Séance n°5 : Test de non-conformité (test de Kolmogorof)

Module de Génie Informatique en Licence 3 d'électronique

`Laboratoire LTSI - UMR INSERM 642 - Université de Rennes1`

Supposons que l'on dispose d'une fonction sur calculateur qui permette de générer des réalisations d'une variable aléatoire X de loi donnée par la fonction de répartition théorique F_X. Se pose alors la question de vérifier expérimentalement cette hypothèse, que l'on appellera *hypothèse nulle* et que l'on notera H_0, à partir d'une suite de N_e réalisations x_n de X, tirées indépendamment.

Une première manière de répondre à cette question est de considérer, comme nous l'avons fait dans la séance n°4, les représentations graphiques des fonctions de répartition expérimentale et théorique et de les comparer visuellement. Ceci peut cependant être considéré comme une méthodologie insatisfaisante sur le plan qualitatif, et peu convaincante quand N_e n'est pas très élevé.

Une approche plus *carrée* consiste à introduire une procédure qui admet en entrée les valeurs x_1,...,x_{N_e} et qui en sortie fournit une décision : soit l'acceptation, soit le refus de H_0. Une telle procédure sera appelée test de conformité des observations à la loi théorique correspondant à H_0. Pour cela Kolmogorof a proposé de comparer quantitativement la fonction de répartition théorique attendue (ou non) F à la fonction de répartition expérimentale $\widehat{F}_X$ au moyen d'une distance définie comme suit :

$$d_X = d(\widehat{F}_X, F) = \sup_{\alpha \in \mathbb{R}} \left\{ \left| \widehat{F}_X(\alpha) - F(\alpha) \right| \right\}$$

qui mesure l'écart maximal, quand α parcourt $\mathbb{R}$, entre les valeurs de la fonction de répartition théorique et celles de la fonction de répartition expérimentale. Le test de conformité de Kolmogorof est alors décrit par :

$$\boxed{\begin{array}{l} \text{si } d_X > \mu_0 \text{ alors } H_0 \text{ est rejetée,} \\ \text{si } d_X \leq \mu_0 \text{ alors } H_0 \text{ est acceptée,} \end{array}}$$

où la valeur de seuil μ_0 doit être préalablement fixée par l'utilisateur. En effet, on comprend bien que si la mesure d_X est trop grande alors l'ypothèse H_0 ne peut être retenue. Pour obtenir d'un point de vue pratique la mesure d_X à partir des N_e données x_n on peut montrer qu'il suffit de calculer :

$$d_X = \max_{k \in \{1,2,\ldots,N_a\}} \left\{ \max \left\{ \left| \widehat{F}_X(\alpha_{k+1}) - F(\alpha_k) \right|, \left| \widehat{F}_X(\alpha_k) - F(\alpha_k) \right| \right\} \right\}$$

où rappelons-le α_k est donné dans l'énoncé de TP de la séance n°4. L'approche décisionnelle décrite par l'encadrée situé ci-dessus est purement déterministe. Une autre approche, quant-à-elle stochastique, profite du résultat suivant démontré par Kolmogorof :

Autrement dit, la fonction de répartition de la variable D_Y, sous l'hypothèse H_0 pour Y (i.e. lorsque D_Y est calculée en utilisant la fonction de répartition théorique de Y), la mesure de probabilité $P(D_Y > \lambda/H_0)$ est indépendante de la loi de Y pour toute valeur réelle λ. Kolmogorov a d'ailleurs donné, pour N_e suffisamment grand, une approximation de cette mesure de probabilité en fonction du paramètre λ' défini par $\lambda' = \lambda\sqrt{N_e}$.

Soit $\widehat{F}_Y$ le processus aléatoire de longueur N_e dont une réalisation représente la fonction de répartition expérimentale obtenue à partir de N_e réalisations indépendantes d'une variable aléatoire Y de fonction de répartition théorique F_Y. Pour N_e fixé, la variable aléatoire D_Y définie par $D_Y = d(\widehat{F}_Y, F_Y)$ suit une loi de probabilité qui est toujours la même quelle que soit la loi de Y.

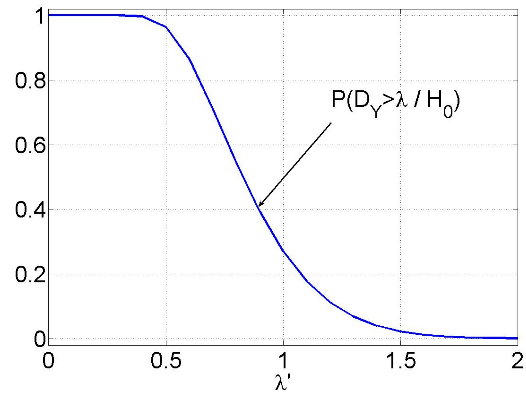

FIG. 1 – Graphe de l'approximation de Kolmogorov où $\lambda' = \lambda\sqrt{N_e}$ pour N_e suffisamment grand

Le graphe de cette approximation est représenté sur la figure **??**. Comme nous le voyions sur la figure **??**, la mesure de probabilité décroit et ce rapidement lorsque λ' est supérieur à 0.5. En outre, nous constatons que la probabilité que D_Y prenne des valeurs supérieures à 1.5 est très faible. Par conséquent, comme la mesure de probabilité précédente – rappelons-le – est indépendante de la loi de la variable considérée, la valeur d_X pour une valeur de $\sqrt{N_e}d_X$ supérieure à 1.5 aurait peu de chance d'être prise par la variable aléatoire D_X sous H_0, ce qui nous pousserait à rejeter H_0 si une telle valeur de d_X devait ressortir de nos N_e réalisations x_n.

Notons que nous avons choisi 1.5 comme valeur de seuil pour $\sqrt{N_e}d_X$ dans l'explication précédente. En pratique, on se donne plutôt un seuil de probabilité, probabilité en dessous de laquelle on décide de rejetter l'hypothèse H_0. Cette probabilité est généralement appelée *risque de première espèce*[1] ; elle correspond en fait au risque de rejeter à tort H_0 si H_0 est en fait vraie, d'où l'intérêt de la choisir. D'après l'approximation de Kolmogorov, pour N_e suffisamment grand, nous avons $P(D_Y > \frac{1.5}{\sqrt{N_e}}/H_0) \approx 0.022$, ce qui correspond à un risque seuil de 2.2%. Souvent, un risque de 5% est considéré comme acceptable (c'est-à-dire que dans 5% des cas quand H0 est vraie, l'expérimentateur se trompera et la rejettera). Mais le choix du risque à employer dépendra de la certitude désirée et de la vraisemblance des alternatives.

1) Programmer sous Matlab la distance entre une fonction de répartition théorique et une fonction de répartition expérimentale.

2) Calculer sous Matlab la table de mesures $P(D_Y > \lambda/H_0)$ pour une variable aléatoire Y de loi uniforme sur $[0, 1]$ et ce pour différentes valeurs de λ et de N_e (on pourra se limiter à des valeurs de $\lambda\sqrt{N_e}$ supérieures à 1 et considérer $N_e = 10, 100, 1000, 10000$).

3) Evaluer la non-conformité des fonctions de répartition expérimentale et théorique d'une loi exponentielle de paramètre $a = 0.2$ en fonction du nombre de réalisations N_e et pour un risque de deuxième espèce de 5% en prenant comme fonction de répartition théorique i) celle d'une loi exponentielle de paramètre $a = 0.2$ et ii) celle d'une loi exponentielle de paramètre $a = 0.8$.

[1]La probabilité pour que H_0 soit acceptée alors qu'elle est fausse est appelée *risque de deuxième espèce*. C'est le risque de ne pas rejeter H_0 quand on devrait la rejeter. Sa valeur dépend du contexte, et est très difficilement évaluable (voire impossible à évaluer), c'est pourquoi seul le risque a est utilisé comme critère de décision.