Résumé des Statistiques en Data Science

by REDA OUZIDANE

3 avril 2025

1 Statistiques Descriptives

Mesures de tendance centrale :
— **Moyenne** : $\bar{x} = \frac{\sum x_i}{n}$
— **Médiane** : Valeur centrale d'un ensemble trié.
— **Mode** : Valeur la plus fréquente.
Mesures de dispersion :
— **Variance** : $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$
— **Écart-type** : $\sigma = \sqrt{\text{Variance}}$
— **Intervalle interquartile (IQR)** : $Q3 - Q1$

2 Probabilité et Distributions

Règles de probabilité :
— Addition : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
— Multiplication : $P(A \cap B) = P(A)P(B)$ (si indépendant)
— Théorème de Bayes : $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
Distributions de probabilité :
— **Binomiale** : $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
— **Normale** : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
— **Poisson** : $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

3 Inférence Statistique

Tests d'hypothèses :

— Test Z (grand échantillon, variance connue)
— Test T (petit échantillon, variance inconnue)
— Test Khi-deux (variables catégorielles)
— ANOVA (comparaison de plusieurs moyennes)

Intervalle de confiance :

$$CI = \bar{x} \pm Z \times \frac{\sigma}{\sqrt{n}} \tag{1}$$

4 Corrélation et Régression

Corrélation :

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}} \tag{2}$$

Régression linéaire :

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{3}$$

5 Standardisation et Normalisation

Standardisation (Z-score) :

$$Z = \frac{X - \mu}{\sigma} \tag{4}$$

Normalisation (Min-Max Scaling) :

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5}$$

Encodage des variables catégoriques :
— **One-Hot Encoding** (vecteurs binaires)
— **Label Encoding** (conversion en indices numériques)

6 Conclusion

Les statistiques sont essentielles en Data Science pour l'analyse de données et l'optimisation des modèles de Machine Learning.
By REDA OUZIDANE.