Test: Types de Variables en Analyse de Données

Niveau: Intermédiaire | Durée: 30 minutes | 25 Questions

Section A: Identification des Types (10 questions)

Question 1

Identifiez le type de chaque variable dans le contexte bancaire:

Variable	Type à identifier
Numéro de compte client	?
Score de satisfaction (1-5)	?
Montant du prêt	?
Type de compte (Épargne/Courant/DAT)	?
Nombre de transactions mensuelles	?

Voir la réponse

Variable	Type
Numéro de compte client	**Nominale** (identifiant, pas numérique!)
Score de satisfaction (1-5)	**Ordinale** (ordre naturel)
Montant du prêt	**Continue (Ratio)**
Type de compte	**Nominale polytomique**
Nombre de transactions	**Discrète** (entiers)

Question 2

Quel type de variable est le “Rating de crédit” (AAA, AA, A, BBB, BB, B)?

A) Nominale

B) Ordinale

C) Discrète

D) Continue

Voir la réponse

B) Ordinale

Le rating de crédit a un ordre naturel (AAA > AA > A > BBB > BB > B), mais les intervalles entre les ratings ne sont pas nécessairement égaux. C’est donc une variable ordinale.

Question 3

Pourquoi ne doit-on JAMAIS traiter un numéro de compte comme une variable numérique?

Voir la réponse

Parce que: 1. Les opérations arithmétiques n'ont pas de sens (compte 1234 + compte 5678 = ?) 2. La moyenne des numéros de compte n'a aucune signification 3. C'est un **identifiant unique**, pas une mesure 4. Doit être traité comme une chaîne de caractères (string)

Le numéro de compte est une variable **nominale** servant uniquement à identifier de manière unique chaque compte.

Question 4

Classez ces variables dans l'ordre croissant de leur niveau de mesure:

- Revenu mensuel (en HTG)
- Région géographique
- Niveau d'éducation (Primaire < Secondaire < Universitaire)
- Température (en °C)

Voir la réponse

Ordre croissant (du moins informatif au plus informatif):

1. **Région géographique** → Nominale (= ≠ seulement)
2. **Niveau d'éducation** → Ordinale (= ≠ < >)
3. **Température** → Intervalle (+ - , zéro arbitraire)
4. **Revenu mensuel** → Ratio (× ÷, zéro absolu = pas de revenu)

Question 5

Quelle est la différence entre une variable discrète et une variable continue?

Voir la réponse

Variable Discrète: - Valeurs dénombrables (on peut les compter) - Généralement des entiers - Exemples: nombre de transactions, nombre de produits, nombre de jours de retard

Variable Continue: - Peut prendre n'importe quelle valeur dans un intervalle - Valeurs décimales possibles - Exemples: montant en HTG, taux d'intérêt, durée en secondes

Règle simple: Si on peut avoir 2.5 de quelque chose, c'est continu. Si 2.5 n'a pas de sens (2.5 transactions?), c'est discret.

Question 6

Le code postal (ex: HT6110) est quel type de variable?

A) Quantitative discrète

B) Nominale

C) Ordinale

D) Continue

Voir la réponse

B) Nominale

Même si un code postal contient des chiffres, c'est une variable nominale car: - Il sert à identifier une zone géographique - Les opérations arithmétiques n'ont pas de sens - Il n'y a pas d'ordre naturel (HT6110 n'est pas "supérieur" à HT6100)

Question 7

Pour chaque situation, indiquez si la variable est Stock ou Flux:

Mesure	Stock ou Flux?
Solde du compte au 31/12	?
Nombre de transactions du mois	?
Encours total de prêts	?
Volume de dépôts du trimestre	?

Voir la réponse

Mesure	Type
Solde du compte au 31/12	**Stock** (à un instant T)
Nombre de transactions du mois	**Flux** (sur une période)
Encours total de prêts	**Stock**
Volume de dépôts du trimestre	**Flux**

Question 8

Quelle mesure de tendance centrale utiliser pour chaque type de variable?

Type de variable	Mesure appropriée
Nominale	?
Ordinale	?
Quantitative avec outliers	?
Quantitative symétrique	?

Voir la réponse

Type de variable	Mesure appropriée
Nominale	**Mode** (seule option possible)
Ordinale	**Médiane** (préserve l'ordre)
Quantitative avec outliers	**Médiane** (robuste)

Type de variable	Mesure appropriée
Quantitative symétrique	**Moyenne** (utilise toutes les valeurs)

Question 9

Qu’est-ce qu’une variable binaire? Donnez 3 exemples bancaires.

Voir la réponse

Une **variable binaire** (ou dichotomique) est une variable nominale qui ne peut prendre que 2 valeurs mutuellement exclusives.

Exemples bancaires: 1. **Défaut de paiement:** Oui / Non 2. **Client actif:** Actif / Inactif 3. **Propriétaire:** Oui / Non 4. **Transaction frauduleuse:** Fraude / Légitime 5. **Compte joint:** Oui / Non

---

Question 10

Le taux d’intérêt (ex: 8.5%) est quel type de variable?

A) Ordinale

B) Discrète

C) Continue (Intervalle)

D) Continue (Ratio)

Voir la réponse

D) Continue (Ratio)

Le taux d’intérêt: - Peut prendre des valeurs décimales (8.5%, 8.75%) - A un zéro absolu (0% = pas d’intérêt) - Les ratios ont du sens (10% est le double de 5%)

Section B: Encodage et Transformation (8 questions)

Question 11

Quel encodage utiliser pour la variable “Région” (Nord, Sud, Est, Ouest)?

A) Label Encoding

B) One-Hot Encoding

C) Ordinal Encoding

D) Aucun encodage nécessaire

Voir la réponse

B) One-Hot Encoding

“Région” est une variable nominale sans ordre naturel. One-Hot Encoding crée une colonne binaire par catégorie:

```
Region_Nord  Region_Sud  Region_Est  Region_Ouest
    1            0           0            0
    0            1           0            0
```

Label Encoding (0,1,2,3) impliquerait un ordre inexistant et biaiserait les modèles.

Question 12

Pour le “Niveau de risque” (Faible, Modéré, Élevé, Critique), quel encodage?

Voir la réponse

Ordinal Encoding avec ordre explicite:

```
ordre = {'Faible': 1, 'Modéré': 2, 'Élevé': 3, 'Critique': 4}
df['risque_encoded'] = df['niveau_risque'].map(ordre)
```

Ou avec sklearn:

```
from sklearn.preprocessing import OrdinalEncoder
categories = [['Faible', 'Modéré', 'Élevé', 'Critique']]
encoder = OrdinalEncoder(categories=categories)
```

L’ordre est préservé car c’est une variable ordinale.

Question 13

Pourquoi faut-il “fit” le StandardScaler uniquement sur les données d’entraînement?

Voir la réponse

Pour éviter le **Data Leakage** (fuite de données):

1. Si on fit sur toutes les données, les statistiques (moyenne, écart-type) incluent des informations du test set
2. Le modèle “voit” indirectement les données de test pendant l’entraînement
3. Les performances sont artificiellement gonflées
4. Le modèle ne généralisera pas bien en production

Bonne pratique:

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)  # FIT + TRANSFORM
X_test_scaled = scaler.transform(X_test)        # TRANSFORM seulement
```

Question 14

Quel encodage pour ces variables?

Variable	Encodage
Genre (M/F)	?
Satisfaction (1-5 étoiles)	?

Variable	Encodage
Pays (Haïti, USA, France…)	?
Score de crédit (300-850)	?

Voir la réponse

Variable	Encodage
Genre (M/F)	**Label (0/1)** ou One-Hot (2 catégories = binaire)
Satisfaction (1-5)	**Ordinal** ou garder tel quel (déjà numérique ordonné)
Pays	**One-Hot** (nominale, pas d'ordre)
Score de crédit	**Aucun** - déjà numérique continu

Question 15

Quand utiliser pd.cut() vs pd.qcut()?

Voir la réponse

pd.cut() - Intervalles de taille ÉGALE:

```
# Tranches fixes
pd.cut(df['age'], bins=[0, 25, 50, 75, 100])
# [0-25], [25-50], [50-75], [75-100]
```

Usage: Quand les bornes ont une signification métier.

pd.qcut() - Quantiles (effectifs ÉGAUX):

```
# Quartiles (25% dans chaque)
pd.qcut(df['age'], q=4)
```

Usage: Quand on veut des groupes de taille similaire.

Question 16

Comment transformer une variable continue très asymétrique (ex: revenus)?

Voir la réponse

Transformation logarithmique:

```
import numpy as np

# log(1+x) pour gérer les valeurs 0
df['log_revenu'] = np.log1p(df['revenu'])

# Ou Box-Cox (requiert valeurs > 0)
from scipy.stats import boxcox
df['boxcox_revenu'], lambda_param = boxcox(df['revenu'])
```

Pourquoi: - Réduit l'asymétrie - Réduit l'impact des outliers - Améliore les performances des modèles linéaires - Les revenus suivent souvent une loi log-normale

---

Question 17

Quelle normalisation choisir?

Situation	Normalisation
Données avec outliers	?
Données normales	?
Besoin de valeurs [0,1]	?

Voir la réponse

Situation	Normalisation
Données avec outliers	**RobustScaler** (utilise médiane et IQR)
Données normales	**StandardScaler** (z-score)
Besoin de valeurs [0,1]	**MinMaxScaler**

```
from sklearn.preprocessing import StandardScaler, MinMaxScaler, RobustScaler

# Standard: z = (x - μ) / σ
StandardScaler()

# MinMax: x' = (x - min) / (max - min)
MinMaxScaler()

# Robust: x' = (x - median) / IQR
RobustScaler()
```

Question 18

Créez 3 features pertinentes à partir de "date_ouverture_compte":

Voir la réponse

```
# 1. Ancienneté en jours
df['anciennete_jours'] = (pd.Timestamp.now() - df['date_ouverture']).dt.days

# 2. Ancienneté en années
df['anciennete_annees'] = df['anciennete_jours'] / 365

# 3. Mois d'ouverture (saisonnalité)
df['mois_ouverture'] = df['date_ouverture'].dt.month

# 4. Jour de la semaine
df['jour_semaine_ouverture'] = df['date_ouverture'].dt.dayofweek

# 5. Année d'ouverture
```

```
df['annee_ouverture'] = df['date_ouverture'].dt.year

# 6. Est-ce un weekend?
df['ouverture_weekend'] = (df['date_ouverture'].dt.dayofweek >= 5).astype(int)
```

Section C: Implications pour l'Analyse (7 questions)

Question 19

Quel test statistique pour comparer les moyennes de revenus entre clients avec et sans défaut?

A) Chi-carré

B) t-test indépendant

C) ANOVA

D) Corrélation de Pearson

Voir la réponse

B) t-test indépendant

- Variable cible: Défaut (binaire: 2 groupes)
- Variable à comparer: Revenu (continue)
- Objectif: Comparer les moyennes de 2 groupes indépendants

Le t-test indépendant est approprié.

Question 20

Pour analyser la relation entre "Type de compte" et "Défaut de paiement", quel test?

Voir la réponse

Test du Chi-carré (χ^2)

Les deux variables sont catégorielles: - Type de compte: Nominale - Défaut: Binaire (nominale)

Le Chi-carré teste l'indépendance entre deux variables catégorielles.

```
from scipy.stats import chi2_contingency
table = pd.crosstab(df['type_compte'], df['defaut'])
chi2, p_value, dof, expected = chi2_contingency(table)
```

Question 21

Pourquoi la corrélation de Pearson n'est-elle pas appropriée pour une variable ordinale?

Voir la réponse

Pearson mesure la **relation linéaire** et suppose: 1. Variables continues 2. Intervalles égaux entre valeurs 3. Distribution normale

Pour les variables ordinales: - Les intervalles ne sont pas égaux (la distance entre “Satisfait” et “Très satisfait” ≠ “Insatisfait” et “Neutre”) - La relation peut être monotone mais non linéaire

Alternative: Corrélation de **Spearman** (basée sur les rangs)

```
from scipy.stats import spearmanr
correlation, p_value = spearmanr(df['satisfaction'], df['nb_achats'])
```

Question 22

Comment détecter les outliers pour une variable continue?

Voir la réponse

Méthode IQR (Interquartile Range):

```
Q1 = df['col'].quantile(0.25)
Q3 = df['col'].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

outliers = df[(df['col'] < lower_bound) | (df['col'] > upper_bound)]
```

Méthode Z-score:

```
from scipy import stats
z_scores = np.abs(stats.zscore(df['col']))
outliers = df[z_scores > 3]  # Plus de 3 écarts-types
```

Visuelle: Box plot

Question 23

Une variable “Montant” a un skewness de 2.5. Que signifie cela?

Voir la réponse

Interprétation: - Skewness > 0 = **Asymétrie positive** (queue à droite) - |Skewness| > 1 = **Fortement asymétrique** - Skewness = 2.5 = Distribution très asymétrique à droite

Implications: 1. Beaucoup de petites valeurs, quelques très grandes 2. Moyenne > Médiane > Mode 3. La moyenne est tirée vers les grandes valeurs 4. Préférer la **médiane** comme mesure de tendance centrale 5. Envisager une **transformation log** avant modélisation

Typique pour: Revenus, montants de transactions, valeurs de prêts

---

Question 24

Quel graphique pour chaque type de variable?

Type	Graphique approprié
Nominale	?
Ordinale	?
Continue	?
2 continues	?

Voir la réponse

Type	Graphique approprié
Nominale	**Bar chart**, Pie chart (< 6 catégories)
Ordinale	**Bar chart ordonné**, Likert scale
Continue	**Histogramme**, Box plot, Density plot
2 continues	**Scatter plot**, Heatmap corrélation

Question 25

Situation: Vous devez construire un modèle de scoring. Décrivez comment traiter ces variables:

Variable	Traitement
ID_Client	?
Age	?
Profession (50 catégories)	?
Rating actuel (A/B/C/D)	?
Revenus (très asymétriques)	?

Voir la réponse

Variable	Traitement
ID_Client	**Supprimer** - identifiant, aucune valeur prédictive
Age	Garder tel quel ou créer des tranches (binning)
Profession	**Target encoding** ou grouper en catégories (trop de modalités pour one-hot)
Rating actuel	**Ordinal encoding** (A=4, B=3, C=2, D=1)
Revenus	**Log-transform** + StandardScaler

Code:

```
# Supprimer ID
df.drop('ID_Client', axis=1, inplace=True)

# Log transform revenus
df['log_revenu'] = np.log1p(df['revenus'])

# Ordinal encoding rating
df['rating_encoded'] = df['rating'].map({'A': 4, 'B': 3, 'C': 2, 'D': 1})
```

```python
# Target encoding profession
mean_by_prof = df.groupby('profession')['defaut'].mean()
df['profession_encoded'] = df['profession'].map(mean_by_prof)
```

Barème

Section	Points
Section A (Q1-10)	40 points
Section B (Q11-18)	32 points
Section C (Q19-25)	28 points
Total	**100 points**

Seuils: - < 50: À revoir - 50-69: Passable - 70-84: Bien - 85+: Excellent

---

Mnémotechniques à Retenir

NOIR pour les types de variables: - **N**ominale = Noms, catégories sans ordre - **O**rdinale = Ordre naturel - **I**ntervalle = Intervalles égaux, zéro arbitraire - **R**atio = Ratios possibles, zéro absolu

“La Médiane est Ma Meilleure Amie” - Pour les outliers et données asymétriques

“One-Hot pour les Noms, Ordinal pour l’Ordre” - Choix d’encodage