 \leq i \leq n} S_i.$$

Pour des lois de probabilité usuelles, les quantités $F_0(x_{(i)}), i = 1, \ldots, n$, peuvent être déterminées au moyen soit d'une table soit d'un logiciel statistique.

3.2 Mise en œuvre dans SAS

Réalisation d'un programme spécifique

Il n'est pas difficile d'écrire un programme SAS permettant de calculer la statistique D_n pour une loi de probabilité courante. Il suffit ensuite d'utiliser une table du test de Kolmogorov-Smirnov pour conclure. À titre d'illustration, on pourra copier la macro `~baccini/macros/kstest`.

Elle permet de tester l'ajustement d'une série d'observations à la loi normale. Les paramètres en sont les suivants :
`donn`, nom de la table SAS contenant les données ;
`varX`, nom de la variable, dans la table `donn`, sur laquelle porte le test ;
`n`, nombre d'observations, donc de lignes de la table `donn` ;
`moy`, moyenne estimée des observations ;
`sigma`, écart-type estimé des observations.

Les valeurs `moy` et `sigma` peuvent être déterminées au moyen de la procédure `means`.

Utilisation de SAS/INSIGHT

La façon la plus attractive de réaliser le test de Kolmogorov-Smirnov dans SAS consiste en fait à utiliser SAS/INSIGHT. Pour cela, disposant encore d'une table SAS appelée `donn` et contenant les données avec la variable à tester dans la colonne `varX`, on ouvre SAS/INSIGHT comme indiqué à la fin de la feuille de T.P. numéro 09. On "clique" ensuite sur `varX`, puis sur le menu `Analyse/Distribution (Y)` : une nouvelle fenêtre, appelée `Distribution`, s'ouvre. Elle contient 2 graphiques (diagramme en boîte et histogramme) ainsi que 2 tableaux relatifs à `varX`. En "cliquant" alors sur le menu `Curves/Test for a Specific Distribution...` de cette dernière fenêtre, une nouvelle fenêtre temporaire s'ouvre. On doit alors choisir la distribution à laquelle les données seront ajustées (parmi les distributions normale, lognormale, exponentielle et de Weibull) et donner les paramètres de la distribution choisie (moyenne et écart-type dans le cas de la normale) afin d'obtenir, dans la fenêtre `Distribution`, d'une part un nouveau graphique contenant les courbes superposées des fonctions $F_n(x)$ et $F_0(x)$, d'autre part la statistique D_n ainsi que la *p-value* associée, ce qui permet de conclure sur H_0.

Si l'on souhaite archiver les résultats de cette analyse, dans la fenêtre `Distribution`, faire d'abord `File/Save/Graphics files` ; dans la fenêtre temporaire qui s'ouvre, "cliquer" sur `ps` et sur `One Per File` : on archive ainsi le graphique des courbes superposées dans un fichier au format *postscript* appelé *scatter*. Ensuite, faire `File/Save/Tables` : les résultats numériques de l'analyse viennent se rajouter dans la fenêtre `OUTPUT` de SAS.

3.3 Illustrations

On pourra illustrer la réalisation du test de Kolmogorov-Smirnov en copiant les fichiers

```
~baccini/tpsas/exolim/data/normal.txt
~baccini/tpsas/exolim/data/normal.don
~baccini/tpsas/exolim/exo11_3.sas
```

relatifs à un petit exemple fictif de 12 observations.

On pourra également utiliser le fichier

```
~baccini/tpsas/exolim/data/notes.don
```

contenant des données réelles plus volumineuses.