Le test d'Anderson-Darling

Étude du cheminement des auteurs

Reda Mdair

2024

1 Introduction

Introduit en 1954, le test d'Anderson-Darling est une approche non paramétrique permettant de vérifier si un échantillon suit bien une certaine loi de probabilité. Son cas d'application le plus populaire est la vérification de la normalité.

Initialement, ce test supposait de connaître les paramètres de l'échantillon. Quelques années plus tard, le statisticien Michael A. Stephens a appronfondi cette approche en utilisant des estimations pour des paramètres inconnus. Avec ces avancées, il devient aujourd'hui un outil complet et puissant dans les tests de conformité.

Ayant les mêmes principes que le test de Kolmogorov-Smirnov, il est considéré comme une variante de ce dernier. La principale différence est dans la pondération de la statistique de test. Celle de Kolmogorov-Smirnov est plus sensible aux écarts près de la médiane, à l'inverse de celle d'Anderson-Darling qui est plus impactée par les valeurs extrêmes.

L'objectif de ce projet est d'étudier les deux articles publiés par les auteurs de ce test. Nous interpréterons les principales informations dans le but de répondre à la problématique : quels sont les avantages et les limites du test d'Anderson-Darling par rapport aux autres approches classiques ?

2 Construction du test

Soit $X = (X_1, \ldots, X_n)$ un échantillon de n variables aléatoires i.i.d d'une loi inconnue P. On cherche à tester si cette distribution suit une loi spécifiée P^1 continue. À partir d'une observation $x = (x_1, \ldots, x_n)$, on pose les hypothèses :

$$(H_0) : P = P^1 \qquad\qquad (H_1) : P \neq P^1$$

En notant respectivement F et F^1 les fonctions de répartitions associées à P et P^1, Anderson et Darling ont proposé les deux statistiques de test :

$$W_n^2(X) = n \int_{\mathbb{R}} [F_n(x) - F^1(x)]^2 \psi[F^1(x)]\, dF^1 \qquad (1.0)$$

$$K_n(X) = \sup_{x \in \mathbb{R}} \sqrt{n} |F_n(x) - F^1(x)| \sqrt{\psi[F^1(x)]} \qquad (2.0)$$

où $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq x}$ est la fonction de répartition empirique de l'échantillon, et $\psi(t)_{0 \leq t \leq 1}$ est une fonction de pondération positive pré-établie. Cette dernière permet de mettre l'accent sur certaines parties de la distribution.

L'objectif est alors de trouver les lois asymptotiques associées à W_n^2 et K_n. Ceci permettra de calculer leurs points critiques respectifs z_1 et z_2, de manière précise pour des échantillons de grande taille.

La principale clé réside notamment dans le choix de la fonction $\psi(t)_{0\leq t\leq 1}$. Celle-ci est laissée au choix du statisticien afin de pondérer les écarts selon l'importance accordée aux différentes parties de la distribution. Par exemple, pour $\psi(t) = 1$, on retrouve les critères de Cramér-Von-Mises via W_n^2 et de Kolmogorov-Smirnov via K_n.

En ordonnant les observations, les statistiques de test se simplifient. Soit $(X_{(1)}, \ldots, X_{(n)})$ la statistique d'ordre associé à l'échantillon X. Étant donné que $F^1(x)$ est supposée continue, le changement de variable $u = F^1(x)$ est bien défini, ce qui induit les observations $u_i = F^1(x_i), i \in \{1, \ldots, n\}$. Sous l'hypothèse (H_0), ces dernières peuvent être considérées comme des tirages selon la loi uniforme sur $[0, 1]$. En notant $G_n(u)$ la fonction de répartition empirique des $(u_i)_{1\leq i\leq n}$, on obtient les nouvelles formules :

$$W_n^2(X) = n \int_0^1 [G_n(u) - u]^2 \psi(u)\, du \qquad (1.1)$$

$$K_n(X) = \sup_{0\leq u\leq 1} \sqrt{n}|G_n(u) - u|\sqrt{\psi[u]} \qquad (2.1)$$

Afin d'étudier les propriétés asymptotiques des statistiques, Anderson et Darling définissent : $Y_n(u) = \sqrt{n}[G_n(u) - u], u \in [0, 1]$. L'avantage de cette transformation est que la loi jointe de $Y_n(u_1), \ldots, Y_n(u_k)$ (pour $u_1, \ldots, u_k$ fixés) approche la loi normale multivariée à l'infini :

$$(Y_n(u_1), \ldots, Y_n(u_k)) \xrightarrow[n\to\infty]{\mathcal{L}} (y(u_1), \ldots, y(u_k))$$

où les $y(u_i)$ sont des variables de loi normale centrée et de covariance :

$$Cov(y(u), y(v)) = E[y(u)y(v)] = min(u, v) - uv$$

Cette étape est cruciale dans leurs recherches vers la loi asymptotique, car la loi normale permet l'application de nombreux théorèmes sur les processus stochastiques. Dans cette voie, ils définissent ainsi les deux quantités :

$$A_n(z) = \mathbb{P}(W_n^2 \leq z) = \mathbb{P}\left(\int_0^1 Y_n^2(u)\psi(u)\, du \leq z\right)$$

$$B_n(z) = \mathbb{P}(K_n \leq z) = \mathbb{P}\left(\sup_{0\leq u\leq 1} |Y_n(u)|\sqrt{\psi[u]} \leq z\right)$$

L'objectif est alors de trouver :

$$A_n(z) = a(z) = \mathbb{P}\left(\int_0^1 y^2(u)\psi(u)\, du \leq z\right)$$

$$B_n(z) = b(z) = \mathbb{P}\left(\sup_{0\leq u\leq 1} |y(u)|\sqrt{\psi[u]} \leq z\right)$$

afin d'obtenir des bases solides pour le calcul des distributions limite de W_n^2 et K_n.

Le théorème de Donsker leur permet de conclure que $A_n(z) = a(z)$. Ce dernier est construit sur l'hypothèse que la fonction $\psi(u)$ est bornée. Les deux auteurs parviennent à le généraliser, sous certaines hypothèses, à des fonctions de poids $\psi(u)$ plus courantes.
Il en suit ainsi plusieurs tests sur la fonction ψ à travers des exemples. Pour la statistique de test W_n^2, ils étudient :

$$\psi_1(u) \equiv 1 \qquad\qquad \psi_2(u) = \frac{1}{u(1-u)}$$

Le second choix se révèle alors plus intéressant. En effet, celui-ci permet d'accentuer les écarts aux queues de la distribution. Autrement dit, la fonction $\psi_2(u) = \frac{1}{u(1-u)}$ détecte plus facilement les déviations observées près de 0 et de 1. Le premier choix, quant à lui, ne privilégie aucune partie de la distribution et accorde un poids uniforme sur tout l'intervalle [0,1].
Concernant la statistique de test K_n, les fonctions étudiées sont plus techniques :

$$\psi_3(u) = q_k \geq 0 \qquad \forall u_k \leq u \leq u_{k+1}, \quad k \in \{1, \ldots, n\}, \quad u_0 = 0, u_{n+1} = 1$$

$$\psi_4(u) = \frac{1}{u(1-u)} \mathbb{1}_{0<a\leq u\leq b<1}$$

Bien que ces choix semblent intéressants, l'interprétation des résultats s'avère très difficile. De plus, certaines incertitudes théoriques réduisent la fiabilité de la loi asymptotique associée à K_n. Cette dernière sera donc écartée.

C'est finalement la statistique de test W_n^2 qui est retenue dans le second article d'Anderson et Darling, avec la fonction de poids $\psi(u) = \frac{1}{u(1-u)}$. Sa formule donnée à (1.0) se simplifie alors sous :

$$\boxed{W_n^2(X) = -n - \tfrac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln\left(F^1(X_{(i)})\right) + \ln\left(F^1(X_{(n-i+1)})\right)\right]}$$

avec sa fonction caractéristique :

$$\phi(t) = \sqrt{\frac{-2\pi i t}{cos(\frac{\pi}{2}\sqrt{1+8it})}}$$

Ces données ont ainsi permis aux auteurs de construire la table statistique qui porte désormais leurs noms.

Remarque : En 1979, le statisticien Michael A. Stephens a proposé, dans le cas où les paramètres sont inconnus, les estimateurs $\mu = \bar{X}$ et $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$ pour le test de normalité. Il a ensuite défini la statistique de test modifiée :

$$W^* = W_n^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$$

3 Discussions

Rappelons la formule de la statistique associée à l'approche de Kolmogorov-Smirnov :

$$T_{KS}(X) = \sqrt{n} \max_{1 \leq i \leq n} \left\{ max \left\{ \left| F^1(X_{(i)}) - \frac{i}{n} \right|, \left| F^1(X_{(i)}) - \frac{i-1}{n} \right| \right\} \right\}$$

Sa compréhension est très simple : la distance est basée sur l'écart le plus important entre les fonctions de répartition empirique et hypothétique.
Une question se pose alors : l'information exploitée par cette statistique est-elle suffisante ? Afin d'y répondre, il est intéressant d'introduire le test de Cramér-Von-Mises (1928) basé sur la distance :

$$T_{CVM}(X) = n \int_{\mathbb{R}} \left[F_n(x) - F^1(x) \right]^2 \, dF^1$$

Cette fonction semble plus qualitative que celle de K-S. En effet, plutôt que de mesurer le plus grand écart vertical, elle mesure l'écart d'ajustement sur tout l'intervalle des courbes. Ainsi, elle peut paraître plus fiable pour la construction d'un test d'ajustement.

Cependant, cette approche de C-V-M relève une légère imprécision. En effet, elle ne prend pas en compte que la fonction de répartition empirique est plus variable vers la médiane plutôt que les extrémités. C'est par ce raisonnement que l'on peut deviner les intentions d'Anderson et Darling via leur approche. La fonction $\psi(u) = \frac{1}{u(1-u)}$ (inverse de la variance de la f.d.r empirique) vient alors pondérer la statistique $T_{CVM}(X)$ afin d'accorder plus de poids aux extrémités.
Il est souvent préférable de privilégier la précision dans les régions peu denses (extrémités) plutôt qu'au niveau de la région dense (médiane).

La technique d'Anderson-Darling est donc la mieux adaptée pour les tests sur les lois à queues lourdes. Par exemple, on peut citer la loi de Pareto, celle de Cauchy ou encore la loi de Lévy. De plus, de nombreux problèmes demandent une attention particulière aux régions extrêmes (comme les événements rares) et l'approche d'A-D y répond efficacement.

Cependant, pour vérifier la conformité à une loi uniforme, le test de Kolmogorov-Smirnov est clairement avantageux. Étant donné que la f.d.r hypothétique croît sous forme de droite, alors mesurer les écarts cumulés de manière égale sur l'ensemble de l'intervalle y est adapté. Une utilisation d'Anderson-Darling est inutile dans ce cas et sa distance calculée peut être excessivement sensible.

4 Synthèse des résultats

Le test d'Anderson-Darling est un outil très puissant qui est venu renforcer les options de choix dans les tests de conformité. Sa formule mesure les écarts sur l'ensemble de l'intervalle des distributions empiriques et hypothétiques, en accordant davantage de poids aux extrémités. Il est donc particulièrement adapté pour les lois à queues lourdes. Également, il peut servir d'analyse des événements rares qui nécessitent une attention sur les observations éloignées de la médiane.

Cependant, il reste important de savoir alterner avec d'autres tests tels que K-S et C-V-M. L'étude de la conformité à la loi uniforme illustre bien les limites de l'approche d'Anderson-Darling. Dans cette situation, ce test peut fournir une statistique trop élevée en amplifiant le poids sur des régions non critiques. Ceci le rendrait donc trop prudent, en rejetant facilement l'hypothèse (H_0) de conformité.

Les deux articles liés à cette étude représentent bien les compromis auxquels sont confrontés les statisticiens dans la construction des tests d'hypothèses. Entre choisir les zones de mesure d'écarts, choisir le poids adapté à chacune de ces régions et minimiser le coût computationnel, la recherche devient très complexe.

5 Bibliographie

[I] Anderson, T. W., and Darling, D. A. - "Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes" - Annals of Mathematical Statistics, 23 (1952), 193-212.

[II] "A Test of Goodness of Fit" - T. W. Anderson ; D. A. Darling - Journal of the American Statistical Association, Vol. 49, No. 268. (Dec., 1954), pp. 765-769.

[III] " The Anderson-Darling Statistic " - Michael A. Stephens - Technical report no. 39 - October 31, 1979.
`https://apps.dtic.mil/sti/pdfs/ADA079807.pdf`