UNIVERSITÉ D'ÉBOLOWA — ESTLC
Filière GSI — 3e Année

CONTRÔLE DE CONNAISSANCES

Introduction à la Science des Données

Document préparatoire — 20 Questions QCM + 5 Exercices avec corrigés

PARTIE I — QUESTIONNAIRE À CHOIX MULTIPLES (20 questions)

Consigne : Pour chaque question, entourez la lettre correspondant à la bonne réponse. La réponse correcte est indiquée en vert après chaque question.

Question 1 — Quelle est la définition correcte de la statistique ?

A. Un langage de programmation dédié à l'analyse des données
✔ B. La discipline qui étudie des phénomènes via la collecte, le traitement, l'analyse et l'interprétation de données
C. Un outil exclusivement réservé aux mathématiques pures
D. Une technique de stockage de bases de données massives
■ Justification : La statistique est définie comme la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation.

Question 2 — Le terme « statisticum » apparaît pour la première fois :

A. Au XV° siècle
B. Au XIX° siècle
✔ C. À la fin du XVII° siècle
D. Au XX° siècle
■ Justification : Le terme statisticum apparaît à la fin du XVIIe siècle, issu du latin status (état).

Question 3 — Pourquoi l'analyse des données est-elle nécessaire face à un grand nombre de caractéristiques ?

A. Parce que les ordinateurs ne peuvent pas stocker trop de données
✔ B. Parce qu'il faut réduire l'espace multidimensionnel en un espace lisible à 2 ou 3 dimensions
C. Parce que la loi l'exige dans certains pays
D. Parce que les données qualitatives sont trop nombreuses
■ Justification : L'analyse des données permet de réduire l'espace multidimensionnel en un espace à peu de dimensions (2 ou 3) où l'information devient lisible, tout en conservant une part importante de l'information d'origine.

Question 4 — Quelles sont les deux familles de techniques de l'analyse des données ?

A. Techniques descriptives et techniques prescriptives
B. Techniques quantitatives et techniques qualitatives
✔ C. Techniques d'analyses factorielles et techniques de classification automatique
D. Techniques de sondage et techniques de recensement
■ Justification : L'analyse des données rassemble les techniques d'analyses factorielles (relations entre variables) et les techniques de classification automatique (regroupement homogène des données).

Question 5 — Qui a développé l'analyse factorielle en composantes principales en 1933 ?

A. Jean-Paul Benzécri
B. Charles Spearman
C. Cyril Burt
✔ D. Harold Hotelling
■ Justification : En 1933, Harold Hotelling développe l'analyse factorielle en composantes principales (ACP), où les individus sont considérés comme des vecteurs à plusieurs dimensions.

Question 6 — Qu'est-ce qu'un recensement ?

A. Une étude réalisée sur un échantillon de la population
B. Un sondage téléphonique
✔ C. Une étude réalisée sur toutes les unités statistiques d'une population
D. Une méthode de collecte par questionnaire en ligne
■ Justification : Le recensement est une étude réalisée sur TOUTES les unités statistiques d'une population, contrairement au sondage qui porte sur un échantillon.

Question 7 — Dans un sondage sur 2 000 électeurs de Yaoundé, qu'est-ce que l'échantillon ?

A. Tous les électeurs de Yaoundé
✔ B. Les 2 000 électeurs choisis au hasard
C. L'intention de vote
D. Un électeur
■ Justification : L'échantillon correspond aux 2 000 électeurs choisis au hasard parmi l'ensemble des électeurs de Yaoundé (la population).

Question 8 — Quelle est la variable qui correspond à « les diplômes (CEP, BEPC, Bac, Licence...) » ?

A. Variable quantitative discrète
B. Variable qualitative nominale
✔ C. Variable qualitative ordinale
D. Variable quantitative continue
■ Justification : Les diplômes sont une variable qualitative ordinale car leurs modalités peuvent être ordonnées (du moins élevé au plus élevé).

Question 9 — Laquelle de ces variables est quantitative continue ?

A. Le nombre d'enfants par foyer
B. L'âge exprimé en années entières
✔ C. La taille d'une personne (en cm)
D. Le genre (homme / femme)
■ Justification : La taille est une variable quantitative continue car son domaine est continu (peut prendre toute valeur réelle dans un intervalle). L'âge en années entières est discret, le genre est qualitatif nominal, et le nombre d'enfants est discret.

Question 10 — Quelle est la première étape de la démarche scientifique face à un problème ?

A. Collecter et prétraiter les données
B. Appliquer des techniques statistiques
C. Interpréter les résultats
✔ D. Formuler le problème et établir des hypothèses
■ Justification : Selon le cours, la démarche scientifique commence toujours par : 1. Formuler le problème et établir des hypothèses, avant de choisir et collecter les données.

Question 11 — Quelles sont les trois catégories principales de sources de données ?

A. Données publiques, privées et secrètes

✔ B. Données écrites, saisies et générées automatiquement

C. Données quantitatives, qualitatives et mixtes

D. Données primaires, secondaires et tertiaires

■ Justification : Le cours distingue trois catégories : données écrites (manuscrites), données saisies (ordinateur/tablette/téléphone) et données générées automatiquement par des processus informatisés.

Question 12 — L'analyse des données a connu un grand essor grâce à :

A. L'invention de l'imprimerie au XVe siècle

✔ B. L'expansion de l'informatique dans les années 1960-1970

C. La Révolution française en 1789

D. L'invention d'Internet dans les années 1990

■ Justification : L'analyse des données a connu un grand essor du fait de l'expansion de l'informatique dans les années 1960-1970, permettant les calculs sur de grands volumes de données.

Question 13 — Qui est le précurseur ayant introduit le concept de « facteur » en analyse factorielle en 1904 ?

A. Harold Hotelling

B. Jean-Paul Benzécri

✔ C. Charles Spearman

D. Louis Leon Thurstone

■ Justification : En 1904, Charles Spearman introduit le concept de facteur, en cherchant une variable explicative cachée derrière les notes obtenues par de nombreux sujets à de nombreux tests.

Question 14 — Selon le cours, à partir de quel niveau de diplôme peut-on accéder aux métiers de la Data Science ?

A. Baccalauréat

B. BTS

C. Licence

✔ D. Master en mathématiques/informatique ou analyse des données

■ Justification : Le cours précise qu'il faut avoir au moins un Master en mathématique/informatique ou en analyse des données/mathématiques appliquées pour accéder à la profession.

Question 15 — Dans l'exemple de l'entreprise de commerce, que représente le fait que « moins de 10% des clients constituent plus de 70% des bénéfices » ?

A. Une erreur dans les données

✔ B. La loi de Pareto (règle 80/20) appliquée à la clientèle

C. Un problème de sondage mal réalisé

D. Un résultat impossible à observer en pratique

■ Justification : Ce résultat illustre la loi de Pareto (règle 80/20) : une minorité de clients génère la majorité des bénéfices, information clé pour orienter les décisions commerciales.

Question 16 — Quel outil Python est cité dans le cours parmi les technologies maîtrisées par un Data Scientist ?

A. JavaScript

B. PHP

✔ C. Python

D. Swift

■ **Justification : Le cours cite Python, R, Hadoop et SQL comme technologies indispensables à maîtriser pour les métiers de la Data Science.**

Question 17 — L'accessibilité aux données peut se faire de deux manières. Lesquelles ?

A. Données en ligne et données hors ligne

✔ **B. Données centralisées et données distribuées**

C. Données nationales et données internationales

D. Données structurées et données non structurées

■ **Justification : Selon le cours, les données peuvent être centralisées (dans une base de données unique) ou distribuées (dans plusieurs unités de stockage nécessitant un mécanisme de regroupement avant analyse).**

Question 18 — Quelle variable est qualitative NOMINALE parmi les suivantes ?

A. Les mois de l'année (Janvier, Février, ...)

B. Les diplômes (CEP, BEPC, Bac, Licence...)

✔ **C. Le genre (homme, femme)**

D. La note d'un étudiant sur 20

■ **Justification : Le genre est une variable qualitative nominale car ses modalités (homme, femme) ne peuvent pas être ordonnées. Les mois et diplômes sont ordinaux ; la note est quantitative.**

Question 19 — L'analyse factorielle des correspondances a été introduite en 1962 par :

A. Harold Hotelling

B. Charles Spearman

C. Cyril Burt

✔ **D. Jean-Paul Benzécri**

■ **Justification : En 1962, Jean-Paul Benzécri introduit l'analyse factorielle des correspondances, qui fournit sans hypothèses a priori des représentations simplifiées faciles à interpréter.**

Question 20 — Laquelle de ces affirmations sur le métier de Data Scientist est correcte selon le cours ?

A. C'est un métier en déclin depuis 2010

B. Il est limité au domaine de la finance

✔ **C. Il a été qualifié de « métier le plus sexy du 21e siècle » par Harvard Business Review en 2012**

D. Il ne nécessite aucune compétence en programmation

■ **Justification : Le cours cite Harvard Business Review (2012) qui a qualifié le Data Scientist de « métier le plus sexy du 21e siècle », soulignant l'épanouissement et les opportunités nombreuses dans ce domaine.**

PARTIE II — EXERCICES APPLIQUÉS (5 exercices)

Consigne : Lisez attentivement chaque énoncé et répondez aux questions. Les éléments de correction sont indiqués en vert.

Exercice 1 — Terminologie statistique

Énoncé :
Un chercheur veut étudier le niveau de satisfaction des étudiants de l'ESTLC vis-à-vis des enseignements dispensés. Il interroge 150 étudiants tirés au sort parmi les 600 inscrits. Il recueille pour chaque étudiant : son âge, sa filière (GSI, GSA, TRL), son niveau de satisfaction (Très satisfait, Satisfait, Peu satisfait, Insatisfait), sa moyenne générale et son sexe.

Questions :

1. Identifiez la population et l'unité statistique de cette étude.

✔ Population : l'ensemble des 600 étudiants inscrits à l'ESTLC. Unité statistique : un étudiant.

2. S'agit-il d'un recensement ou d'un sondage ? Justifiez.

✔ C'est un sondage, car l'étude porte sur un sous-ensemble (150 étudiants) de la population totale (600 étudiants), et non sur tous les étudiants.

3. Identifiez et classifiez chacune des 5 variables relevées (type et sous-type).

✔ • Âge → Variable quantitative discrète (valeurs entières dénombrables)

✔ • Filière (GSI, GSA, TRL) → Variable qualitative nominale (catégories sans ordre)

✔ • Satisfaction (Très satisfait ... Insatisfait) → Variable qualitative ordinale (modalités ordonnables)

✔ • Moyenne générale → Variable quantitative continue (valeur réelle dans [0;20])

✔ • Sexe → Variable qualitative nominale (catégories sans ordre)

4. Peut-on généraliser les résultats obtenus à l'ensemble des 600 étudiants ? Pourquoi ?

✔ Oui, car l'échantillon a été constitué par tirage au sort (aléatoire). Un échantillon aléatoire permet de généraliser les résultats à l'ensemble de la population avec un certain niveau de confiance.

Exercice 2 — Démarche scientifique

Énoncé :
Un hôpital de district remarque une augmentation des cas de paludisme dans sa zone de desserte. Le responsable médical souhaite mener une étude statistique pour comprendre et réduire ce phénomène.

Questions :

1. Formulez un problème précis et proposez deux hypothèses de recherche.

✔ Problème : Quels sont les facteurs qui expliquent l'augmentation des cas de paludisme dans la zone de desserte de l'hôpital ?

✔ Hypothèses : H1 : La proximité de points d'eau stagnante favorise l'apparition des cas. H2 : Le non-usage des moustiquaires imprégnées est corrélé à un risque plus élevé d'infection.

2. Quelles données faudrait-il collecter ? Citez au moins 5 variables pertinentes et précisez leur type.

✔ Variables à collecter :

✔ • Âge du patient → quantitative discrète

✔ • Sexe → qualitative nominale

✔ • Distance domicile-point d'eau → quantitative continue

✔ • Utilisation de moustiquaire (Oui/Non) → qualitative nominale

✔ • Nombre de cas dans le foyer → quantitative discrète

✔ • Zone géographique (quartier) → qualitative nominale

3. Décrivez les 5 étapes complètes de la démarche scientifique appliquées à ce cas.

✔ Étape 1 : Formuler le problème → identifier les facteurs de risque du paludisme.

✔ Étape 2 : Choisir les données → sélectionner les variables pertinentes listées ci-dessus.

✔ Étape 3 : Collecter et prétraiter les données → enquêtes, nettoyage des données manquantes.

✔ Étape 4 : Appliquer les techniques statistiques → tableaux croisés, graphiques, corrélations.

✔ Étape 5 : Interpréter et recommander → tirer des conclusions et proposer des actions.

4. Donnez deux exemples de recommandations découlant des résultats possibles.

✔ Recommandation 1 : Si les résultats montrent que les zones proches des points d'eau stagnante concentrent 80% des cas → mener des campagnes d'assainissement ciblées dans ces zones.

✔ Recommandation 2 : Si la non-utilisation des moustiquaires est fortement corrélée aux cas → intensifier la distribution et la sensibilisation à l'usage des moustiquaires imprégnées.

Exercice 3 — Classification des variables

Énoncé :

Pour chacune des variables suivantes, indiquez : (a) son type (qualitative ou quantitative), (b) son sous-type (nominale, ordinale, discrète ou continue), (c) un exemple de modalité ou de valeur possible.

Questions :

1. La langue parlée par un étudiant (français, anglais, espagnol...)

✔ Qualitative nominale — les langues ne peuvent pas être ordonnées. Modalité : français.

2. Le nombre de cours suivis par semaine

✔ Quantitative discrète — le nombre de cours est entier et dénombrable. Valeur : 5 cours/semaine.

3. Le niveau d'étude (Licence 1, Licence 2, Licence 3, Master 1, Master 2)

✔ Qualitative ordinale — les niveaux d'étude sont ordonnés du plus bas au plus élevé. Modalité : Licence 2.

4. Le poids corporel d'un individu (en kg)

✔ Quantitative continue — le poids peut prendre toute valeur réelle dans un intervalle. Valeur : 68,5 kg.

5. La couleur préférée d'un individu

✔ Qualitative nominale — les couleurs ne peuvent pas être ordonnées. Modalité : bleu.

6. La température corporelle d'un patient (en °C)

✔ Quantitative continue — la température peut prendre toute valeur réelle dans un intervalle. Valeur : 37,8°C.

Exercice 4 — Application à un cas réel

Énoncé :

Une société de microfinance camerounaise dispose d'un historique de 5 000 dossiers de prêts accordés sur 3 ans. Elle souhaite réduire ses risques d'impayés en identifiant le profil des clients qui remboursent et de ceux qui ne remboursent pas.

Questions :

1. Dans quel domaine d'application de l'analyse des données ce cas s'inscrit-il ? Justifiez.

✔ Ce cas s'inscrit dans le domaine de la Microfinance. Le cours cite explicitement : « pour évaluer les risques et définir les populations emprunteuses (critères de ceux qui remboursent leurs prêts) ».

2. Quelle est la population ? L'unité statistique ? S'agit-il d'un recensement ou d'un sondage ?

✔ Population : l'ensemble des 5 000 dossiers de prêts. Unité statistique : un dossier de prêt (ou un client). Il s'agit d'un recensement car on dispose de TOUS les dossiers des 3 dernières années.

3. Proposez 4 variables pertinentes à analyser et précisez leur type.

✔ • Montant du prêt (en FCFA) → quantitative continue

✔ • Durée du prêt (en mois) → quantitative discrète

✔ • Statut de remboursement (Remboursé / En cours / Impayé) → qualitative nominale

✔ • Secteur d'activité du client (commerce, agriculture, services...) → qualitative nominale

4. Quelle famille de techniques d'analyse des données est la plus adaptée pour regrouper les clients en profils homogènes ? Justifiez.

✔ La technique de classification automatique est la plus adaptée, car elle permet de regrouper les clients en classes homogènes selon leurs comportements (ceux qui remboursent vs ceux qui ne remboursent pas), afin d'établir des typologies et d'aider à la décision.

Exercice 5 — Analyse et recommandations

Énoncé :

Un supermarché de Yaoundé veut optimiser ses bénéfices. Son analyse de données révèle que : 72% des produits ont un poids quasi identique de 1,2% sur le chiffre d'affaires ; 8% des produits génèrent 75% des bénéfices totaux ; les produits (riz, huile, tomates) sont très souvent achetés ensemble ; deux groupes de clients aux comportements d'achat similaires ont été identifiés : « ménages » (achats hebdomadaires massifs) et « célibataires » (achats quotidiens petites quantités).

Questions :

1. Décrivez les étapes de la démarche scientifique qui ont conduit à ces résultats.

✔ Étape 1 : Formuler le problème → le supermarché veut augmenter ses bénéfices.

✔ Étape 2 : Choisir les données → historique des ventes, achats par client, marges par produit.

✔ Étape 3 : Collecter et prétraiter → regrouper les données de caisse, nettoyer les anomalies.

✔ Étape 4 : Appliquer les techniques → calcul des poids sur CA, analyse des corrélations entre produits, classification des clients.

✔ Étape 5 : Interpréter → identifier les produits clés, les associations et les segments de clientèle.

2. Formulez une recommandation concrète pour les 8% de produits qui génèrent 75% des bénéfices.

✔ Recommandation : Concentrer les efforts sur ces 8% de produits : augmenter les stocks pour éviter les ruptures, intensifier les promotions, les placer en tête de gondole, et négocier de meilleurs prix d'achat auprès des fournisseurs pour maximiser la marge.

3. Que peut faire le supermarché avec les 3 produits souvent achetés ensemble (riz, huile, tomates) ? Donnez deux stratégies opposées et justifiez chacune.

✔ Stratégie 1 (faciliter la vente) : Créer des bouquets/offres groupées « riz + huile + tomates » à prix réduit et les rapprocher dans les rayons pour encourager l'achat simultané et augmenter le panier moyen.

✔ Stratégie 2 (promouvoir d'autres produits) : Les éloigner dans les rayons pour forcer le client à traverser tout le magasin et découvrir d'autres produits, augmentant ainsi les achats non planifiés.

4. Comment exploiter les deux groupes de clients identifiés pour augmenter les ventes ?

✔ Pour les ménages : proposer des formats familiaux, des cartes de fidélité avec réductions sur les achats massifs hebdomadaires, et envoyer des offres promotionnelles en début de semaine.

✔ Pour les célibataires : proposer des petits formats, des offres « produit du jour », et des promotions flash sur les produits de première nécessité en petites quantités.
