TAMSIR NDONG ISEP2

EXERCICE R : RESUME DES EXPOSES

**ANOVA ET TESTS NON PARAMETRIQUES**

L’Analyse de la Variance (ANOVA) et les tests non paramétriques sont des outils statistiques essentiels pour l’analyse de données. L’ANOVA est couramment utilisée pour comparer les moyennes de trois groupes ou plus tandis que les tests non paramétriques offrent une alternative robuste lorsque les données ne répondent pas aux hypothèses de normalité requises par l’ANOVA.

**ANOVA**

Le groupe qui a exposé a pris l’exemple des olympiades avec comme correcteur Tamsir, Awa et Ameth et ils se sont intéressés aux variables : note, correcteur, sexe du candidat. Ainsi ils ont procédé à un test pour étudier ces différentes relations : lien la note d’un candidat et son sexe, lien entre la note d’un candidat et son correcteur. Dans le premier cas, ils ont formé 2 groupes et dans le second 6 groupes en fonction du correcteur et du sexe. Dans chaque cas ils émettent 2 hypothéses H0 et H1.Dans les 2 cas  :

H0 : les moyennes des différents groupes sont identiques

H1 : les moyennes des différents groupes ne sont pas identiques

Pour cela ils suivent 6 étapes  à savoir : Hypothéses, Calcul des moyennes empiriques des différents groupes et de la moyenne totale, somme des carrés des écarts intra classe, somme des carrés des écarts inter classe, Calcul de la statistique T, Décision .

Ainsi les conclusions tirées à l’issue de cette étude sont : Autrement dit le sexe d’un candidat n’influence pas sa note. Cela voudrait dire qu’il n’y a pas de correcteur plus sévère ou au contraire plus gentil que les autres.

TESTS NON PARAMETRIQUES

Un test non paramétrique est un test ne nécessitant pas d’hypothèse sur la loi des données. Les données sont alors remplacées par des statistiques ne dépendant pas des moyennes/variances des données initiales (tableau de contingence, statistique d’ordre comme les rangs. . . ).Il y a le test de : Test de rang signé de Wilcoxon, Test du khi-carré pour la qualité d’ajustement, Test U de Mann-Whitney(Wilcoxon/Mann-Whitney), Test de Kruskal-Wallis).

TRAITEMENT DES VALEURS MANQUANTES ET DES VALEURS ABERRANTES AVEC R

La présence de valeurs aberrantes et manquantes dans les données est un problème fréquent qui peut affecter la qualité des analyses statistiques. Les valeurs aberrantes peuvent fausser les résultats et les conclusions, tandis que les valeurs manquantes peuvent réduire la puissance statistique et la précision des estimations. Il y a 3 types de valeurs manquantes : **MCAR** (Donnée Manquante Complètement Aléatoire) ; **MAR** (Donnée Manquante Aléatoire Conditionnelle) ; **MNAR** (Donnée Manquante Non Aléatoire At Random) et 2 catégories de méthodes (Suppression des observations incomplètes ; Imputation). La première méthode peut causer une perte énorme de données.

**La méthode de suppression** ne doit être utiliser que pour des certaines situations : nombre de données faible, manque d’informtions sur les données manquantes, analyse trop robuste pour les données manquantes.

Il existe plusieurs méthodes d’imputation :

**Méthode d’imputation par la moyenne** : Elle consiste à remplacer les valeurs manquantes par la moyenne de la variable concernée.

**Méthode d’imputation par la médiane**  : Elle consiste à remplacer les valeurs manquantes par la médiane de la variable concernée.

**Méthode d’imputation par régression :** Elle se déroule en 3 étapes :selection des variables explicatives pertinentes ; Construction d’un modèle de regression; Prédiction des valeurs manquantes.

**Méthode d’imputation par HOT DECK** : Elle se déroule en 3 étapes : Définition des groupes d’imputation ; Sélection d’un donneur ; Imputation de la valeur manquante.

**Méthode d’imputation par KNN** : Elle fonctionne en identifiant les K observations les plus proches (en termes de distance euclidienne) de l’observation avec la valeur manquante et en utilisant la moyenne des valeurs KNN pour imputer la valeur manquante. Elle se déroule en 3 étapes : Définition des groupes d’imputation ; Sélection d’un donneur ; Imputation de la valeur manquante.

**Méthode d’imputation par LOCF :** LOCF, ou “Last Observation Carried Forward” (la dernière observation reportée), est une méthode qui consiste à remplir les valeurs manquantes par la dernière observation disponible dans la série chronologique.

**Détection des valeurs aberrantes**

Elle se fait suivant 3 possibilités : Détection suivant le domaine des valeurs ; Détection graphique ; Autres méthodes statistiques de détection.

**Les méthodes d’imputation** généralement utilisés sont : la winsorisation; le box-plot ; la transformation algorithmique.

**Statistiques descriptives et visualisation des variables catégorielles avec ggplot2()**

Les variables en général. . .

**Quantitatives**: discrètes et continues

**Qualitatives**: nominales et ordinales Les variables sur R. . .

Sous forme de vecteurs double, float, character

Les factors: En R, un facteur (ou factor en anglais) est un vecteur contenant uniquement certaines valeurs prédéfinies. Les valeurs pré-définies sont appelées des niveaux.

1. **La composante eesthétique (aes)**
2. **Ajout d’élèments de Base (titre)**

Pour que les noms de nos variables correspondent au nom de nos axes, il nous suffit d’utiliser la commande aes (x=variable1,y=Variable2) dans le cas de graphiques bivariés. On peut utiliser la commande labs.

1. **Le mappage color**

Dans ggplot2, le mappage color (ou colour) de la composante aes est utilisé pour attribuer une couleur différente à chaque niveau (modalité) d’une variable catégorielle qui est la variable de mappage.

1. **Le mappage shape**

Dans ggplot2, le mappage shape est utilisé pour attribuer une forme différente à chaque niveau d’une variable catégorielle. Ainsi, utiliser shape = Reg\_rec fera que chaque région sera représentée par une forme de point différente sur le graphique.

1. **Le mappage size**

L’esthétique size contrôle la taille des points. Elle détermine la taille physique des points sur le graphique.

1. **Le mappage alpha**

L’esthétique alpha contrôle la transparence des points. Elle va de 0 à 1, où 0 signifie complètement transparent (invisible) et 1 signifie complètement opaque (visible). En utilisant alpha, nous pouvons rendre les points plus ou moins transparents en fonction de la valeur de la variable associée.

1. **La composante theme**

La composante theme est utilisée pour contrôler l’apparence de différents éléments dans notre graphique tels que les axes, les légendes, les titres, les marges, etc. Elle comprend 4 parties les elements de theme, les fonctions d’élèment, la foncction theme et les thèmes complets.

1. **Les thèmes complets**

Ce sont des thèmes pré-définis dans R qui définissent tous les éléments du thème pour une apparence cohérente. Ces thèmes offrent une solution rapide pour obtenir un aspect visuel uniforme pour tout le graphique.

1. **Les élèments de thème**

Les éléments de thème permettent de contrôler les différents composants visuels d’un graphique, tels que les titres, les axes, les légendes, les graduations, etc. Ils sont regroupés en plusieurs groupes.

1. **Les fonctions d’élèment**

Les fonctions d’élément sont utilisées pour définir les propriétés visuelles des éléments individuels d’un graphique dans ggplot2. Chaque élément (titre, axes, légendes, etc.) est associé à une fonction d’élément de thème spécifique qui permet de définir ses caractéristiques telles que la couleur, la taille, la police, etc.

1. **Quelques exemples pour bien comprendre les thèmes**

Il est nécessaire de connaitre les élèments de thème et les fonction d’élèments avant de se lancer dans les représentations graphiques étant donné que les deux vont de pair. Ainsi, nous pouvons montrer quelques exemples pratiques de l’utilisation simultanée des élèments de thème et fonctions d’élèment.

1. **La composante geom**

C’est elle qui précise le type de graphique à représenter. Il existe une multitude de geoms. Nous allons en présenter certains.

1. **Les types de geoms :** geom\_hist() et geom\_freqpoly() : histogramme, polygone de fréquences; geom\_bar() et geom\_col() : diagrammes à barres; geom\_point(): pour les nuages de points; geom\_dotplot(): graphique fait de points; geom\_path() et geom\_line(); (time series) geom\_smooth(): régression; geom\_density() geom\_area(): graphique à aire; geom\_boxplot(); geom\_text() et geom\_label() . . . Voir la fiche récapitulative suivante: <https://thinkr.fr/pdf/ggplot2-french-cheatsheet.pd>.
2. **Quelques exemples de geoms : geom\_bar() ; geom\_hist() ; geom\_density() ;geom\_boxplot() ; geom\_text() ;geom\_label()**
3. **La composante facet**

facet\_wrap est une fonction de ggplot2 qui permet de diviser un graphique en plusieurs panneaux en fonction des niveaux d’une ou plusieurs variables. Cette fonction crée une disposition de panneaux en “enroulement” (wrap), où chaque niveau de la variable spécifiée crée un panneau distinct.

1. **La composante scale**

La fonction scale est utilisée pour modifier les échelles des esthétiques visuelles dans un graphique ggplot. A la composante scale, on peut associer les attributs scale\_color, scale\_size, scale\_x, scale\_y, scale\_area, scale\_shape, chacun permettant de modifier les échelles de l’esthétique correspondante. Prenons comme exemple scale\_color. Les échelles de couleur (scale\_color): Ces échelles attribuent des couleurs à des valeurs spécifiques de la variable. Dans scale\_color, nous avons scale\_color\_radient, scale\_color\_manual ou encore. . .