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 :

$$\mathbf{Z = (X - \mu) / \sigma}$$

Où :

- $\mathbf{X}$ = valeur analysée
- $\boldsymbol{\mu}$ = moyenne de l'ensemble de données
- $\boldsymbol{\sigma}$ = 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.

Name	Formula	Use
Standard score	$\frac{X - \mu}{\sigma}$	Normalizing errors when population parameters are known. Works well for populations that are normally distributed[2]
Student's t-statistic	$\frac{\widehat{\beta} - \beta_0}{\mathrm{s.e.}(\widehat{\beta})}$	the departure of the estimated value of a parameter from its hypothesized value, normalized by its standard error.
Studentized residual	$\frac{\hat{\varepsilon}_i}{\hat{\sigma}_i} = \frac{X_i - \hat{\mu}_i}{\hat{\sigma}_i}$	Normalizing residuals when parameters are estimated, particularly across different data points in regression analysis.
Standardized moment	$\frac{\mu_k}{\sigma^k}$	Normalizing moments, using the standard deviation σ as a measure of scale.
Coefficient of variation	$\frac{\sigma}{\mu}$	Normalizing dispersion, using the mean μ as a measure of scale, particularly for positive distribution such as the exponential distribution and Poisson distribution.
Min-max feature scaling	$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$	Feature scaling is used to bring all values into the range [0,1]. This is also called unity-based normalization. This can be generalized to restrict the range of values in the dataset between any arbitrary points a and b, using for example $X' = a + \frac{(X - X_{\min})(b - a)}{X_{\max} - X_{\min}}$.

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

Hypothèse à tester(test bilatéral):

$$H_0 : "\mu \text{ et } \mu' \text{ ne sont pas significativement différentes}"$$

$$H_1 : "\mu \text{ et } \mu' \text{ sont significativement différentes}"$$

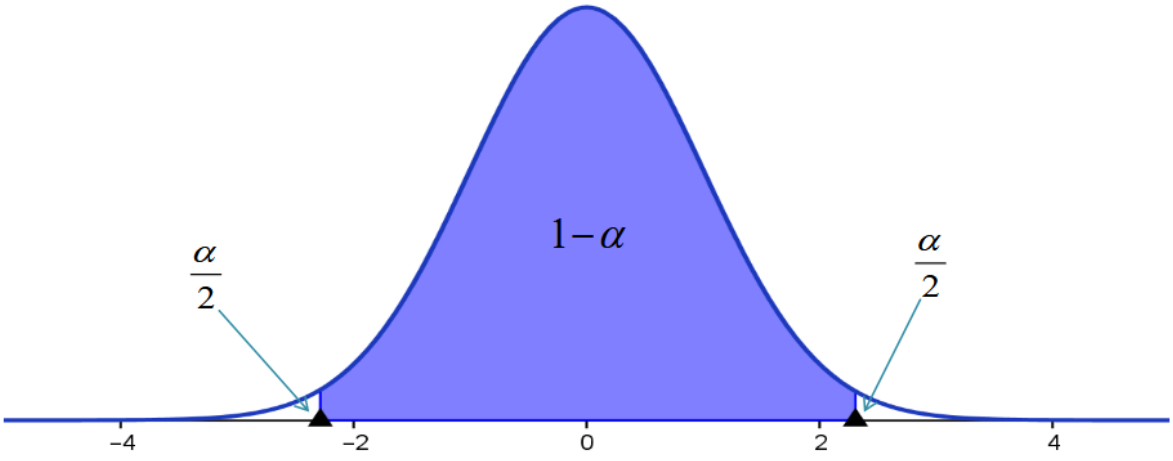
$$\text{alors, sous } H_0 \text{ la v.a } T = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma'^2}{n_B}}} \rightarrow N(0,1)$$

Fixons un seuil de risque α et soit l'unique réel strictement positif t_α telque $P(|T| \leq t_\alpha) = 1 - \alpha$

$$(t_\alpha = \Pi^{-1}(1 - \frac{\alpha}{2}))$$

Règle de décision du test:

- *Si* $|T| \leq t_\alpha$, *on ne rejete pas* H_0, *avec un risque* β *de se tromper.*
- *Si* $|T| \succ t_\alpha$, *on rejete* H_0, *avec le risque* α *de se tromper.*



Mise en œuvre du test:

$$On\ calcule\ le\ nombre\ t = \frac{m_A - m_B}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma'^2}{n_B}}}\ et\ on\ compare\ |t|\ à\ t_\alpha\ ;$$

$$et\ on\ utilise\ la\ régle\ de\ décision\ pour\ conclure.$$

$$En\ général\ \sigma\ et\ \sigma'\ sont\ inconnus\ et\ remplacés\ par\ leurs\ estimateurs$$

$$respectifs\ \hat{\sigma}_A = \sigma_A \sqrt{\frac{n_A}{n_A - 1}}\ \ et\ \ \hat{\sigma}_B = \sigma_B \sqrt{\frac{n_B}{n_B - 1}}$$

Bien sûr ! Voici l'explication en français :

---

1. Interpréter l'Erreur Quadratique Moyenne (MSE)

- **L'Erreur Quadratique Moyenne (MSE)** mesure la différence moyenne au carré entre les valeurs prédites et réelles. Plus la MSE est faible, meilleure est la performance du modèle.
- Si 0.14370.1437 est une "bonne" erreur dépend du **scale** (échelle) de votre variable cible yy :
  - Si les valeurs de yy varient entre 0 et 1, un MSE de 0.1437 est relativement élevé.
  - Si yy varie entre 0 et 100, cet MSE pourrait être acceptable.

Action : Vérifiez la plage des valeurs de yy pour savoir si cette erreur est tolérable dans ce contexte.

---

2. Interpréter le R2R^2

- **Le R2R^2** (coefficient de détermination) mesure dans quelle mesure le modèle explique la variance de la variable cible :
  - R2=1.0R^2 = 1.0 : Ajustement parfait.
  - R2>0.5R^2 > 0.5 : Indique généralement un bon modèle.
  - R2=0R^2 = 0 : Le modèle ne prédit pas mieux que la moyenne.
  - R2<0R^2 < 0 : Le modèle est pire qu'une simple prédiction par la moyenne.

Votre Résultat : R2=0.16R^2 = 0.16 :

- Cela signifie que seulement **16 % de la variance de yy** est expliquée par le modèle.
- C'est assez faible dans la plupart des applications, ce qui suggère que le modèle ne capture pas bien la relation.

---

3. Pourquoi le modèle pourrait sous-performer ?

1. **Ordre du modèle incorrect (p,d,qp, d, q) :**
   - L'ordre ARIMAX choisi `(1, 0, 0)` peut ne pas être adapté à vos données.
   - Utilisez des outils comme les graphiques ACF/PACF pour guider la sélection de pp et qq, et effectuez des tests de stationnarité pour déterminer dd.
2. **Variables Exogènes :**
   - Si les variables exogènes n'ont pas une forte relation avec yy, elles n'aideront pas beaucoup.
   - Évaluez la corrélation entre xx et yy. Envisagez d'ajouter d'autres variables exogènes ou de les supprimer totalement.
3. **Stationnarité :**
   - Si vos données ne sont pas stationnaires (moyenne et variance constantes dans le temps), le modèle pourrait avoir du mal à prédire correctement.
   - La différenciation (en définissant d>0d > 0) pourrait aider.
4. **Relations complexes :**
   - ARIMAX suppose des relations linéaires. Si la relation entre xx et yy ou entre yty_t et yt−1y_{t-1} est non-linéaire, ARIMAX peut sous-performer.
   - Envisagez d'autres modèles comme LSTM ou la régression non linéaire si vous suspectez des relations non-linéaires.
5. **Bruit dans les données :**
   - Si les données contiennent beaucoup de bruit ou ont un faible rapport signal/bruit, la prédiction devient plus difficile.

---

4. Étapes Suivantes pour Améliorer le Modèle

- **Vérifiez la stationnarité des données :**

```
from statsmodels.tsa.stattools import adfuller

result = adfuller(df['y'])
print(f"Statistique ADF : {result[0]}")
print(f"P-value : {result[1]}")
```

  - Si la p-value > 0.05, vos données ne sont pas stationnaires. Envisagez de les différencier.

- **Utilisez les graphiques ACF/PACF pour ajuster pp et qq :**

```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

plot_acf(df['y'])
```

- `plot_pacf(df['y'])`
- `plt.show()`
- **Essayez différents ordres ARIMAX :** Testez des combinaisons comme `(1, 1, 1)`, `(2, 1, 0)`, etc.
- **Génération de nouvelles caractéristiques :**
    - Si xx n'est pas fortement corrélé avec yy, essayez d'ajouter d'autres variables pertinentes.
    - Exemple : utilisez des caractéristiques basées sur le temps comme le jour de la semaine, la saison, etc.

---

5. Autres Approches

- Si ARIMAX ne s'améliore pas, explorez d'autres modèles de séries temporelles :
    - **Prophet** (idéal pour les tendances et la saisonnalité).
    - **Modèles de Machine Learning** (comme les forêts aléatoires ou LSTM pour les séries temporelles).

En résumé :

- Votre R2R^2 et MSE suggèrent que le modèle actuel n'est pas optimal.
- Affinez le modèle en vérifiant la stationnarité, en ajustant p,d,qp, d, q, et en évaluant les variables exogènes.
- Faites-moi savoir si vous avez besoin d'aide avec ces étapes !