Test A/B Testing et Expérimentation - Niveau 1 (Intermédiaire)

UniBank Haiti - Data Analyst

Durée: 40 minutes
Questions: 25
Niveau: Intermédiaire
Sujets: Design expérimental, taille d'échantillon, analyse de résultats

---

Q1. Qu'est-ce qu'un A/B test dans le contexte bancaire?

A) Une comparaison de deux agences bancaires

B) Une expérience contrôlée randomisée comparant deux versions d'un élément

C) Un test de conformité réglementaire

D) Une analyse de deux trimestres consécutifs

Réponse: B) Une expérience contrôlée randomisée comparant deux versions d'un élément

L'A/B test est une méthode expérimentale où les sujets sont randomisés en deux groupes: contrôle (version actuelle) et traitement (nouvelle version) pour mesurer l'impact d'un changement.

Q2. Pourquoi la randomisation est-elle essentielle dans un A/B test?

A) Pour accélérer le test

B) Pour garantir que les groupes sont comparables et éliminer les biais de sélection

C) Pour réduire la taille d'échantillon

D) Pour satisfaire une exigence légale

Réponse: B) Pour garantir que les groupes sont comparables et éliminer les biais de sélection

La randomisation assure que les différences observées entre groupes sont dues au traitement et non à des caractéristiques préexistantes différentes (confondants).

Q3. UniBank veut tester un nouvel email marketing pour les prêts. Le taux de conversion actuel est 4%. Quelle est la "baseline" dans ce contexte?

A) Le nombre total de clients

B) Le taux de conversion actuel de 4% (groupe contrôle)

C) Le nombre d'emails envoyés

D) Le budget marketing

Réponse: B) Le taux de conversion actuel de 4% (groupe contrôle)

La baseline est le taux de performance actuel contre lequel on mesure l'amélioration. Ici, 4% de conversion avec l'email actuel.

Q4. Qu'est-ce que l'Effet Minimal Détectable (MDE)?

A) L'effet maximum possible

B) La plus petite différence pratiquement significative qu'on souhaite détecter

C) L'erreur de mesure

D) La taille de l'échantillon

Réponse: B) La plus petite différence pratiquement significative qu'on souhaite détecter

Le MDE définit le seuil d'amélioration minimum intéressant pour le business. Si on ne peut pas détecter une différence de +1%, un résultat de +0.5% sera non-significatif.

Q5. Avec une puissance statistique de 80%, quelle est la probabilité de détecter un effet réel s'il existe?

A) 20%

B) 80%

C) 95%

D) 50%

Réponse: B) 80%

La puissance = 1 - β = P(rejeter H_0 | H_1 vraie) = P(détecter un vrai effet). 80% est le standard conventionnel.

Q6. À quoi correspond une erreur de Type I dans un A/B test?

A) Déployer une version qui n'est pas meilleure (faux positif)

B) Ne pas déployer une version qui est meilleure (faux négatif)

C) Avoir trop de données

D) Mal calculer la taille d'échantillon

Réponse: A) Déployer une version qui n'est pas meilleure (faux positif)

Type I = rejeter H_0 à tort = conclure que B est meilleur alors qu'il ne l'est pas. On déploie un changement inutile ou contre-productif.

Q7. Vous planifiez un A/B test avec α=0.05, puissance=80%, baseline=5%, MDE=1%. Qu'augmente la taille d'échantillon nécessaire?

A) Augmenter α à 0.10

B) Augmenter la puissance à 90%

C) Augmenter le MDE à 2%

D) Augmenter la baseline à 10%

Réponse: B) Augmenter la puissance à 90%

Plus de puissance nécessite plus de données. Inversement, augmenter α (moins strict) ou MDE (effet plus grand à détecter) réduit la taille nécessaire.

Q8. Quelle est la durée minimale recommandée pour un A/B test, même si la taille d'échantillon est atteinte avant?

A) 1 jour

B) 7 jours minimum (un cycle complet)

C) 1 mois obligatoire

D) Aucune durée minimale

Réponse: B) 7 jours minimum (un cycle complet)

Un cycle complet (semaine) capture les variations jour/jour (lundi ≠ dimanche). Arrêter trop tôt peut biaiser les résultats si certains jours sont surreprésentés.

Q9. Le "peeking" dans un A/B test fait référence à:

A) Espionner la concurrence

B) Regarder les résultats avant la fin prévue et potentiellement arrêter le test précocement

C) Vérifier les données manquantes

D) Analyser les segments

Réponse: B) Regarder les résultats avant la fin prévue et potentiellement arrêter le test précocement

Le peeking répété augmente la probabilité de faux positifs. Si on vérifie tous les jours et arrête dès que $p < 0.05$, le vrai taux d'erreur peut atteindre 30%+.

Q10. UniBank teste une nouvelle interface d'ouverture de compte. 5000 clients sont exposés à chaque version. Version A: 250 comptes ouverts. Version B: 300 comptes ouverts. Quels sont les taux de conversion?

A) A: 5%, B: 6%

B) A: 4%, B: 5%

C) A: 2.5%, B: 3%

D) A: 50%, B: 60%

Réponse: A) A: 5%, B: 6%

Taux A = 250/5000 = 5%, Taux B = 300/5000 = 6%. Le lift est (6%-5%)/5% = 20% relatif ou +1pp absolu.

Q11. Pour l'exemple précédent (A: 5%, B: 6%), le test z pour proportions donne p = 0.03. Au seuil α = 0.05, quelle décision?

A) Pas de différence significative

B) Différence significative - B est meilleur

C) Besoin de plus de données

D) Le test est invalide

Réponse: B) Différence significative - B est meilleur

$p = 0.03 < \alpha = 0.05$, donc on rejette H_0. La différence est statistiquement significative. B a un taux de conversion significativement plus élevé.

Q12. Qu'est-ce que le "lift" en A/B testing?

A) Le nombre absolu de conversions supplémentaires

B) L'amélioration relative du traitement par rapport au contrôle

C) La taille d'échantillon

D) La durée du test

Réponse: B) L'amélioration relative du traitement par rapport au contrôle

Lift = (Taux_B - Taux_A) / Taux_A × 100%. Si A=5% et B=6%, lift = (6-5)/5 = 20%.

Q13. Un A/B test sur les frais de service montre un lift de -3% (B pire que A) avec p = 0.04. Quelle décision?

A) Déployer B quand même

B) Ne pas déployer B - le changement est significativement négatif

C) Continuer le test indéfiniment

D) Le résultat n'est pas interprétable

Réponse: B) Ne pas déployer B - le changement est significativement négatif

$p < 0.05$ avec un lift négatif signifie que B est significativement PIRE que A. On conserve la version actuelle.

Q14. La stratification dans un A/B test signifie:

A) Supprimer certains clients

B) Randomiser au sein de sous-groupes (segments) pour assurer l'équilibre

C) Analyser après coup par segment

D) Augmenter la durée du test

Réponse: B) Randomiser au sein de sous-groupes (segments) pour assurer l'équilibre

La stratification assure que chaque segment (Retail, Premium, etc.) est équitablement réparti entre A et B, améliorant la précision.

Q15. Avant de lancer un A/B test, vous effectuez une vérification "A/A". Quel est l'objectif?

A) Tester la nouvelle version deux fois

B) Vérifier que le système de randomisation fonctionne (pas de différence entre groupes identiques)

C) Doubler la taille d'échantillon

D) Comparer deux anciennes versions

Réponse: B) Vérifier que le système de randomisation fonctionne (pas de différence entre groupes identiques)

Un test A/A expose deux groupes à la même version. Si une différence significative apparaît, il y a un problème de randomisation ou de mesure.

Q16. Quelle métrique est appropriée comme métrique primaire pour un test d'email marketing crédit?

A) Taux d'ouverture de l'email

B) Taux de souscription au crédit (conversion finale)

C) Nombre d'emails envoyés

D) Durée de lecture de l'email

Réponse: B) Taux de souscription au crédit (conversion finale)

La métrique primaire doit être alignée avec l'objectif business. Ici, l'objectif est la souscription crédit, pas l'ouverture (qui est une métrique secondaire/intermédiaire).

Q17. Pourquoi définir les métriques AVANT de lancer le test?

A) Pour accélérer le test

B) Pour éviter le "cherry-picking" de métriques favorables après coup

C) C'est une exigence technique

D) Les métriques ne peuvent pas changer

Réponse: B) Pour éviter le "cherry-picking" de métriques favorables après coup

Définir les métriques à l'avance évite le biais de "trouver" une métrique significative parmi plusieurs testées (problème de comparaisons multiples implicite).

Q18. Un effet de "novelty" (nouveauté) dans un A/B test signifie:

A) Le test est trop court

B) L'effet initial peut être gonflé car les utilisateurs réagissent à la nouveauté, pas à la valeur réelle

C) La version B est toujours meilleure

D) Il n'y a pas assez de données

Réponse: B) L'effet initial peut être gonflé car les utilisateurs réagissent à la nouveauté, pas à la valeur réelle

L'effet de nouveauté peut biaiser les résultats. Un nouveau design peut attirer l'attention temporairement. L'effet réel à long terme peut être différent.

Q19. Comment le biais de survivant peut-il affecter un A/B test sur le churn?

A) Aucun effet

B) Si les clients qui churnt disparaissent du test, on ne mesure que les survivants (biais)

C) Le test dure trop longtemps

D) La randomisation échoue

Réponse: B) Si les clients qui churnt disparaissent du test, on ne mesure que les survivants (biais)

Pour mesurer le churn, on doit suivre tous les clients initialement assignés, y compris ceux qui partent. Ne mesurer que les restants biaise les résultats.

Q20. UniBank teste 3 versions d'email (A, B, C). C'est un:

A) A/B test standard

B) Test A/B/n ou multivarié

C) Test invalide

D) Test de corrélation

Réponse: B) Test A/B/n ou multivarié

Avec plus de 2 variantes, on parle de test A/B/n ou MVT (multivariate test). L'analyse nécessite des corrections pour comparaisons multiples.

Q21. Pour un test A/B/C avec 3 variantes, combien de comparaisons par paires sont possibles?

A) 2

B) 3 (A-B, A-C, B-C)

C) 6

D) 9

Réponse: B) 3 (A-B, A-C, B-C)

Nombre de paires = k(k-1)/2 = 3×2/2 = 3. Attention au problème de comparaisons multiples (inflation erreur Type I).

Q22. Quelle correction appliquer pour 3 comparaisons avec α global = 0.05?

A) Utiliser $\alpha = 0.05$ pour chaque

B) Bonferroni: $\alpha = 0.05/3 \approx 0.017$ par comparaison

C) Utiliser $\alpha = 0.15$

D) Aucune correction nécessaire

Réponse: B) Bonferroni: $\alpha = 0.05/3 \approx 0.017$ par comparaison

Bonferroni maintient l'erreur familywise à 5% en divisant α par le nombre de comparaisons. Chaque test doit avoir $p < 0.017$.

Q23. L'intervalle de confiance à 95% sur la différence de taux (B-A) est [0.5%, 2.5%]. Que concluez-vous?

A) B n'est pas significativement différent de A

B) B est significativement meilleur que A (l'IC ne contient pas 0)

C) Le test est invalide

D) Besoin de plus de données

Réponse: B) B est significativement meilleur que A (l'IC ne contient pas 0)

L'IC [0.5%, 2.5%] est entièrement positif, signifiant que B est entre 0.5 et 2.5 points de pourcentage meilleur que A avec 95% de confiance.

Q24. Un résultat est "statistiquement significatif" mais le lift n'est que de 0.1%. Quelle est la bonne interprétation?

A) Déployer immédiatement

B) La différence est réelle mais peut-être pas pratiquement importante (significativité statistique ≠ importance business)

C) Le test a échoué

D) Le résultat est faux

Réponse: B) La différence est réelle mais peut-être pas pratiquement importante (significativité statistique ≠ importance business)

Avec de grands échantillons, même de très petites différences deviennent significatives. Il faut évaluer si 0.1% justifie le coût du changement.

Q25. Quel outil statistique calculer en plus du test de significativité pour évaluer l'impact business?

A) Plus de p-values

B) L'intervalle de confiance sur l'effet et l'impact projeté en valeur (ex: HTG de revenus supplémentaires)

C) La taille d'échantillon

D) Uniquement le lift

Réponse: B) L'intervalle de confiance sur l'effet et l'impact projeté en valeur (ex: HTG de revenus supplémentaires)

Au-delà de la significativité, traduire l'effet en impact business (ex: "+1% de conversion × 100,000 clients × 50,000 HTG = X millions HTG/an") aide la décision.

Résumé des Concepts Clés

Terminologie

Terme	Définition
Baseline	Taux de performance actuel (contrôle)
MDE	Effet minimal qu'on veut détecter
Lift	Amélioration relative (%)
Puissance	P(détecter un vrai effet) = 80% standard
α	P(faux positif) = 5% standard

Étapes d'un A/B Test

1. **Hypothèse:** "B augmentera le taux de X%"
2. **Design:** Calculer taille, durée, métriques
3. **Randomisation:** Assigner clients aléatoirement
4. **Exécution:** Collecter données sans peeking
5. **Analyse:** Test statistique + décision

Erreurs à Éviter

Erreur	Conséquence
Peeking	Inflation faux positifs
Durée trop courte	Biais jour de semaine
Pas de stratification	Groupes déséquilibrés
Cherry-picking métriques	Conclusions biaisées

Score: ___/25