Manuel de Préparation: Concepts Clés de Probabilité

Introduction

La probabilité est le fondement mathématique de l'incertitude. Pour un Data Analyst en milieu bancaire, comprendre les probabilités est essentiel pour évaluer les risques, interpréter les statistiques, et prendre des décisions basées sur des données incomplètes.

Partie 1: Concepts Fondamentaux

1.1 Définitions de Base

Expérience Aléatoire Processus dont le résultat ne peut être prédit avec certitude.

```
Exemple bancaire: Un client rembourse-t-il son prêt?
```

Espace Échantillonnal (Ω) Ensemble de tous les résultats possibles.

```
Ω = {Remboursement complet, Défaut partiel, Défaut total}
```

Événement Sous-ensemble de l'espace échantillonnal.

```
A = "Le client fait défaut" = {Défaut partiel, Défaut total}
```

Probabilité Mesure de la vraisemblance d'un événement, entre 0 et 1.

```
P(A) = Nombre de cas favorables / Nombre de cas possibles

0   P(A)   1
P(Ω) = 1 (certitude)
P( ) = 0 (impossibilité)
```

1.2 Types de Probabilités

Type	Définition	Exemple
Classique	Cas équiprobables	Tirage d'une carte
Fréquentiste	Fréquence à long terme	Taux de défaut historique
Subjective	Degré de croyance	Expert estime 70% de succès

Probabilité Fréquentiste (la plus utilisée en Data)

```
P(A) = lim(n→∞) [Nombre de fois où A se produit / n]

Exemple:
Sur 10,000 prêts historiques, 500 ont fait défaut
P(Défaut) = 500 / 10,000 = 0.05 = 5%
```

1.3 Règles Fondamentales

Complément

```
P(A') = 1 - P(A)

Exemple: Si P(Défaut) = 5%, alors P(Non-défaut) = 95%
```

Addition (Union)

```
P(A  B) = P(A) + P(B) - P(A  B)

Si A et B mutuellement exclusifs:
P(A  B) = P(A) + P(B)

Exemple:
P(Client Jeune OU Client Senior) = P(Jeune) + P(Senior) - P(Jeune ET Senior)
```

Multiplication (Intersection)

```
P(A  B) = P(A) × P(B|A)

Si A et B indépendants:
P(A  B) = P(A) × P(B)

Exemple:
P(Défaut Prêt1 ET Défaut Prêt2) = P(Défaut1) × P(Défaut2) si indépendants
```

Partie 2: Probabilité Conditionnelle

2.1 Définition

Probabilité d'un événement A sachant que B s'est produit.

```
P(A|B) = P(A  B) / P(B)

Exemple:
P(Défaut | Revenu < 20K) = P(Défaut ET Revenu < 20K) / P(Revenu < 20K)

Sur 1000 clients:
- 200 ont Revenu < 20K
- 30 ont Revenu < 20K ET Défaut

P(Défaut | Revenu < 20K) = 30/200 = 15%
```

2.2 Indépendance

Deux événements sont indépendants si:

```
P(A|B) = P(A)  OU  P(A  B) = P(A) × P(B)

Si P(Défaut | Revenu)  P(Défaut), alors Défaut et Revenu sont DÉPENDANTS
```

2.3 Théorème de Bayes

Permet d’inverser les probabilités conditionnelles.

```
P(A|B) = [P(B|A) × P(A)] / P(B)

Ou avec normalisation:
P(A|B) = [P(B|A) × P(A)] / [P(B|A) × P(A) + P(B|A') × P(A')]
```

Exemple Bancaire: Détection de Fraude

```
Données:
- P(Fraude) = 0.01 (1% des transactions sont frauduleuses)
- P(Alerte | Fraude) = 0.95 (95% des fraudes déclenchent une alerte)
- P(Alerte | Non-fraude) = 0.05 (5% de faux positifs)

Question: Si une alerte est déclenchée, quelle est la probabilité de fraude?

P(Fraude | Alerte) = [P(Alerte|Fraude) × P(Fraude)] / P(Alerte)

P(Alerte) = P(Alerte|Fraude)×P(Fraude) + P(Alerte|Non-fraude)×P(Non-fraude)
P(Alerte) = 0.95 × 0.01 + 0.05 × 0.99 = 0.0095 + 0.0495 = 0.059

P(Fraude | Alerte) = (0.95 × 0.01) / 0.059 = 0.161   16%

Interprétation: Même avec une bonne détection, seulement 16% des alertes
sont de vraies fraudes (à cause de la faible prévalence)
```

Partie 3: Variables Aléatoires

3.1 Définition

Fonction qui associe une valeur numérique à chaque résultat d’une expérience aléatoire.

Type	Valeurs	Exemple
Discrète	Dénombrables	Nombre de transactions
Continue	Intervalle	Montant d’un prêt

3.2 Variable Aléatoire Discrète

Fonction de Masse de Probabilité (PMF)

```
P(X = x) pour chaque valeur x

Exemple: Nombre de produits détenus
X = 1: P(X=1) = 0.30
X = 2: P(X=2) = 0.40
X = 3: P(X=3) = 0.20
X = 4+: P(X 4) = 0.10
```

Fonction de Répartition (CDF)

```
F(x) = P(X   x)

F(2) = P(X   2) = P(X=1) + P(X=2) = 0.30 + 0.40 = 0.70
```

3.3 Variable Aléatoire Continue

Fonction de Densité de Probabilité (PDF)

```
f(x) où P(a   X   b) =  [a,b] f(x) dx

Propriété: f(x)   0 et  f(x)dx = 1
```

Partie 4: Espérance et Variance

4.1 Espérance (Moyenne Théorique)

Discrète

```
E[X] = Σ x  × P(X = x )

Exemple: Gain attendu d'un investissement
Gain 100K avec proba 0.3
Gain 50K avec proba 0.5
Perte 20K avec proba 0.2

E[Gain] = 100×0.3 + 50×0.5 + (-20)×0.2 = 30 + 25 - 4 = 51K
```

Continue

```
E[X] =   x × f(x) dx
```

Propriétés

```
E[aX + b] = aE[X] + b
E[X + Y] = E[X] + E[Y]
E[XY] = E[X]×E[Y] si X et Y indépendants
```

4.2 Variance et Écart-Type

Variance

```
Var(X) = E[(X - E[X])²] = E[X²] - (E[X])²

Mesure la dispersion autour de l'espérance
```

Écart-Type

```
  = √Var(X)
```

Propriétés

```
Var(aX + b) = a² × Var(X)
Var(X + Y) = Var(X) + Var(Y) si X et Y indépendants
```

4.3 Covariance et Corrélation

Covariance

```
Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]

Cov > 0: X et Y varient dans le même sens
Cov < 0: X et Y varient en sens opposé
Cov = 0: Pas de relation linéaire (pas nécessairement indépendants)
```

Corrélation (Coefficient de Pearson)

```
 = Cov(X,Y) / ( X ×  Y)

-1      1
```

Partie 5: Distributions Importantes

5.1 Distribution de Bernoulli

Usage: Événement binaire (succès/échec)

```
X ~ Bernoulli(p)

P(X = 1) = p (succès)
P(X = 0) = 1 - p (échec)

E[X] = p
Var(X) = p(1-p)

Exemple: Défaut d'un prêt (oui/non)
```

5.2 Distribution Binomiale

Usage: Nombre de succès sur n essais indépendants

```
X ~ Binomiale(n, p)

P(X = k) = C(n,k) × p^k × (1-p)^(n-k)

E[X] = np
Var(X) = np(1-p)

Exemple: Sur 100 prêts, combien feront défaut?
n = 100, p = 0.05
E[Défauts] = 100 × 0.05 = 5
```

```python
from scipy.stats import binom

# Probabilité d'exactement 3 défauts sur 100 prêts
p_3_defauts = binom.pmf(3, n=100, p=0.05)

# Probabilité de 5 défauts ou moins
p_5_ou_moins = binom.cdf(5, n=100, p=0.05)
```

5.3 Distribution de Poisson

Usage: Nombre d'événements rares sur une période

```
X ~ Poisson( )

P(X = k) = ( ^k × e^(- )) / k!

E[X] =
Var(X) =

Exemple: Nombre de fraudes par jour
  = 2 fraudes/jour en moyenne
P(3 fraudes) = (2³ × e^(-2)) / 3! = 0.18
```

```python
from scipy.stats import poisson

# Probabilité de 3 fraudes
p_3 = poisson.pmf(3, mu=2)

# Probabilité de plus de 5 fraudes
p_plus_5 = 1 - poisson.cdf(5, mu=2)
```

5.4 Distribution Normale (Gaussienne)

Usage: Variable continue symétrique (très commune)

```
X ~ N( ,  ²)

f(x) = (1 /  √2 ) × exp(-(x- )²/2 ²)

E[X] =
Var(X) =  ²

Propriété: 68-95-99.7 rule
```

Loi Normale Centrée Réduite

```
Z = (X -  ) /   ~ N(0, 1)

Permet de calculer des probabilités avec les tables Z
```

```python
from scipy.stats import norm

# P(X < 50) pour X ~ N(45, 10²)
p = norm.cdf(50, loc=45, scale=10)

# Valeur telle que P(X < x) = 0.95
x = norm.ppf(0.95, loc=45, scale=10)
```

5.5 Distribution Exponentielle

Usage: Temps entre événements (durée de vie, temps d'attente)

```
X ~ Exp( )
```

```
f(x) =  × e^(- x) pour x  0

E[X] = 1/
Var(X) = 1/ ²

Propriété: Sans mémoire
```

```python
from scipy.stats import expon

# Temps moyen entre transactions = 5 minutes ( = 1/5)
# P(temps > 10 minutes)
p = 1 - expon.cdf(10, scale=5)
```

5.6 Distribution Log-Normale

Usage: Variables positives asymétriques (montants, revenus)

```
Si Y ~ N( , ²), alors X = e^Y ~ LogNormale

Caractéristique: Asymétrie positive typique des données financières
```

Partie 6: Théorèmes Importants

6.1 Loi des Grands Nombres

```
Quand n → ∞, la moyenne échantillonnale converge vers l'espérance.

x̄ → E[X]

Implication: Plus on a de données, plus nos estimations sont précises.
```

6.2 Théorème Central Limite (TCL)

```
Pour n suffisamment grand (n  30), la distribution de la moyenne
échantillonnale suit approximativement une loi normale:

X̄ ~ N( , ²/n)

Quelle que soit la distribution originale!
```

Implication Pratique

```
Permet d'utiliser la distribution normale pour l'inférence statistique,
même si la population n'est pas normale.
```

Partie 7: Applications Bancaires

7.1 Probabilité de Défaut (PD)

```python
def calculate_pd(historical_defaults, historical_loans):
    """Calcul simple de la probabilité de défaut"""
    return historical_defaults / historical_loans
```

```python
def expected_loss(pd, lgd, ead):
    """
    Perte attendue
    PD: Probability of Default
    LGD: Loss Given Default (% de perte si défaut)
    EAD: Exposure at Default (montant exposé)
    """
    return pd * lgd * ead

# Exemple
pd = 0.05  # 5% de probabilité de défaut
lgd = 0.45  # 45% de perte si défaut
ead = 100000  # Exposition de 100K

el = expected_loss(pd, lgd, ead)
print(f"Perte attendue: {el:,.0f} HTG")  # 2,250 HTG
```

7.2 Value at Risk (VaR)

```python
from scipy.stats import norm

def calculate_var(mean_return, std_return, confidence=0.95, investment=1000000):
    """
    Value at Risk paramétrique
    Perte maximale avec un niveau de confiance donné
    """
    z = norm.ppf(1 - confidence)
    var = investment * (mean_return + z * std_return)
    return -var  # Perte (valeur positive)

# Exemple
var = calculate_var(mean_return=0.001, std_return=0.02, confidence=0.95)
print(f"VaR 95%: {var:,.0f} HTG")
```

7.3 Simulation Monte Carlo

```python
import numpy as np

def monte_carlo_default(num_loans, pd, num_simulations=10000):
    """
    Simulation du nombre de défauts dans un portefeuille
    """
    defaults = np.random.binomial(num_loans, pd, num_simulations)

    return {
        'mean': defaults.mean(),
        'std': defaults.std(),
        'percentile_95': np.percentile(defaults, 95),
        'percentile_99': np.percentile(defaults, 99)
    }

# Exemple: 1000 prêts avec PD = 5%
results = monte_carlo_default(1000, 0.05)
```

```python
print(f"Défauts attendus: {results['mean']:.0f}")
print(f"Défauts (95%): {results['percentile_95']:.0f}")
```

7.4 Scoring de Crédit

```python
def score_to_pd(score, base_pd=0.10, score_range=(300, 850)):
    """
    Convertit un score de crédit en probabilité de défaut
    Score élevé = faible PD
    """
    min_score, max_score = score_range
    normalized = (score - min_score) / (max_score - min_score)

    # Transformation logistique
    odds = (1 - base_pd) / base_pd  # Odds de base
    adjusted_odds = odds * np.exp(3 * (normalized - 0.5))
    pd = 1 / (1 + adjusted_odds)

    return pd

# Exemple
for score in [400, 550, 700, 800]:
    pd = score_to_pd(score)
    print(f"Score {score}: PD = {pd:.2%}")
```

Partie 8: Formules Essentielles

Probabilités de Base

```
P(A') = 1 - P(A)
P(A   B) = P(A) + P(B) - P(A   B)
P(A   B) = P(A) × P(B|A) = P(B) × P(A|B)
P(A|B) = P(A   B) / P(B)
```

Bayes

```
P(A|B) = [P(B|A) × P(A)] / P(B)
```

Espérance et Variance

```
E[X] = Σ x  P(x )  ou    x f(x) dx
Var(X) = E[X²] - (E[X])²
```

Distributions Clés

```
Binomiale: E = np, Var = np(1-p)
Poisson: E =  , Var =
Normale: E =  , Var =  ²
```

Questions d’Entretien

1. **Qu’est-ce que le théorème de Bayes et comment l’appliquer?** → Permet d’inverser les probabilités conditionnelles; ex: P(Fraude|Alerte)
2. **Différence entre distribution binomiale et Poisson?** → Binomiale: n essais finis; Poisson: événements rares sur une période
3. **Pourquoi le TCL est-il important?** → Permet d’utiliser la normale pour l’inférence même si population non normale
4. **Comment calculer une perte attendue (EL)?** → EL = PD × LGD × EAD
5. **Qu’est-ce qu’une variable aléatoire?** → Fonction qui associe une valeur numérique aux résultats d’une expérience aléatoire

---

Checklist Probabilités

```
Distinguer probabilité marginale, conditionnelle, conjointe
Appliquer Bayes pour inverser les probabilités
Calculer espérance et variance
Identifier la distribution appropriée au problème
Utiliser le TCL pour l'inférence
Interpréter les résultats dans le contexte business
```

Rappel: La probabilité est le langage de l’incertitude. Maîtriser ces concepts permet de quantifier les risques et de prendre des décisions éclairées malgré l’information incomplète.