Test Global Data Analyst - Complet (4/4)

Contexte: Entretien complet pour Data Analyst - UniBank Haiti
Durée estimée: 60-90 minutes
Nombre de questions: 50

Section A: Statistiques et Probabilités (12 questions)

Q1. Définissez et calculez: moyenne, médiane, mode, variance, écart-type pour le jeu de données suivant: [10, 15, 20, 20, 25, 30, 100]

R1. - Moyenne: (10+15+20+20+25+30+100)/7 = 220/7 = **31.43** - Médiane: 7 valeurs, position médiane = 4ème → **20** - Mode: Valeur la plus fréquente → **20** - Variance: $\Sigma(x_i-\bar{x})^2/(n-1)$ = **973.81** - Écart-type: √973.81 = **31.21**

Observation: La moyenne (31.43) est très supérieure à la médiane (20) à cause de l'outlier (100). La médiane est plus représentative.

Q2. Qu'est-ce que la p-value? Un test donne p=0.03. Que concluez-vous avec α=0.05?

R2. La p-value est la probabilité d'obtenir un résultat aussi extrême ou plus extrême que celui observé, si H_0 est vraie.

Avec p=0.03 < α=0.05: - **Rejeter H_0** - Le résultat est statistiquement significatif - Moins de 3% de chances d'observer ce résultat si H_0 est vraie

---

Q3. Expliquez la différence entre corrélation et causalité avec un exemple bancaire.

R3. Corrélation: Relation statistique entre deux variables. **Causalité:** Une variable cause directement l'autre.

Exemple bancaire: - Corrélation: "Les clients avec plus de produits ont moins de défauts" (r = -0.4) - Mais: Avoir plus de produits ne CAUSE pas moins de défauts - Variable confondante: Les clients plus stables financièrement ont tendance à prendre plus de produits ET à moins faire défaut

---

Q4. Comment calculer un intervalle de confiance à 95% pour une moyenne?

R4.

```
IC 95% = x̄ ± 1.96 × (s/√n)

Exemple:
x̄ = 1000, s = 200, n = 100
IC = 1000 ± 1.96 × (200/10)
IC = 1000 ± 39.2
IC = [960.8, 1039.2]
```

Q5. Quelle distribution utiliseriez-vous pour modéliser le nombre de clients visitant une agence par heure?

R5. Distribution de Poisson car: - Événements discrets (comptage) - Taux moyen stable (λ clients/heure) - Événements indépendants - Intervalle de temps fixe

$P(X=k) = (\lambda^k \times e^{-\lambda}) / k!$

Q6. Qu'est-ce que le théorème de Bayes? Appliquez-le à la détection de fraude.

R6. $P(A|B) = P(B|A) \times P(A) / P(B)$

Exemple fraude: - P(Fraude) = 1% - P(Alerte|Fraude) = 90% - P(Alerte|Non-fraude) = 5%

$P(Fraude|Alerte) = (0.90 \times 0.01) / (0.90 \times 0.01 + 0.05 \times 0.99) = 0.009 / 0.0585 =$ **15.4%**

Q7-12. [Suite questions statistiques...]

Section B: SQL pour Data Analyst (12 questions)

Q13. Écrivez une requête pour trouver les clients dans le top 10% par solde.

R13.

```sql
WITH percentiles AS (
    SELECT
        client_id,
        solde,
        NTILE(10) OVER (ORDER BY solde) as decile
    FROM comptes
)
SELECT * FROM percentiles WHERE decile = 10;

-- Alternative avec percentile
SELECT *
FROM comptes
WHERE solde >= (SELECT PERCENTILE_CONT(0.9) WITHIN GROUP (ORDER BY solde) FROM comptes);
```

Q14. Calculez le taux de croissance MoM (Month-over-Month) des dépôts.

R14.

```sql
WITH monthly AS (
    SELECT
        DATE_TRUNC('month', date) as mois,
        SUM(solde) as total_depots
    FROM comptes
    GROUP BY DATE_TRUNC('month', date)
)
SELECT
    mois,
    total_depots,
    LAG(total_depots) OVER (ORDER BY mois) as mois_prec,
    ROUND((total_depots - LAG(total_depots) OVER (ORDER BY mois)) * 100.0 /
          NULLIF(LAG(total_depots) OVER (ORDER BY mois), 0), 2) as croissance_pct
```

```sql
FROM monthly
ORDER BY mois;
```

Q15. Identifiez les clients qui ont eu des transactions chaque mois de l'année.

R15.

```sql
SELECT client_id
FROM transactions
WHERE date_tx >= '2024-01-01' AND date_tx < '2025-01-01'
GROUP BY client_id
HAVING COUNT(DISTINCT DATE_TRUNC('month', date_tx)) = 12;
```

Q16. Créez une analyse de cohorte en SQL.

R16.

```sql
WITH cohortes AS (
    SELECT
        client_id,
        DATE_TRUNC('month', date_inscription) as cohorte
    FROM clients
),
activite AS (
    SELECT DISTINCT
        c.client_id,
        c.cohorte,
        DATE_TRUNC('month', t.date_tx) as mois_activite
    FROM cohortes c
    JOIN comptes co ON c.client_id = co.client_id
    JOIN transactions t ON co.compte_id = t.compte_id
)
SELECT
    cohorte,
    DATE_PART('month', AGE(mois_activite, cohorte)) as mois_depuis_inscription,
    COUNT(DISTINCT client_id) as clients_actifs
FROM activite
GROUP BY cohorte, DATE_PART('month', AGE(mois_activite, cohorte))
ORDER BY cohorte, mois_depuis_inscription;
```

Q17-24. [Suite questions SQL...]

Section C: DAX et Power BI (10 questions)

Q25. Créez une mesure pour calculer le NPL ratio.

R25.

```
NPL Ratio =
VAR NPL = CALCULATE(
    SUM(Prets[Solde]),
    Prets[JoursRetard] > 90
```

```
)
VAR Total = SUM(Prets[Solde])
RETURN
DIVIDE(NPL, Total, 0) * 100
```

Q26. Implémentez une comparaison dynamique YTD vs même période année précédente.

R26.

```
YTD Actuel = TOTALYTD(SUM(Ventes[Montant]), Calendrier[Date])

YTD N-1 = CALCULATE(
    TOTALYTD(SUM(Ventes[Montant]), Calendrier[Date]),
    SAMEPERIODLASTYEAR(Calendrier[Date])
)

Var YTD = DIVIDE([YTD Actuel] - [YTD N-1], [YTD N-1], 0) * 100
```

Q27. Créez une mesure qui affiche dynamiquement “N/A” quand il n’y a pas de données.

R27.

```
Resultat =
VAR Valeur = SUM(Table[Montant])
RETURN
IF(ISBLANK(Valeur), "N/A", FORMAT(Valeur, "#,##0"))
```

Q28-34. [Suite questions DAX...]

Section D: Python et Data Analysis (8 questions)

Q35. Écrivez du code Python pour: a) Charger un CSV b) Identifier les valeurs manquantes c) Traiter les outliers

R35.

```python
import pandas as pd
import numpy as np

# a) Charger
df = pd.read_csv('data.csv')

# b) Valeurs manquantes
print(df.isnull().sum())
print(df.isnull().mean() * 100)  # Pourcentage

# c) Outliers (méthode IQR)
def detect_outliers_iqr(df, col):
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
```

```
    upper = Q3 + 1.5 * IQR
    outliers = df[(df[col] < lower) | (df[col] > upper)]
    return outliers, lower, upper

# Traitement: capping
def cap_outliers(df, col):
    outliers, lower, upper = detect_outliers_iqr(df, col)
    df[col] = df[col].clip(lower, upper)
    return df
```

Q36. Comment feriez-vous une analyse RFM en Python?

R36.

```
def rfm_analysis(df, customer_id, date_col, amount_col, reference_date=None):
    if reference_date is None:
        reference_date = df[date_col].max()

    rfm = df.groupby(customer_id).agg({
        date_col: lambda x: (reference_date - x.max()).days,
        customer_id: 'count',
        amount_col: 'sum'
    })
    rfm.columns = ['Recency', 'Frequency', 'Monetary']

    # Scoring
    rfm['R_Score'] = pd.qcut(rfm['Recency'], 5, labels=[5,4,3,2,1])
    rfm['F_Score'] = pd.qcut(rfm['Frequency'].rank(method='first'), 5, labels=[1,2,3,4,5])
    rfm['M_Score'] = pd.qcut(rfm['Monetary'].rank(method='first'), 5, labels=[1,2,3,4,5])

    rfm['RFM_Score'] = rfm['R_Score'].astype(str) + rfm['F_Score'].astype(str) + rfm['M_Score'].astype(s

    return rfm
```

Q37-42. [Suite questions Python...]

Section E: Business et KPIs Bancaires (8 questions)

Q43. Listez les 5 KPIs les plus importants pour le CEO d'une banque et expliquez pourquoi.

R43. 1. **ROE (12-18%):** Rentabilité pour les actionnaires - objectif ultime 2. **NPL Ratio (<5%):** Qualité du portefeuille - principal risque 3. **CAR (>12%):** Solvabilité - exigence réglementaire 4. **CIR (<55%):** Efficacité opérationnelle - compétitivité 5. **Croissance des dépôts:** Ressources pour financer les prêts

---

Q44. Comment calculer l'Expected Loss sur un portefeuille de prêts?

R44.

```
EL = PD × LGD × EAD

Pour un portefeuille:
```

```
EL_total = Σ(PD  × LGD  × EAD )

Exemple:
Prêt A: PD=5%, LGD=45%, EAD=100K → EL = 2,250
Prêt B: PD=3%, LGD=40%, EAD=200K → EL = 2,400
EL_portefeuille = 4,650
```

Q45. Expliquez les 3 lignes de défense en gestion des risques bancaires.

R45. 1. **1ère ligne - Opérationnel:** - Unités métier (agences, crédit) - Propriétaires du risque - Contrôles quotidiens

2. **2ème ligne - Risque & Conformité:**
   - Direction des risques
   - Compliance
   - Politiques et monitoring
3. **3ème ligne - Audit Interne:**
   - Indépendant
   - Évaluation des contrôles
   - Rapporte au Conseil

Q46-50. [Suite questions business...]

Scoring

Score	Niveau
0-15	À améliorer
16-25	Débutant
26-35	Intermédiaire
36-42	Avancé
43-47	Senior
48-50	Expert

Conseils Finaux

1. **Structurez vos réponses:** Problème → Méthode → Résultat → Interprétation
2. **Contextualisez:** Donnez des exemples bancaires
3. **Soyez honnête:** Si vous ne savez pas, dites-le et expliquez comment vous trouveriez
4. **Pensez business:** Les outils servent des objectifs métier