Test A/B Testing et Expérimentation - Niveau 2 (Avancé)

UniBank Haiti - Data Analyst

Durée: 45 minutes
Questions: 25
Niveau: Avancé
Sujets: Calculs avancés, approche bayésienne, pièges, éthique, cas complexes

---

Q1. UniBank veut détecter une amélioration de 2pp (de 5% à 7%) sur le taux de conversion. Avec α=0.05 et puissance=80%, la taille approximative par groupe est:

A) 100 clients

B) 500 clients

C) 1,000-1,500 clients

D) 10,000 clients

Réponse: C) 1,000-1,500 clients

Pour proportions, avec baseline 5%, MDE 2pp, α=0.05, puissance 80%, la formule donne environ 1,000-1,500 par groupe. La formule exacte donne $n \approx 16 \times p(1-p)/\delta^2 \approx 1,200$.

Q2. Si vous doublez le MDE (de 2pp à 4pp) dans l'exemple précédent, la taille d'échantillon nécessaire:

A) Double

B) Est divisée par 4 (proportionnelle à $1/MDE^2$)

C) Reste identique

D) Est divisée par 2

Réponse: B) Est divisée par 4 (proportionnelle à $1/MDE^2$)

La taille est inversement proportionnelle au carré du MDE: $n \propto 1/\delta^2$. Si δ double, n est divisé par 4.

Q3. Vous effectuez un test séquentiel (regarder les résultats à intervalles réguliers). Quelle procédure utiliser pour maintenir α = 0.05 global?

A) Utiliser α = 0.05 à chaque vérification

B) Utiliser une correction de type O'Brien-Fleming ou spending function

C) Ne jamais regarder avant la fin

D) Doubler α

Réponse: B) Utiliser une correction de type O'Brien-Fleming ou spending function

Les méthodes séquentielles (group sequential designs) ajustent les seuils à chaque analyse intermédiaire pour maintenir l'erreur globale. O'Brien-Fleming est plus strict au début.

Q4. L'approche bayésienne de l'A/B testing diffère de l'approche fréquentiste par:

A) Elle ne nécessite pas de données

B) Elle fournit une probabilité directe que B > A (ex: "87% de chances que B soit meilleur")

C) Elle est toujours plus conservative

D) Elle ne peut pas être utilisée pour les proportions

Réponse: B) Elle fournit une probabilité directe que B > A (ex: "87% de chances que B soit meilleur")

L'approche bayésienne calcule $P(\theta_B > \theta_A \mid data)$, interprétable directement. Le fréquentisme donne seulement p-value (probabilité des données sous H_0).

Q5. Dans un test bayésien, le "prior" représente:

A) Les données collectées

B) La croyance initiale sur les paramètres avant le test

C) La décision finale

D) La taille d'échantillon

Réponse: B) La croyance initiale sur les paramètres avant le test

Le prior encode ce qu'on sait avant l'expérience. Avec des données, il est mis à jour en posterior via Bayes: posterior $\propto$ likelihood $\times$ prior.

Q6. Vous obtenez P(B > A) = 0.92 dans une analyse bayésienne. Avec un seuil de décision de 0.95, que faire?

A) Déployer B immédiatement

B) Continuer le test - la probabilité n'atteint pas encore le seuil

C) Abandonner le test

D) Déployer A

Réponse: B) Continuer le test - la probabilité n'atteint pas encore le seuil

92% < 95%, donc on n'a pas suffisamment de certitude. On continue à collecter des données jusqu'à ce que P(B>A) > 0.95 ou P(A>B) > 0.95.

Q7. Le “expected loss” en A/B testing bayésien mesure:

A) Les pertes financières passées

B) La perte moyenne attendue si on prend une mauvaise décision

C) Le coût du test

D) L’erreur de mesure

Réponse: B) La perte moyenne attendue si on prend une mauvaise décision

L’expected loss combine la probabilité de se tromper et l’ampleur de l’erreur. On peut arrêter quand l’expected loss est acceptable (ex: < 0.1%).

Q8. UniBank teste une nouvelle tarification. Le groupe B (nouveaux frais) montre +5% de revenus mais +2% de churn. Comment décider?

A) Déployer car les revenus augmentent

B) Ne pas déployer car le churn augmente

C) Calculer l’impact net sur le LTV/revenus long terme et décider sur cette base

D) Refaire le test

Réponse: C) Calculer l’impact net sur le LTV/revenus long terme et décider sur cette base

Les effets contradictoires nécessitent une analyse business: gain revenus court terme vs perte clients long terme. Calculer l’impact net (revenus - perte LTV) guide la décision.

Q9. Le test de ratio (ex: revenu moyen par client) a souvent plus de variance que le test de proportion. Comment compenser?

A) Utiliser moins de données

B) Augmenter la taille d’échantillon ou utiliser des techniques de réduction de variance (CUPED)

C) Ignorer la variance

D) Convertir en proportion

Réponse: B) Augmenter la taille d’échantillon ou utiliser des techniques de réduction de variance (CUPED)

Les métriques continues ont généralement plus de variance. CUPED (Controlled-experiment Using Pre-Experiment Data) utilise les données pré-test pour réduire la variance.

Q10. CUPED réduit la variance en:

A) Supprimant les outliers

B) Utilisant les données pré-expérience comme covariable pour ajuster les différences préexistantes

C) Doublant la taille d'échantillon

D) Changeant la métrique

Réponse: B) Utilisant les données pré-expérience comme covariable pour ajuster les différences préexistantes

CUPED ajuste pour la variance expliquée par le comportement pré-test: Y_adj = Y - θ(X_pre - E[X_pre]). Cela réduit la variance et augmente la puissance.

Q11. Un test montre un effet significatif globalement mais non significatif pour le segment Premium. Que conclure?

A) L'effet n'existe pas

B) L'effet peut varier par segment - interaction possible (effet hétérogène)

C) Le segment Premium doit être exclu

D) Le test global est faux

Réponse: B) L'effet peut varier par segment - interaction possible (effet hétérogène)

Les effets hétérogènes par segment sont fréquents. L'effet global peut être "dilué" par un segment non-répondant. L'analyse par segment révèle où le traitement fonctionne.

Q12. Vous analysez par 10 segments sans correction. Quel est le risque approximatif d'au moins un faux positif (α=0.05)?

A) 5%

B) 50%

C) 40% ($1 - 0.95^{10}$)

D) 10%

Réponse: C) 40% ($1 - 0.95^{10}$)

Le risque familywise = 1 - (1-α)^k = 1 - 0.95^10 ≈ 40%. C'est le problème des comparaisons multiples.

Q13. L'effet de "network" ou contamination dans un A/B test bancaire se produit quand:

A) Le réseau internet est lent

B) Les clients du groupe A interagissent avec ceux du groupe B (ex: transferts, bouche-à-oreille)

C) Le test dure trop longtemps

D) Les données sont corrompues

Réponse: B) Les clients du groupe A interagissent avec ceux du groupe B (ex: transferts, bouche-à-oreille)

Si un client B convainc un client A d'un nouveau produit, l'effet "contamine" le groupe contrôle. Solutions: randomisation par cluster (agence, région).

Q14. Pour éviter la contamination réseau, vous randomisez par agence. 50 agences: 25 contrôle, 25 traitement. Quel est le problème potentiel?

A) Aucun problème

B) Les agences peuvent différer (taille, région) - besoin de stratification ou cluster size suffisant

C) Trop d'agences

D) Les clients sont mécontents

Réponse: B) Les agences peuvent différer (taille, région) - besoin de stratification ou cluster size suffisant

Avec peu de clusters (50), les différences inter-clusters peuvent biaiser. Il faut stratifier par caractéristiques (taille, région) ou avoir plus de clusters.

Q15. Le "bucket" en A/B testing fait référence à:

A) Un seau physique

B) L'identifiant de groupe (A ou B) assigné à chaque utilisateur

C) Le budget du test

D) La base de données

Réponse: B) L'identifiant de groupe (A ou B) assigné à chaque utilisateur

Chaque utilisateur est "bucketed" dans un groupe. Le bucket est généralement déterminé par un hash de l'ID utilisateur pour être reproductible.

Q16. Un hash de l'ID client pour l'assignation A/B garantit:

A) Des groupes de taille exactement égale

B) Que le même client est toujours dans le même groupe (reproductibilité)

C) L'anonymat

D) Des résultats significatifs

Réponse: B) Que le même client est toujours dans le même groupe (reproductibilité)

hash(client_id) % 100 donne toujours le même résultat pour le même client. Cela assure une expérience cohérente sur plusieurs sessions.

Q17. UniBank doit tester une réduction de taux d'intérêt sur les prêts. Quelle considération éthique est primordiale?

A) Aucune, c'est un test normal

B) Les clients du groupe contrôle ne doivent pas être lésés - considérer une communication transparente

C) Augmenter la durée du test

D) Exclure les clients VIP

Réponse: B) Les clients du groupe contrôle ne doivent pas être lésés - considérer une communication transparente

Les tests de prix/taux peuvent désavantager un groupe. Considérations: durée limitée, communication, remboursement de la différence post-test si bénéfique.

Q18. Un test sur des produits de crédit doit respecter quelle contrainte réglementaire?

A) Aucune

B) Non-discrimination (même traitement indépendant de caractéristiques protégées)

C) Avoir plus de 100 clients

D) Être approuvé par la concurrence

Réponse: B) Non-discrimination (même traitement indépendant de caractéristiques protégées)

La randomisation doit être indépendante des caractéristiques protégées (genre, âge, origine). Un test discriminatoire violerait les réglementations anti-discrimination.

Q19. Le ratio de variance (variance ratio test) avant un A/B test vérifie:

A) Si les groupes ont des tailles égales

B) Si les variances pré-traitement sont similaires entre groupes (validation randomisation)

C) La normalité des données

D) Le taux de conversion

Réponse: B) Si les variances pré-traitement sont similaires entre groupes (validation randomisation)

Comparer les variances (ou moyennes) pré-test entre A et B vérifie que la randomisation a créé des groupes comparables.

Q20. L'effet "primacy" dans un A/B test d'interface signifie:

A) Le test est terminé en premier

B) Les utilisateurs préfèrent ce qu'ils voient en premier (ordre d'exposition importe)

C) La première métrique est la plus importante

D) Le groupe A est toujours premier

Réponse: B) Les utilisateurs préfèrent ce qu'ils voient en premier (ordre d'exposition importe)

Pour les tests où les utilisateurs voient plusieurs options (ex: liste de produits), l'ordre peut biaiser. Solution: randomiser aussi l'ordre.

Q21. Un "guardrail metric" dans un A/B test est:

A) La métrique principale

B) Une métrique qui ne doit pas se dégrader significativement (ex: satisfaction, bugs)

C) La taille d'échantillon

D) Le budget

Réponse: B) Une métrique qui ne doit pas se dégrader significativement (ex: satisfaction, bugs)

Les guardrails protègent contre les effets secondaires négatifs. Même si la métrique primaire s'améliore, on ne déploie pas si un guardrail est violé.

Q22. La correction de Benjamini-Hochberg (FDR) est préférée à Bonferroni quand:

A) On a peu de comparaisons

B) On accepte un certain taux de faux positifs parmi les significatifs (moins conservateur)

C) Les données sont normales

D) Il n'y a qu'une seule métrique

Réponse: B) On accepte un certain taux de faux positifs parmi les significatifs (moins conservateur)

FDR contrôle la proportion de faux positifs parmi les résultats significatifs (pas parmi tous les tests). Moins conservateur, plus de puissance.

Q23. Vous avez un effet lift de +10% mais un intervalle de confiance de [-5%, +25%]. La décision appropriée est:

A) Déployer car le lift est positif

B) Ne pas déployer ou continuer le test - l'IC inclut des effets négatifs possibles

C) L'IC est trop large, le test est invalide

D) Ignorer l'IC

Réponse: B) Ne pas déployer ou continuer le test - l'IC inclut des effets négatifs possibles

Un IC contenant 0 (et des négatifs) signifie qu'on n'est pas certain que B est meilleur. Le vrai effet pourrait être négatif. Plus de données réduiraient l'IC.

Q24. Le "Simpson's Paradox" dans un A/B test peut se produire quand:

A) Les données sont manquantes

B) Un effet global positif peut masquer des effets négatifs dans tous les segments (ou inversement)

C) Le test est trop court

D) La randomisation échoue

Réponse: B) Un effet global positif peut masquer des effets négatifs dans tous les segments (ou inversement)

Simpson's Paradox: une tendance présente dans chaque sous-groupe peut s'inverser dans l'agrégat. Toujours analyser par segment pour détecter ce phénomène.

Q25. Pour un test long-terme sur la rétention client (6 mois), quelle approche est recommandée?

A) A/B test classique de 6 mois

B) Holdout group permanent avec analyse longitudinale

C) Ne pas tester la rétention

D) Tester pendant 1 semaine et extrapoler

Réponse: B) Holdout group permanent avec analyse longitudinale

Pour les effets long-terme, maintenir un petit groupe holdout (contrôle) permanent permet de mesurer les différences sur des mois/années sans contaminer toute la base.

Résumé des Concepts Avancés

Approche Bayésienne

Concept	Description
Prior	Croyance initiale
Posterior	Croyance mise à jour avec données
P(B > A)	Probabilité directe que B est meilleur
Expected loss	Perte moyenne si mauvaise décision

Techniques de Réduction de Variance

Technique	Description
CUPED	Ajustement par données pré-test
Stratification	Équilibre forcé par segment
Blocking	Appariement de sujets similaires

Corrections Comparaisons Multiples

Méthode	Type	Usage
Bonferroni	FWER	Conservative, peu de tests
Holm	FWER step-down	Moins conservative
Benjamini-Hochberg	FDR	Beaucoup de tests, acceptable faux positifs

Considérations Éthiques

- Non-discrimination dans la randomisation
- Durée limitée pour tests de prix
- Transparence avec les clients
- Protection des groupes vulnérables

Score: ___/25