Statistiques - Un Kit de Survie

Table des matières

Qu'est-ce qu'un nombre?

10 min

Qu'est-ce qu'un nombre?

Qu'est-ce qu'un nombre?

Dans le cadre de ce kit,

- un indicateur sur une tendance, une aide à la décision
- un moyen d'appuyer un argumentaire ou de le remettre en question.

À partir d'une proportion de comportements constatés sur un sous-ensemble d'étude on infère une tendance générale.

Échantillon et population

- Un indicateur concernant une population à l'étude n'est pas toujours possible.
- Un échantillon est une sous-partie de cette population.

À partir d'une mesure sur un échantillon d'une population à l'étude on souhaite obtenir une information sur cette population.

Récolter des données

Quand on parle de statistiques, on sous-entend

- une démarche
- des données brutes

Mettez toujours en place dans vos projets un process de récolte et sauvegarde de vos données de travail (tout en étant RGPD compliant bien sûr).

Qu'est-ce qu'un dataset?

Qu'est-ce qu'un dataset?

Un ensemble de ***d'observations*** contenant les valeurs communes de ***caractéristiques*** à l'étude.

- Les observations / les individus
- Les variables / features
  - Numériques
  - Catégorielles (ou qualitatives), la valeur d'une variable catégorielle est une modalité

30 min

Travail pratique, découverte de R et RStudio.

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

Mesures empiriques

- Les mesures de centrage
  - Moyenne
  - Médiane
  - Mode
- Les mesures de dispersion
  - L'intervalle de variation
  - Variance
  - Les écarts inter-quartiles
- La corrélation

Mesures empiriques

Mesures de centrage

Une valeur autour de laquelle se répartissent les données observées

Les plus classiques sont

- la moyenne
- la médiane
- le mode (pour une variable catégorielle).

Quid d'une proportion?

Mesures de dispersion

La disposition des données observées autour d'une mesure de centrage

Elles concernent les variables numériques. Les plus classiques sont

- l'intervalle de variation
- la variance / l'écart-type
- les écarts inter-quartiles.

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

La corrélation

Est-ce que deux features sont liées?

Elles concernent les variables numériques et catégorielles.

- le coefficient de corrélation linéaire entre variables numériques
- le coefficient du Khi deux entre variables catégorielles.

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

20 min

À vous de conjecturer ce que bon vous semble sur les datasets mis à votre disposition

L'inférence statistique

Problème

On a une information observée sur des mesures empiriques liées à un échantillon (ou plusieurs) d'une population, dans quelle mesure celle-ci est valable sur les populations à l'étude?

Dans notre cadre

Dans quelle mesure

- une moyenne conjecturée provient d'un échantillon de la population étudiée?
- la différence entre les moyennes empiriques sur deux sous-populations est une différence reflétée par la population étudiée?
- la corrélation empirique entre deux variables est valable sur la population étudiée?

Formuler un test statistique

On cherche à tester une ***hypothèse*** H0, de complémentaire H1. Dans nos cas l'hypothèse H0 est l'hypothèse qu'on cherche à rejeter. Par exemple

- Tester si une valeur *m* peut-être moyenne d'une population dont on a un échantillon.
- Tester si deux moyennes *m* et *n* sont significativement distinctes.
- Tester si un coefficient de corrélation est significativement non nul.

45 min

Retour sur datasets

Synthèse de la séance

Merci à tous