Résumé des 3 exposés déjà faits en classe

Mame Balla BOUSSO, élève en ISEP2_2024

2024-04-27

Résumé de ma compréhension des différents exposés

Exposé 1 : traitement statistique des valeurs manquantes et aberrantes avec R

Sommaire Les données statistiques sont souvent affectées par la présence de valeurs aberrantes et manquantes, qui peuvent compromettre la qualité des analyses. Les valeurs aberrantes peuvent déformer les résultats en biaisant les estimations, tandis que les valeurs manquantes réduisent la puissance statistique des tests et la précision des estimations. L'objectif principal de cette étude est de discuter des différentes méthodes pour identifier et traiter les valeurs aberrantes et manquantes, en s'adaptant au contexte et aux objectifs spécifiques de chaque analyse. I. Traitement des valeurs manquantes Types de valeurs manquantes : MCAR (Manquantes Complètement Aléatoires) : les données manquent sans lien avec d'autres variables. MAR (Manquantes Aléatoires Conditionnelles) : les données manquent en relation avec d'autres variables observées mais pas les valeurs manquantes elles-mêmes. MNAR (Manquantes Non Aléatoires) : les données manquent en relation avec des variables non observées, rendant leur traitement plus complexe. Méthodes de traitement : Suppression : Élimination des observations incomplètes, simple mais peut entraîner une perte significative de données. Imputation par la moyenne : Remplacement des valeurs manquantes par la moyenne des observations disponibles; adapté pour les données où la variabilité est faible. Imputation par la médiane : Utilisée lorsque les données sont asymétriques; remplace les valeurs manquantes par la médiane pour une robustesse accrue. Imputation par régression : Plus sophistiquée, utilise un modèle de régression pour prédire les valeurs manquantes à partir d'autres variables. Imputation par HOT DECK : Remplace les valeurs manquantes par des valeurs similaires issues d'autres observations dans le dataset. Imputation par KNN (k-nearest neighbors) : Impute les valeurs manquantes en se basant sur les moyennes de k observations les plus proches. LOCF (Last Observation Carried Forward) : Utilisée dans les séries chronologiques pour propager la dernière observation disponible. II. Traitement des valeurs aberrantes Détection des valeurs aberrantes : Détection par domaine de valeurs : Identifie les aberrations en se basant sur les intervalles attendus pour certaines variables. Détection graphique : Utilisation de graphiques comme les boxplots et les nuages de points pour visualiser et identifier les anomalies. Méthode de Tukey (IQR) et Score Z : Détectent les valeurs extrêmes en utilisant des mesures statistiques de dispersion. Méthodes de traitement : Suppression : Élimination des valeurs aberrantes, adaptée pour les erreurs de saisie ou les données avec faible effectif. Transformation : Techniques comme la transformation logarithmique pour minimiser l'influence des extrêmes, ou la winsorisation qui limite les valeurs à un certain percentile pour conserver plus de données. Limitations Le traitement des valeurs aberrantes et manquantes peut introduire des biais et compliquer l'interprétation des résultats. La sélection de la méthode de traitement dépend étroitement de la structure des données et des objectifs de l'analyse, ce qui peut accroître la complexité de l'approche statistique.

Exposé 2: STATISTIQUES DESCRIPTIVES ET VISUALISATION DES VARIABLES CATEGORIELLES AVEC GGPLOT2

La grammaire graphique, telle que mise en œuvre dans ggplot2, fournit un cadre cohérent pour construire des visualisations informatives et esthétiquement plaisantes. Elle se compose de composants de base qui peuvent être combinés pour créer une grande variété de graphiques : Couches (layers) : Les graphiques ggplot2 sont construits à partir de couches. Chaque couche représente un aspect différent de la visualisation, tel que les points de données, les lignes de tendance, les barres d'erreur, etc. L'ajout de couches permet de

superposer différents éléments visuels pour enrichir la représentation graphique. • aes (Aesthetic mappings) : Les aes spécifient comment les variables sont associées aux propriétés visuelles des éléments graphiques. Dans ggplot2, les aes sont définis à l'intérieur de la fonction aes() et peuvent inclure des variables pour contrôler des éléments tels que la taille, la forme, la couleur, le remplissage et la transparence. Dans aes, on a : _Size (taille) : Permet de définir la taille des éléments graphiques, tels que les points ou les lignes, en fonction des valeurs d'une variable. _Shape (forme) : Détermine la forme des éléments graphiques, comme les points, en fonction des niveaux d'une variable. _Color (couleur) : Contrôle la couleur des éléments graphiques en fonction des valeurs d'une variable. _Fill (remplissage) : Utilisé pour remplir les éléments graphiques, tels que les barres dans un histogramme, en fonction des niveaux d'une variable.

Exposé 3: ANOVA et tests non paramétriques

Sommaire L'Analyse de Variance (ANOVA) et les tests non paramétriques sont des outils cruciaux pour l'analyse de données dans divers domaines. L'ANOVA est utilisée pour comparer les moyennes de trois groupes ou plus, tandis que les tests non paramétriques sont utilisés lorsque les données ne satisfont pas aux hypothèses de normalité requises par l'ANOVA.

I. ANOVA Développement et Application : Développée par Ronald Fisher au début du 20ème siècle, l'ANOVA permet de détecter les différences ou similitudes dans une population étudiée, en se concentrant sur l'interaction entre les facteurs de variabilité et une variable quantitative principale. Exemple Pratique : Un exemple est donné avec des scores attribués par différents correcteurs lors des olympiades de mathématiques de ENSAE, analysant l'influence du correcteur et du sexe des candidats sur les notes. Limites : L'ANOVA suppose l'égalité des variances et une distribution normale des données, et elle est sensible aux valeurs aberrantes, nécessitant souvent des tests post-hoc pour identifier les groupes spécifiquement différents.

II. Tests Non Paramétriques

Définition et Utilité : Contrairement à l'ANOVA, les tests non paramétriques ne nécessitent pas d'hypothèse sur la distribution des données. Ils sont utilisés pour des données qui ne suivent pas une distribution spécifique, en utilisant des statistiques d'ordre ou des tableaux de contingence.

Exemples de Tests : Test de Wilcoxon : Compare les distributions de jeux de variables appariées en prenant en compte les signes et amplitudes des différences. Test du khi-carré : Évalue si une distribution de fréquences observées diffère de celle attendue, souvent utilisé pour des données catégorielles. Test U de Mann-Whitney : Compare les rangs de deux groupes pour tester s'ils proviennent de la même population. Test de Kruskal-Wallis : Extension du test U de Mann-Whitney pour plus de deux groupes, détectant des différences dans la distribution des échantillons. III. Applications Pratiques Des liens pour ouvrir des fichiers R d'application de l'ANOVA et des tests non paramétriques sont mentionnés, permettant une mise en pratique directe des méthodes étudiées. Ce résumé enrichi offre une compréhension, dans l'ensemble, de l'ANOVA et des tests non paramétriques, soulignant leur pertinence et leurs applications dans le traitement et l'analyse statistique des données.