Test Statistiques Descriptives - Test 1

Sujet: Statistiques Descriptives
Niveau: Intermédiaire
Nombre de questions: 25

Questions et Réponses

Q1. Quelles sont les quatre échelles de mesure et donnez un exemple bancaire pour chacune?

R1. | Échelle | Caractéristique | Exemple Bancaire | |———|—————-|————-| | **Nominale** | Catégories sans ordre | Type de compte (épargne, courant) | | **Ordinale** | Catégories ordonnées | Rating crédit (A, B, C, D) | | **Intervalle** | Distances égales, pas de zéro absolu | Score de crédit (300-850) | | **Ratio** | Zéro absolu significatif | Montant du prêt (0 = pas de prêt) |

Q2. Quelle est la différence entre moyenne, médiane et mode? Quand utiliser chacun?

R2. | Mesure | Définition | Quand utiliser | |——–|———|————-| | **Moyenne** | $\Sigma x_i/n$ | Distribution symétrique, pas d'outliers | | **Médiane** | Valeur centrale | Distribution asymétrique, outliers | | **Mode** | Valeur la plus fréquente | Données catégorielles, multimodales |

Exemple: Pour les salaires (distribution asymétrique), la médiane est préférable.

Q3. Calculez la moyenne, médiane et mode pour: [5, 7, 8, 8, 10, 12, 15, 45]

R3. - **Moyenne:** (5+7+8+8+10+12+15+45)/8 = 110/8 = **13.75** - **Médiane:** 8 valeurs → moyenne des 4ème et 5ème = (8+10)/2 = **9** - **Mode: 8** (apparaît 2 fois)

Observation: La moyenne (13.75) est tirée vers le haut par l'outlier (45). La médiane (9) est plus représentative.

Q4. Qu'est-ce que la variance et l'écart-type? Donnez les formules.

R4. Variance (échantillon):

$$s^2 = \Sigma(x - \bar{x})^2 / (n-1)$$

Écart-type:

$$s = \sqrt{s^2} = \sqrt{[\Sigma(x - \bar{x})^2 / (n-1)]}$$

Interprétation: Mesure la dispersion des données autour de la moyenne. - Écart-type faible → données concentrées - Écart-type élevé → données dispersées

---

Q5. Pourquoi divise-t-on par (n-1) et non par n pour l'écart-type échantillonnal?

R5. C'est la **correction de Bessel** pour obtenir un estimateur non biaisé de la variance de la population.

- Avec n: On sous-estime la variance (biased)
- Avec (n-1): Estimateur non biaisé (unbiased)

Explication: On perd un degré de liberté car on utilise $\bar{x}$ (calculé à partir des mêmes données) comme référence.

Q6. Qu'est-ce que le coefficient de variation (CV) et comment l'interpréter?

R6. Formule:

$$CV = (s / \bar{x}) \times 100\%$$

Interprétation: - CV < 15%: Faible variabilité (données homogènes) - 15% < CV < 30%: Variabilité modérée - CV > 30%: Forte variabilité (données hétérogènes)

Avantage: Permet de comparer la variabilité entre variables de différentes échelles.

Q7. Définissez les quartiles Q1, Q2, Q3 et l'IQR.

R7. - **Q1 (25ème percentile):** 25% des données sont en dessous - **Q2 (50ème percentile):** Médiane, 50% en dessous - **Q3 (75ème percentile):** 75% des données sont en dessous - **IQR (Interquartile Range):** Q3 - Q1

Utilisation: L'IQR mesure la dispersion de la moitié centrale des données, robuste aux outliers.

Q8. Calculez Q1, Q2, Q3 et IQR pour: [2, 4, 6, 8, 10, 12, 14, 16]

R8. Données ordonnées: 2, 4, 6, 8, 10, 12, 14, 16 (n=8)

- **Q1:** Médiane de [2, 4, 6, 8] = (4+6)/2 = **5**
- **Q2:** Médiane totale = (8+10)/2 = **9**
- **Q3:** Médiane de [10, 12, 14, 16] = (12+14)/2 = **13**
- **IQR:** Q3 - Q1 = 13 - 5 = **8**

Q9. Comment utiliser l'IQR pour détecter les outliers?

R9. Règle des 1.5 IQR:

$$\text{Borne inférieure} = Q1 - 1.5 \times IQR$$
$$\text{Borne supérieure} = Q3 + 1.5 \times IQR$$

Toute valeur hors de ces bornes est considérée comme outlier.

Exemple: Si Q1=5, Q3=13, IQR=8: - Borne inf: 5 - 1.5×8 = -7 - Borne sup: 13 + 1.5×8 = 25 - Outliers: $x < -7$ ou $x > 25$

Q10. Qu'est-ce que le skewness (asymétrie) et comment l'interpréter?

R10. Le **skewness** mesure l'asymétrie de la distribution.

Interprétation: - **Skew = 0:** Distribution symétrique (moyenne ≈ médiane) - **Skew > 0:** Asymétrie positive (queue à droite, moyenne > médiane) - **Skew < 0:** Asymétrie négative (queue à gauche, moyenne < médiane)

Règle pratique: - |skew| < 0.5: Approximativement symétrique - 0.5 < |skew| < 1: Modérément asymétrique - |skew| > 1: Fortement asymétrique

---

Q11. Qu'est-ce que le kurtosis et comment l'interpréter?

R11. Le **kurtosis** mesure l'"épaisseur" des queues de la distribution.

Kurtosis excédentaire (par rapport à la normale): - **Kurt = 0:** Mésokurtique (comme la normale) - **Kurt > 0:** Leptokurtique (queues lourdes, plus de valeurs extrêmes) - **Kurt < 0:** Platykurtique (queues légères, moins de valeurs extrêmes)

Impact bancaire: Un kurtosis élevé sur les pertes indique plus d'événements extrêmes.

Q12. Comment construire un tableau de fréquences?

R12.

```
# Fréquence absolue
freq_abs = df['categorie'].value_counts()

# Fréquence relative
freq_rel = df['categorie'].value_counts(normalize=True)

# Tableau complet
freq_table = pd.DataFrame({
    'Fréquence': freq_abs,
    'Pourcentage': freq_rel * 100,
    'Cumul': freq_rel.cumsum() * 100
})
```

Q13. Qu'est-ce qu'un tableau de contingence et à quoi sert-il?

R13. Un **tableau de contingence** (crosstab) montre la distribution jointe de deux variables catégorielles.

```
contingency = pd.crosstab(df['secteur'], df['defaut'], margins=True)
```

Utilisation: - Analyser la relation entre deux variables catégorielles - Base pour le test du Chi-carré - Calculer des probabilités conditionnelles

---

Q14. Comment calculer les percentiles en Python?

R14.

```
import numpy as np

# Percentile unique
p90 = np.percentile(df['montant'], 90)

# Plusieurs percentiles
percentiles = np.percentile(df['montant'], [10, 25, 50, 75, 90])

# Avec Pandas
df['montant'].quantile([0.1, 0.25, 0.5, 0.75, 0.9])

# Rank en percentile
df['percentile_rank'] = df['montant'].rank(pct=True) * 100
```

Q15. Quelle est la relation entre moyenne, médiane et mode selon la forme de la distribution?

R15. | Distribution | Relation | |———————-|———-| | **Symétrique** | Mode ≈ Médiane ≈ Moyenne | | **Asymétrique droite** | Mode < Médiane < Moyenne | | **Asymétrique gauche** | Moyenne < Médiane < Mode |

Règle empirique (distribution unimodale):

```
Moyenne - Mode   3 × (Moyenne - Médiane)
```

Q16. Comment calculer la moyenne pondérée?

R16. Formule:

```
x̄  = Σ(w  × x ) / Σw
```

Exemple bancaire: Taux moyen pondéré par montant

```
# Méthode 1: np.average
weighted_mean = np.average(df['taux'], weights=df['montant'])

# Méthode 2: calcul manuel
weighted_mean = (df['taux'] * df['montant']).sum() / df['montant'].sum()
```

Q17. Qu'est-ce que l'étendue (range) et quelles sont ses limites?

R17. Formule:

```
Étendue = Max - Min
```

Limites: - Très sensible aux outliers - N'utilise que 2 valeurs - Ne donne pas d'information sur la distribution - Peut augmenter avec la taille de l'échantillon

Alternative robuste: IQR (Interquartile Range)

Q18. Comment interpréter la moyenne et l'écart-type avec la règle empirique (68-95-99.7)?

R18. Pour une distribution **normale:** - ~68% des données sont dans [μ - σ, μ + σ] - ~95% des données sont dans [μ - 2σ, μ + 2σ] - ~99.7% des données sont dans [μ - 3σ, μ + 3σ]

Application bancaire: Si le montant moyen de prêt = 100K avec σ = 20K: - 68% des prêts entre 80K et 120K - 95% entre 60K et 140K - 99.7% entre 40K et 160K

Q19. Comment calculer les statistiques descriptives par groupe?

R19.

```
# Par groupe unique
df.groupby('agence')['montant'].describe()

# Statistiques personnalisées
stats = df.groupby('agence')['montant'].agg([
    'count',
    'mean',
    'median',
    'std',
    ('cv', lambda x: x.std() / x.mean() * 100),
```

```
    ('skew', 'skew')
])
```

Q20. Qu'est-ce que la covariance et comment l'interpréter?

R20. Formule:

$$Cov(X,Y) = \Sigma(x - \bar{x})(y - \bar{y}) / (n-1)$$

Interprétation: - **Cov > 0:** Relation positive (X ↑ quand Y ↑) - **Cov < 0:** Relation négative (X ↓ quand Y ↑) - **Cov ≈ 0:** Pas de relation linéaire

Limite: Dépend des unités, pas comparable entre variables.

Q21. Quelle est la différence entre covariance et corrélation?

R21. | | Covariance | Corrélation | |-|———-|———-| | **Formule** | Cov(X,Y) | Cov(X,Y) / ($s_x \times s_y$) | | **Plage** | -∞ à +∞ | -1 à +1 | | **Unités** | Dépend des variables | Sans unité | | **Interprétation** | Difficile | Facile | | **Comparabilité** | Non | Oui |

Corrélation = Covariance standardisée

Q22. Comment calculer et interpréter le coefficient de corrélation de Pearson?

R22.

```
# Calcul
correlation = df['score_credit'].corr(df['montant'])
# ou
correlation = np.corrcoef(df['score_credit'], df['montant'])[0,1]
```

Interprétation: - **|r| > 0.7:** Forte corrélation - **0.3 < |r| < 0.7:** Corrélation modérée - **|r| < 0.3:** Faible corrélation

Attention: Mesure uniquement les relations LINÉAIRES.

Q23. Quelles sont les mesures robustes aux outliers?

R23. | Mesure Standard | Alternative Robuste | |———-|———-| | Moyenne | **Médiane** | | Écart-type | **IQR** ou **MAD** | | Range | **IQR** | | Corrélation Pearson | **Spearman** |

MAD (Median Absolute Deviation):

$$MAD = median(|x - median(x)|)$$

Q24. Comment présenter un résumé statistique complet (five-number summary)?

R24. Le **five-number summary:** 1. **Minimum** 2. **Q1 (25ème percentile)** 3. **Médiane (Q2)** 4. **Q3 (75ème percentile)** 5. **Maximum**

```
df['montant'].describe()

# Manuel
summary = {
    'min': df['montant'].min(),
    'Q1': df['montant'].quantile(0.25),
```

```
    'median': df['montant'].median(),
    'Q3': df['montant'].quantile(0.75),
    'max': df['montant'].max()
}
```

Q25. Analysez les statistiques suivantes et donnez vos conclusions: - Montant moyen: 85,000 HTG - Montant médian: 52,000 HTG - Écart-type: 95,000 HTG - Skewness: 2.3 - Kurtosis: 8.1

R25. Analyse: 1. **Moyenne » Médiane (85K vs 52K):** Asymétrie droite confirmée 2. **Skewness = 2.3:** Fortement asymétrique positif 3. **CV = 95/85 = 112%:** Très forte variabilité 4. **Kurtosis = 8.1:** Queues très lourdes, valeurs extrêmes

Conclusions: - Distribution très asymétrique avec outliers élevés - **Utiliser la médiane** (52K) comme mesure de tendance centrale - **Utiliser l'IQR** pour la dispersion - Présence probable de quelques très gros prêts - Envisager une **transformation log** pour l'analyse

---

Scoring

Score	Niveau
0-10	À améliorer
11-17	Intermédiaire
18-22	Avancé
23-25	Expert