STATISTIQUE APPLIQUÉE & CONTRÔLE DE QUALITÉ

Le Test d'Adéquation de Kolmogorov-Smirnov (K-S)

Principe, formulation mathématique, éléments apportant la confiance & exemple d'application complet

1. Introduction et Objectif

Dans le cadre de l'analyse statistique et de la maîtrise des procédés, valider si un échantillon de mesures provient d'une distribution de référence spécifique (normale, exponentielle, Weibull, etc.) est une étape critique préalable à toute inférence. Le test d'adéquation de Kolmogorov-Smirnov (K-S) est l'un des outils non paramétriques les plus puissants et répandus pour vérifier la conformité d'une distribution continue empirique avec un modèle théorique spécifié.

2. Définitions et Formulation Mathématique

2.1 Principe général

Soit un échantillon de n observations indépendantes et identiquement distribuées (i.i.d.) noté $X_1, X_2, ..., X_n$, réordonné sous forme de statistiques d'ordre croissantes :

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$$

Le test vise à départager deux hypothèses fondamentales :

- **Hypothèse nulle (H_0) :** Les données observées proviennent de la loi théorique $F_0(x)$.
- **Hypothèse alternative (H_1) :** La distribution réelle $F(x)$ diffère de la loi $F_0(x)$.

2.2 Fonction de répartition empirique et Statistique Dn

La fonction de répartition empirique $F_n(x)$ représente la proportion des observations inférieures ou égales à x :

> **■ FORMULATION MATHÉMATIQUE DE LA STATISTIQUE K-S**
>
> $F_n(x) = (1 / n) \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$
>
> Statistique de Kolmogorov-Smirnov (distance sup) :
>
> $D_n = \sup_{x \in \mathbb{R}} \{ | F_n(x) - F_0(x) | \}$
>
> Calcul algorithmique pratique aux points d'observation $X_{(i)}$:
>
> $D_n = \max_{1 \leq i \leq n} \{ \max(|(i / n) - F_0(X_{(i)})|, |((i - 1) / n) - F_0(X_{(i)})|) \}$

La statistique D_n mesure le plus grand écart vertical absolu entre la courbe en escalier des fréquences observées et la courbe continue de la fonction théorique candidate.

3. Éléments Apportant de la Confiance dans le Test

Plusieurs théorèmes et propriétés fondamentales confèrent au test de Kolmogorov-Smirnov une robustesse et une rigueur mathématique indispensables pour les analyses d'ingénierie :

Propriété Fondamentale	Enoncé / Théorème	Rôle dans la Confiance
Théorème de Glivenko-Cantelli	$\sup_{x \in \mathbb{R}} \lvert F_n(x) - F(x) \rvert \xrightarrow{p.s.} 0$ quand $n \to \infty$	Garantit la convergence presque sûre de la fonction empirique vers la réalité. À grand échantillon, l'erreur d'échantillonnage s'annule.
Distribution Distribution-Free (Loi Libre)	La loi de D_n sous H_0 ne dépend aucunement de F_0 si celle-ci est continue.	Permet de tabuler des valeurs critiques universelles $d_{n, \alpha}$ valables pour toute loi continue sans recalculer la distribution sous-jacente.
Normalisation Asymptotique de Kolmogorov	$P(\sqrt{n} D_n \leq z) \longrightarrow K(z) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2 k^2 z^2}$	Fournit une approximation analytique extrêmement précise des p-values dès que $n > 35$, sans dépendre de simulations de Monte-Carlo lourdes.
Non-paramétrique sans découpage	Travaille directement sur les observations individuelles continues.	Élimine le biais subjectif de partitionnement en classes (contrairement au test du Chi-deux qui dépend du nombre et de la largeur des intervalles).

3.1 Règle de décision et Seuil de signification

Pour un seuil de risque α fixé (usuellement 5% ou 1%) et une taille d'échantillon n :

> **■ RÈGLE DE DÉCISION STATISTIQUE**
>
> Valeur critique asymptotique (pour $n > 35$) au seuil $\alpha = 5\%$: $d_{n, 0.05} \approx 1.36 / \sqrt{n}$
>
> Valeur critique asymptotique (pour $n > 35$) au seuil $\alpha = 1\%$: $d_{n, 0.01} \approx 1.63 / \sqrt{n}$
>
> Règle de décision :
>
> • Si $D_n \leq d_{n, \alpha}$ (ou p-value $\geq \alpha$) : Non-rejet de H_0. Les données sont compatibles avec la loi F_0.
>
> • Si $D_n > d_{n, \alpha}$ (ou p-value $< \alpha$) : Rejet de H_0. L'adéquation au modèle est rejetée avec un risque d'erreur de première espèce α.

3.2 Point de vigilance : Paramètres estimés (Test de Lilliefors)

Une précaution méthodologique majeure s'impose : lorsque les paramètres de la loi théorique ne sont pas fixés a priori mais sont **estimés à partir du même échantillon** (par exemple μ et σ estimés par la moyenne et l'écart-type de l'échantillon), la fonction théorique $F_0(x ; \hat{\theta})$ se 'rapproche' artificiellement des données observées. Le test standard devient alors conservateur (il rejette moins facilement H_0). On recourt dans ce cas à des tables adaptées,

notamment la correction de **Lilliefors**, où la valeur critique à 5% passe typiquement de 1.36 / √n à 0.886 / √n pour une loi normale.

4. Exemple Pratique : Test d'Adéquation à une Loi Exponentielle

On étudie les temps de défaillance (durée de vie en milliers d'heures) de n = 5 composants mécaniques soumis à un test d'endurance. Les mesures relevées sont les suivantes :

Échantillon ordonné $X_{(i)}$: [0.25, 0.6, 1.1, 1.8, 2.75]

On souhaite tester au seuil $\alpha = 0.05$ si ces défaillances suivent une loi exponentielle théorique de paramètre spécifié $\lambda_0 = 0.8\ h^{-1}$, dont la fonction de répartition est $F_0(x) = 1 - e^{\{-0.8\ x\}}$ pour $x \geq 0$.

Rang i	$X_{(i)}$	$F_0(X_{(i)})$	$(i - 1)/n$	i/n	$D_i^- = \lvert F_0 - (i-1)/n \rvert$	$D_i^+ = \lvert (i/n) - F_0 \rvert$
1	0.25	0.1813	0.00	0.20	0.1813	0.0187
2	0.60	0.3812	0.20	0.40	0.1812	0.0188
3	1.10	0.5852	0.40	0.60	0.1852	0.0148
4	1.80	0.7631	0.60	0.80	0.1631	0.0369
5	2.75	0.8892	0.80	1.00	0.0892	0.1108

4.1 Calcul de la statistique observée et conclusion

1. Statistique calculée D_5 : Le maximum des écarts absolus sur l'ensemble du tableau donne :

> **■ RÉSULTAT DE LA STATISTIQUE OBSERVÉE**
>
> $D_5 = \max_{1 \leq i \leq 5} \{ D_i^-, D_i^+ \} = \max(0.1813, 0.0187, 0.1805, 0.0195, 0.1866, 0.0134, 0.1633, 0.0367, 0.0899, 0.1101)$
>
> $\Rightarrow D_5 = 0.1866$

2. Comparaison à la valeur critique : D'après la table exacte de Kolmogorov-Smirnov pour n = 5 au seuil de risque $\alpha = 0.05$, la valeur critique est :

$$d_{5,\ 0.05} = 0.565$$

Conclusion statistique : Comme $D_5 = 0.1866 \leq 0.565$, nous ne rejetons pas l'hypothèse H_0 au seuil de 5%. L'écart maximal entre la distribution empirique des composants et la loi théorique exponentielle $\mathcal{E}(\lambda=0.8)$ est tout à fait compatible avec les fluctuations aléatoires d'échantillonnage.

5. Synthèse des Forces et Limites du Test K-S

Avantages Décisifs	Limites & Points d'Attention

✓ Exact même pour les petits échantillons (n < 30), avec des tables statistiques tabulées précises.	⚠ Réservé exclusivement aux variables aléatoires à distribution continue (inapplicable directement sur des variables discrètes).
✓ Non-paramétrique : ne nécessite aucune hypothèse de départ sur la normalité globale des données.	⚠ Plus sensible au centre de la distribution qu'aux extrémités (queues de distribution) en raison de la distance suprémum uniforme.
✓ Pas de perte d'information liée au regroupement en classes d'intervalles (contrairement au Chi-deux).	⚠ Nécessite l'ajustement de Lilliefors si les paramètres du modèle sont estimés sur l'échantillon.