Test de Kolmogorov-Smirnov : Théorie et Propriétés

RAHMANI Naceur

16/10/2025

Introduction

Le test de Kolmogorov-Smirnov (K-S) est une méthode statistique non paramétrique qui permet de :

- Comparer deux distributions de probabilité.
- Tester si un échantillon suit une distribution théorique spécifique.

C'est un test non paramétrique, ce qui signifie qu'il ne nécessite aucune hypothèse sur la forme de la distribution sous-jacente des données.

Formulation du Test

Le test de Kolmogorov-Smirnov peut être formulé de deux manières principales :

- **Test à un échantillon** : Tester si un échantillon suit une distribution théorique.
- **Test pour deux échantillons** : Comparer deux échantillons pour déterminer s'ils proviennent de la même distribution.

Test à un échantillon

Ce test est utilisé pour tester si un échantillon $X_1, X_2, \ldots, X_n$ suit une distribution théorique $F(x)$.

Hypothèses du test :

- **Hypothèse nulle (H)** : Les données suivent la distribution théorique $F(x)$.
- **Hypothèse alternative (H)** : Les données ne suivent pas la distribution théorique $F(x)$.

Statistique du test :

$$D_n = \sup_x |F_n(x) - F(x)|$$

où $F_n(x)$ est la fonction de répartition empirique et $F(x)$ la fonction de répartition théorique.

Test à un échantillon (suite)

Décision :

- Si D_n est plus grand que la valeur critique obtenue à partir de la table de Kolmogorov-Smirnov, l'hypothèse nulle est rejetée.
- Le test permet de conclure que l'échantillon ne suit pas la distribution théorique spécifiée.

Test pour deux échantillons

Ce test est utilisé pour comparer deux échantillons et déterminer s'ils proviennent de la même distribution.

Hypothèses du test :

- **Hypothèse nulle (H)** : Les deux échantillons proviennent de la même distribution.
- **Hypothèse alternative (H)** : Les deux échantillons proviennent de distributions différentes.

Statistique du test :

$$D_{12} = \sup_x |F_{n_1}(x) - F_{n_2}(x)|$$

où $F_{n_1}(x)$ et $F_{n_2}(x)$ sont les fonctions de répartition empiriques des deux échantillons.

Test pour deux échantillons (suite)

Décision :

- Si D_{12} dépasse la valeur critique pour un certain niveau de signification α, l'hypothèse nulle est rejetée.
- Les deux échantillons ne proviennent donc probablement pas de la même distribution.

Propriétés du Test

- **Non paramétrique** : Le test ne dépend pas d'hypothèses sur la forme de la distribution sous-jacente.
- **Puissance** : Le test est sensible aux différences dans la forme des distributions, mais moins puissant pour des différences petites, surtout dans les queues des distributions.
- **Consistance** : Le test devient plus fiable avec des tailles d'échantillon plus grandes.
- **Symétrie** : Le test pour deux échantillons est symétrique. La comparaison de A et B donne la même statistique que la comparaison de B et A.

Calcul de la Statistique du Test

Test à un échantillon :

- Triez les données de l'échantillon $X_1, X_2, \ldots, X_n$.
- Calculez la fonction de répartition empirique $F_n(x)$ pour chaque observation x.
- Calculez la fonction de répartition théorique $F(x)$.
- Calculez la statistique $D_n = \sup_x |F_n(x) - F(x)|$.

Test pour deux échantillons :

- Calculez les fonctions de répartition empiriques $F_{n_1}(x)$ et $F_{n_2}(x)$.
- Calculez la statistique $D_{12} = \sup_x |F_{n_1}(x) - F_{n_2}(x)|$.

Applications du Test

- **Vérification de la normalité** : Tester si un échantillon suit une distribution normale.
- **Comparaison de deux échantillons** : Tester si deux groupes de données proviennent de la même distribution.
- **Analyse des résidus** : Tester si les résidus d'un modèle de régression suivent une distribution spécifique.

Exemple Pratique

Supposons que nous avons un échantillon de $n = 50$ observations provenant d'un test psychologique, et nous souhaitons tester si cet échantillon suit une distribution normale avec une moyenne de 100 et un écart-type de 15.

- La fonction de répartition théorique serait la fonction de répartition de la loi normale $F(x)$ avec les paramètres spécifiés.
- Nous calculons la fonction de répartition empirique $F_n(x)$ de l'échantillon.
- Ensuite, nous calculons la statistique D_n et la comparons à la valeur critique pour un niveau de signification donné.

Conclusion

Le test de Kolmogorov-Smirnov est un outil statistique puissant et flexible qui permet de :

- Tester la conformité des données à une distribution théorique.
- Comparer deux distributions empiriques.

Il est largement utilisé en raison de sa nature non paramétrique et de sa capacité à s'appliquer à divers types de distributions.

Région Critique du Test de Kolmogorov-Smirnov

RAHMANI Naceur

Statistique du Test de Kolmogorov-Smirnov

La statistique D mesure la plus grande différence entre la fonction de répartition empirique $F_n(x)$ et la fonction de répartition théorique $F(x)$.

- Pour un test à un échantillon :

$$D = \sup_x |F_n(x) - F(x)|$$

où $F_n(x)$ est la fonction de répartition empirique et $F(x)$ la fonction de répartition théorique.

- Pour un test à deux échantillons :

$$D = \sup_x |F_{n_1}(x) - F_{n_2}(x)|$$

Cette statistique mesure la plus grande distance entre les deux fonctions de répartition (empirique et théorique, ou entre deux échantillons).

Région Critique du Test

La **région critique** est l'ensemble des valeurs de D pour lesquelles l'hypothèse nulle (H) est rejetée.

- **Hypothèse nulle (H)** : Les données suivent la distribution théorique $F(x)$ (ou les deux échantillons proviennent de la même distribution).
- **Hypothèse alternative (H)** : Les données ne suivent pas la distribution théorique (ou les deux échantillons proviennent de distributions différentes).

Le test rejette H_0 si la statistique D est supérieure à la valeur critique D_α, correspondant au niveau de signification α.

Valeurs Critiques et Niveau de Signification

La **valeur critique** D_α dépend de deux facteurs :

- Le **niveau de signification** α (souvent $\alpha = 0.05$ ou $\alpha = 0.01$).
- La **taille de l'échantillon** n.

Les valeurs critiques de D pour un test à un échantillon sont données par des tables spécifiques du test de Kolmogorov-Smirnov, ou peuvent être calculées à l'aide de la formule :

$$D_\alpha = \sqrt{-\ln(\alpha/2)/(2n)}$$

où α est le niveau de signification et n la taille de l'échantillon.

Procédure de Décision

La procédure de décision du test de Kolmogorov-Smirnov est la suivante :

1. **Calculer la statistique** D pour l'échantillon ou les deux échantillons.
2. **Trouver la valeur critique** D_α à partir des tables en fonction de la taille de l'échantillon n et du niveau de signification α.
3. **Comparer D à D_α :**
   - Si D ¿ D_α, rejetez l'hypothèse nulle H_0 (les données ne suivent pas la distribution théorique).
   - Si $D \leq D_\alpha$, ne rejetez pas l'hypothèse nulle H_0.

Cela permet de prendre la décision statistique de rejeter ou de ne pas rejeter l'hypothèse nulle.

Exemple de Test à un Échantillon

Supposons que nous avons un échantillon de taille $n = 50$ et que nous souhaitons tester si cet échantillon suit une distribution normale avec une moyenne de 0 et un écart-type de 1.

- **Hypothèse nulle (H)** : L'échantillon suit une distribution normale standard.
- La **statistique du test** D est calculée en comparant la fonction de répartition empirique $F_n(x)$ à la fonction de répartition théorique $F(x)$ de la loi normale.
- En utilisant une table de Kolmogorov-Smirnov, nous trouvons que pour $n = 50$ et $\alpha = 0.05$, $D_\alpha \approx 0.223$.

Si la statistique D est supérieure à 0.223, nous rejetons H_0. Sinon, nous ne rejetons pas H_0.

Exemple de Test pour Deux Échantillons

Supposons que nous avons deux échantillons A et B, et que nous souhaitons tester si ces deux échantillons proviennent de la même distribution.

- **Hypothèse nulle (H)** : Les deux échantillons proviennent de la même distribution.
- La **statistique du test** D est calculée en comparant les fonctions de répartition empiriques $F_{n_1}(x)$ et $F_{n_2}(x)$ des deux échantillons.
- Si D est supérieur à la valeur critique D_α, nous rejetons H_0.

Valeurs Critiques pour Test à Deux Échantillons

La valeur critique D_α pour le test à deux échantillons dépend également de n_1 et n_2 (les tailles des deux échantillons).

- Une fois les tailles des échantillons connues, la valeur critique peut être obtenue à partir des tables de Kolmogorov-Smirnov ou calculée à l'aide de la formule appropriée.
- Le niveau de signification α influence également la valeur critique.

Conclusion

Le test de Kolmogorov-Smirnov est un test non paramétrique puissant pour comparer une fonction de répartition empirique à une fonction de répartition théorique ou pour comparer deux échantillons.

- La **région critique** est l'ensemble des valeurs de la statistique D pour lesquelles l'hypothèse nulle est rejetée.
- La **valeur critique** dépend de la taille de l'échantillon n et du niveau de signification α.
- La procédure de décision repose sur la comparaison de la statistique D avec la valeur critique.

Le test de Kolmogorov-Smirnov est largement utilisé pour tester la normalité des données ou comparer des distributions empiriques.

Tests statistiques utilisant la fonction de répartition empirique

RAHMANI Naceur

Introduction

La fonction de répartition empirique ($F_n(x)$) est une estimation de la fonction de répartition d'une population à partir d'un échantillon de données.

- Ces tests non paramétriques utilisent $F_n(x)$ pour comparer des distributions ou tester des hypothèses sur les données.
- Ils sont utiles pour tester si un échantillon suit une distribution théorique ou pour comparer deux échantillons.

Les tests basés sur la fonction de répartition empirique incluent :

- Test de Cramer-von Mises
- Test d'Anderson-Darling
- Test de Kuiper
- Test de Watson

Test de Cramer-von Mises

Le test de Cramer-von Mises est similaire au test de Kolmogorov-Smirnov, mais il repose sur une somme pondérée des différences entre les fonctions de répartition empirique et théorique.

- La statistique du test est :

$$W^2 = \sum_{i=1}^{n} [F_n(x_i) - F(x_i)]^2$$

où $F_n(x_i)$ et $F(x_i)$ sont respectivement les fonctions de répartition empirique et théorique.

- Ce test mesure la "magnitude" des écarts entre les deux distributions sur toute la gamme des données.

Décision :

- Si W^2 dépasse une valeur seuil, l'hypothèse nulle est rejetée.

Test d'Anderson-Darling

Le test d'Anderson-Darling est une variante du test de Cramer-von Mises, mais il met plus de poids sur les différences dans les queues de la distribution.

- La statistique du test est donnée par :

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i-1) \left(\ln F(x_i) + \ln(1 - F(x_{n-i+1})) \right) \right]$$

où $F(x)$ est la fonction de répartition théorique et $F_n(x)$ est la fonction empirique.

- Ce test est plus puissant pour détecter des écarts dans les queues de la distribution.

Décision :

- Si A^2 est supérieur à la valeur critique, l'hypothèse nulle est rejetée.

Test de Kuiper

Le test de Kuiper est une variante du test de Kolmogorov-Smirnov qui est symétrique et mesure les écarts dans les deux directions (positives et négatives).

- La statistique de test est :

$$V = \sup_x |F_n(x) - F(x)| + \sup_x |F(x) - F_n(x)|$$

où $F_n(x)$ et $F(x)$ sont respectivement la fonction de répartition empirique et théorique.

Ce test est particulièrement utile lorsque l'on veut tester des écarts dans les deux directions de la fonction de répartition.

Décision :

- Si V dépasse la valeur seuil, l'hypothèse nulle est rejetée.

Test de Watson

Le test de Watson est basé sur la fonction de répartition empirique, mais il est spécifiquement utilisé pour tester la normalité des données.

- La statistique du test est calculée en utilisant les différences pondérées entre $F_n(x)$ et $F(x)$, en appliquant des poids dépendant de la forme de la distribution théorique.
- Ce test est souvent plus puissant que le test de Kolmogorov-Smirnov lorsqu'il s'agit de tester la normalité des données.

Décision :

- Si la statistique dépasse la valeur critique, l'hypothèse nulle (normalité) est rejetée.

Conclusion

Les tests utilisant la fonction de répartition empirique sont des outils puissants pour tester des hypothèses sur les distributions des données. Ces tests sont non paramétriques et ne nécessitent aucune hypothèse préalable sur la forme de la distribution sous-jacente des données.

- Le test de Kolmogorov-Smirnov est le plus connu, mais d'autres tests comme le test de Cramer-von Mises, Anderson-Darling, Kuiper et Watson peuvent être plus adaptés selon le contexte.
- Ces tests sont particulièrement utiles pour tester la normalité des données ou pour comparer des échantillons.

Applications : Analyse des résidus, test de normalité, comparaison de distributions empiriques et théoriques.