POLYTECH LILLE

Enseignant : Sofiane MAAZI

Introduction aux logiciels pour les statistiques (IS2A3)

2024-2025

L'environnement de travail

1. Créez un dossier *CoursR* sur votre poste de travail.
2. Décrivez les fenêtres du logiciel RStudio.
3. Créez un script *MonPremierTP*
4. Importez le jeu de données *mtcars.*

Interroger une table de données

1. Importez les tables *immeubles* et *communes*.
2. Donnez le nombre d'observations et de variables et la structure de la table *communes*.
3. Donnez le nom des variables de *communes*.
4. Donnez les propriétés qui caractérisent une table de données en R.
5. Sélectionnez une variable de *communes* en utilisant 3 façons différentes.
6. Remplacez les valeurs manquantes de communes.
7. Fusionnez les tables *immeubles* et *communes.*
8. Combien de monuments historiques des Hauts-de-France sont protégés depuis 1945 ?
9. Combien de villes de Normandie ont des monuments historiques ?
10. Donnez le nombre moyen d'habitants par région.
11. Donnez la commune ayant le plus d'habitants.
12. Une commune reçoit une dotation de 100 € par habitant. Créez la variable *Dotation* afférente.

Variables quantitatives continues

1. Importez le jeu de données iris.
2. Donnez la médiane, les quartiles, l'étendue et l'intervalle interquartile de *Petal.Length*.
3. Donnez la moyenne, la variance, l'écart-type de *Petal.Length*.
4. Donnez l'intervalle de confiance de la moyenne de *Petal.Length*.
5. Quel est le rôle de la commande *summary()* ?

Variables qualitatives et quantitatives discrètes

1. Importez le jeu de données *TitanicSurvival*. On s'intéresse à la variable *passengerClass*.
2. Donnez les effectifs de chacune des classes de regroupement de la variable.
3. Donnez la fréquence relative de chacune des classes de regroupement de la variable.
4. Donnez les effectifs totaux et effectifs cumulés des classes de regroupement de la variable.
5. Donnez les proportions cumulées des classes de regroupement de la variable.
6. En déduire comment construire le tableau de contingence associé à la variable.
7. Donnez l'intervalle de confiance de la proportion de femmes sur le Titanic.

Graphiques

1. Importez le jeu de données *diamonds.*
2. Représentez graphiquement la variable *clarity* par un diagramme en barre puis un camembert.

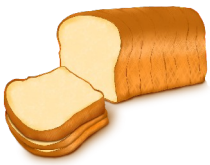
3. Représentez graphiquement la variable *table* par un histogramme puis diagramme de Tuckey.
4. Enregistrez l'histogramme de la question précédente au format pdf.
5. Expliquer les avantages et les inconvénients du package *ggplot2* utilisé en l'état.
6. En déduire le rôle de l'addin *esquisse*.

Tests statistiques

Le principe

Le Bureau Communautaire de Référence de la Communauté Européenne établit des références pour les traces inorganiques dans les denrées alimentaires. Ainsi, pour la quantité de zinc présent dans le pain complet, la moyenne de référence est de 19.5 microgrammes par gramme. Un expérimentateur prélève 8 échantillons de pain complet dans une production et mesure la teneur en zinc dans chacun d'entre eux. Les résultats sont :

$$21.1 - 21.5 - 21 - 21.6 - 19.1 - 20.3 - 22.2 - 21.5$$



On suppose que la teneur de zinc en microgramme (mg) présent dans le pain complet peut être modélisée par une variable aléatoire Y suivant une loi normale. La problématique est la suivante : Peut-on affirmer, avec un faible risque de se tromper, que la production de pain complet n'est pas conforme aux règles européennes concernant la teneur en zinc ?

1. Explicitez les hypothèses à tester.
2. Déterminez la *p-valeur* associée en utilisant la commande ($\Rightarrow$ *t.test*).
3. Répondez à la problématique en précisant le degré de significativité en cas de rejet de H_0.

Quelques tests

Pour chacune des situations présentées dans le diaporama du cours :
1. Choisissez le but du test dans la liste ci-dessous :
   (a) Comparer un pourcentage d'un échantillon à celui d'une population.
   (b) Vérifier l'indépendance entre 2 variables.
   (c) Vérifier que les répartitions de différents effectifs sont équivalentes.
   (d) Évaluer les effets d'un nouveau test diagnostic à ceux d'un test de référence.
   (e) Existence d'un lien entre une variable quantitative et une variable qualitative.
   (f) Existence d'un lien entre 2 variables quantitatives.
2. Formulez H_0 .
3. Choisissez le test adéquat dans la liste ci-dessous :
   (a) χ^2 de conformité
   (b) χ^2 d'indépendance
   (c) χ^2 d'homogénéité
   (d) Mac Nemar
   (e) tests paramétriques
4. Vérifiez que les conditions de réalisation du test sont remplies.
5. Quantifiez le risque d'incertitude en fixant un seuil.
6. Calculez la p-value et concluez.

Sources des images : pixabay.com et freepik.com