**Quel méthode d'échantillonnage utiliser ?**

En fonction des exigences détaillées de votre projet, je recommande d'utiliser **l'échantillonnage aléatoire stratifié** pour les raisons suivantes :

1. Vous avez trois catégories de produits distinctes (pantalons, robes, pulls) qui doivent être analysées séparément et comparées.
2. Vos objectifs incluent la compréhension des performances par catégorie et la prise de décisions stratégiques.
3. Vous prévoyez de réaliser des tests d'hypothèses entre les catégories ultérieurement.
4. Les données serviront à des prédictions de ventes, donc préserver la représentativité est crucial.

**Taille de l'échantillon : tout ou un pourcentage ?**

Étant donné que vos ensembles de données sont très petits (37, 38, 37 entrées), je recommande fortement d'utiliser **l'ensemble des données** plutôt qu'un échantillon. Voici pourquoi :

1. **Considérations pour petits ensembles de données** :
   * Avec seulement ~37 entrées par catégorie, un échantillon réduit diminuerait significativement la puissance statistique.
   * Vous risqueriez de perdre des motifs ou des variations importants dans les données.
   * Pour l'apprentissage automatique et l'analyse statistique (que vous envisagez), avoir plus de points de données fournira des résultats plus fiables.

**Expliquer le Z-score dans ce contexte**

Le Z-score sert à détecter les valeurs aberrantes dans l'analyse de vos données de ventes.  
Il mesure à combien d'écarts-types un point de données se situe par rapport à la moyenne, c'est-à-dire à quel point une valeur est "inhabituelle" dans votre jeu de données.

La formule du Z-score est :

**Z = (X - μ) / σ**

Où :

* **X** = valeur analysée
* **μ** = moyenne de l'ensemble de données
* **σ** = écart-type de l'ensemble de données

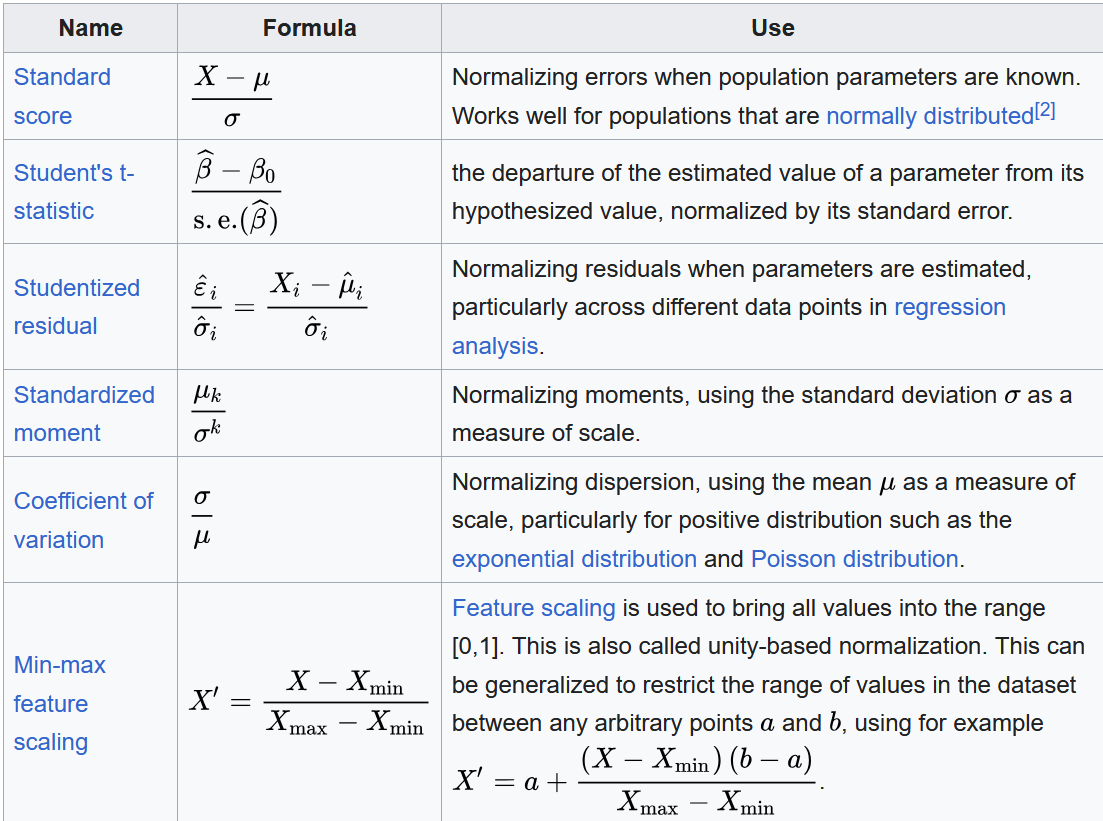
**Comment choisir le seuil (seuil) ?**

Le seuil du Z-score est généralement choisi en fonction de plusieurs critères :

1. **Théorie statistique** :
   * **Z > 3** : Capture environ 0,3 % des valeurs les plus extrêmes (règle des 99,7 %).
   * **Z > 2** : Capture environ 5 % des valeurs les plus extrêmes (règle des 95 %).
   * **Z > 1** : Capture environ 32 % des valeurs les plus extrêmes (règle des 68 %).
2. **Contexte métier** : Pour vos données e-commerce :
   * **Ventes (S)** : Un seuil élevé (3-3,5) pourrait être pertinent, car les ventes varient souvent naturellement.
   * **Pages vues (X4)** : Un seuil standard (3) peut suffire, car les patterns de trafic sont généralement prévisibles.
   * **Indices clients/transactions (X9, X10)** : Un seuil plus bas (2,5) peut être utilisé si une certaine stabilité est attendue.
3. **Taille de l'échantillon** :
   * **Petits jeux de données (<100, comme le vôtre)** : Envisagez des seuils plus bas (2,5-2,7).
   * **Grands jeux de données (>1000)** : Utilisez des seuils standards plus élevés (3-3,5).
   * **Très grands jeux de données (>10000)** : Vous pouvez envisager des seuils encore plus élevés (4+).

**Mise à l’échelle des caractéristiques : pourquoi ?**

* Les variables de votre dataset (par exemple, ventes, pages vues, indices de transaction) sont exprimées dans des unités différentes : milliards de Yuans, millions, centaines de milliers, etc.
* Ces différences d'échelle peuvent **biaiser les résultats des analyses** (corrélations, ACP, etc.) et **affecter la performance des algorithmes d'apprentissage automatique** (régressions, réseaux neuronaux, etc.).
* Exemple : une variable mesurée en milliards aura une influence disproportionnée par rapport à une autre mesurée en centaines de milliers.



# Hypothèse Statistique et Test de Normalité

Si vous devez effectuer un test d’hypothèse mais que vous ne savez pas si votre échantillon suit une loi normale, voici les étapes à suivre :

## 1. Vérifier la Normalité de Vos Données

Pour déterminer si vos données suivent une distribution normale, utilisez :

* **Méthodes Visuelles**
  + Graphiques comme les Q-Q plots ou les histogrammes
* **Tests Statistiques**
  + Test de Shapiro-Wilk (adapté aux petits échantillons)
  + Test de Kolmogorov-Smirnov

## 2. Que Faire si Vos Données Ne Sont Pas Normales ?

Deux options s'offrent à vous :

1. Utiliser des **tests non paramétriques** qui ne supposent pas la normalité.
2. Appliquer le **Théorème Central Limite (TCL)** si votre échantillon est suffisamment grand (généralement n > 30).

## Échantillon Supérieur à 30 : Utilisation du TCL

Si votre échantillon contient plus de 30 observations, vous pouvez vous fier au **Théorème Central Limite (TCL)** :

* **Pourquoi ?**
  + Le TCL indique que, quelle que soit la distribution initiale de vos données, la distribution de l’échantillon moyen sera approximativement normale pour un grand échantillon.
* **Conséquence :**
  + Vous pouvez procéder à des tests paramétriques (comme les tests t ou z), même si vos données d'origine ne sont pas parfaitement normales.

## 3. Tester une Hypothèse Spécifique

**Comparaison des Moyennes de Deux Groupes** :

Si vous comparez les moyennes de deux groupes indépendants, suivez ces étapes :

1. **Calculer les Moyennes et Écarts-Types** des deux groupes.
2. **Tester l’Égalité des Variances** :
   * Utilisez le test de Levene pour vérifier si les variances sont égales. Cela déterminera si vous devez utiliser des variances combinées ou non.
3. **Effectuer un Test t** (ou test de Student pour deux échantillons indépendants) :
   * Basé sur :
     + La différence entre les moyennes des échantillons
     + L’erreur standard de la différence entre moyennes
     + Les degrés de liberté appropriés

