
STATISTIQUE
Tout ce que le developpeur data doit savoir

Cette veille  porte sur les notions de statistique utiles.
Tout d’abord, qu’est-ce que la statistique ?
La statistique est la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles par tous.
Par exemple, les relevés des nombres de pannes observées dans une unité de production constituent une statistique.

Les concepts développés en statistique sont utiles dans de nombreux domaines et font partie des connaissances de base de l’ingénieur, de l’économiste et du scientifique en général. Parmi les nombreuses applications dans l’industrie, on peut citer la fiabilité, le contrôle de qualité, la maîtrise statistique des procédés.

# Statistiques descriptives:

Les statistiques descriptives permettent de décrire et de résumer des données à l'aide de différentes mesures de tendance centrale comme la moyenne, le mode et la médiane

- Moyenne: mesure de tendance centrale qui indique où se situe le centre d'un ensemble de données.
- Formule: moyenne = **Σ**x / n (où **Σ**x représente la somme des scores et n représente le nombre total de scores)
- Mode: la valeur qui apparaît le plus souvent dans un ensemble de données.
- Médiane: la valeur centrale d'un ensemble de données trié.

# Correlations

La corrélation permet d'analyser la relation entre deux variables et le coefficient de corrélation peut être positif, négatif ou nul

Coefficient de corrélation de Pearson (r): mesure la relation linéaire entre deux variables. Il varie de -1 (corrélation négative parfaite) à 1 (corrélation positive parfaite).

**Cas 1**

## Coefficient de corrélation positif

les deux variables évoluent dans la même direction

**Cas 2**

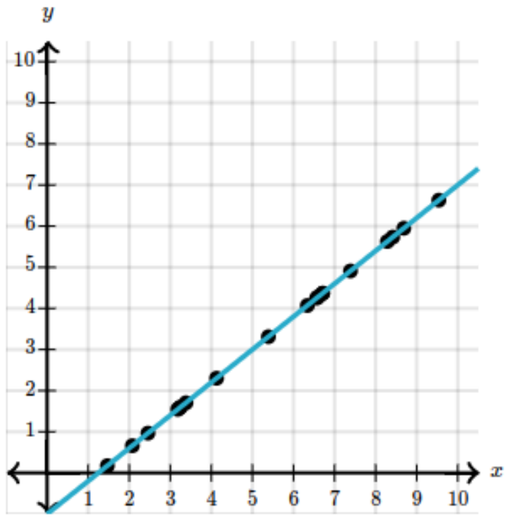
## Coefficient de corrélation négatif

les deux variables évoluent dans des directions opposées.

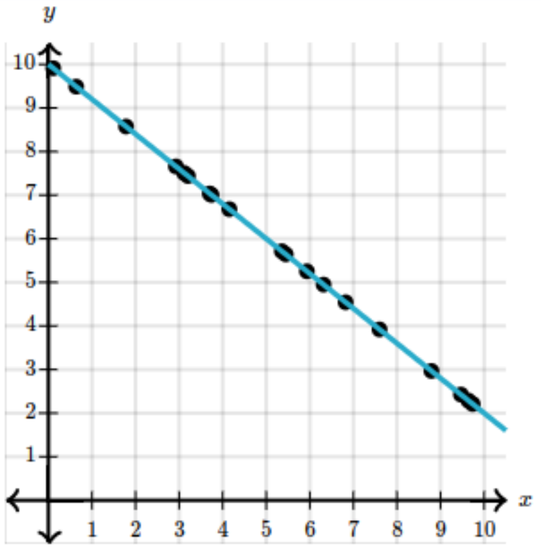
**Cas 3**

## Coefficient de corrélation nul

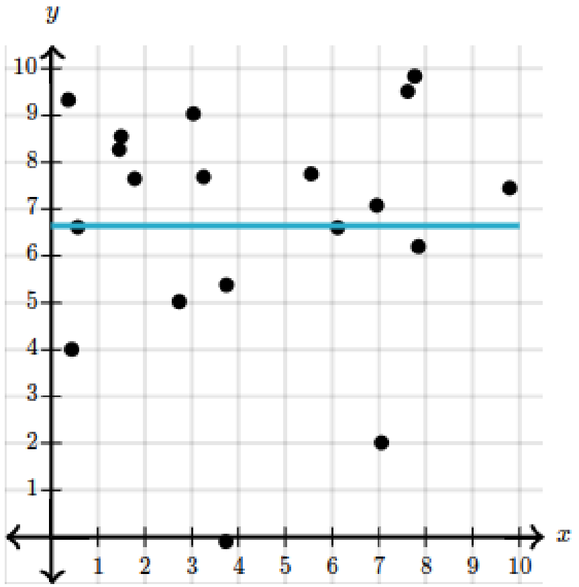
Il n'existe pas de corrélation




ici, $r = 1$ : corrélation positive parfaite entre les deux variables



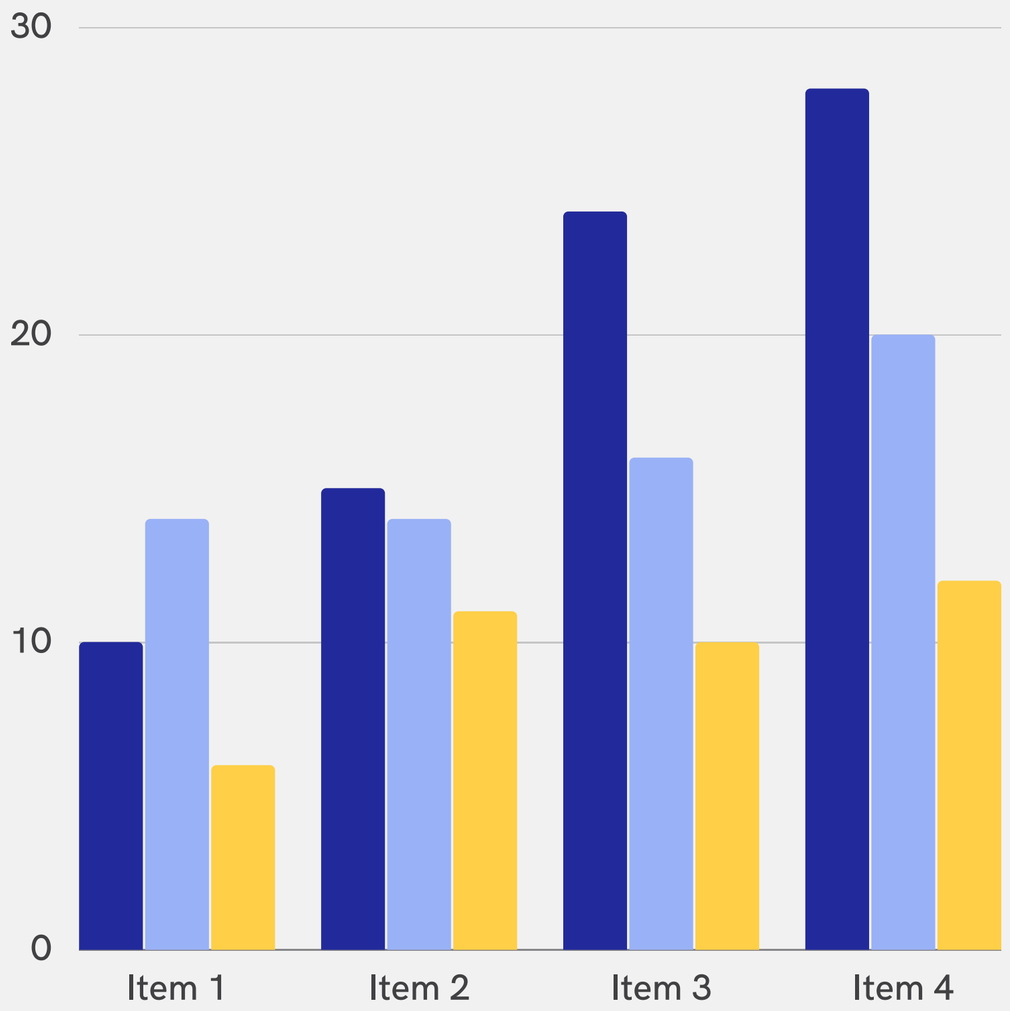

ici, $r = -1$ : corrélation négative parfaite entre les deux variables




ici, $r = 0$ : absence totale de corrélation, les deux variables sont

# Variabilité

La variabilité mesure l'étendue et la dispersion des données à travers des mesures telles que l'écart-type, la variance, la plage, les percentiles, les quartiles et l'intervalle interquartile.



# Ecart-Type

Mesure de dispersion des données par rapport à la moyenne. Il indique la quantité de variation ou d'étalement des données.
Plus l'écart-type est faible, plus la population est homogène.

# Variance

La variance est utilisée dans le domaine de la statistique et de la probabilité en tant que mesure servant à caractériser la dispersion d’une distribution ou d’un échantillon. Il est possible de l’interpréter comme la dispersion des valeurs par rapport à la moyenne.

# Plage

La plage est une valeur numérique qui indique la différence entre la valeur maximale et minimale d'une population ou d'un échantillon statistique.
La plage est généralement utilisée pour obtenir la dispersion totale. C'est-à-dire que si nous avons un échantillon avec deux observations : 10 et 100 euros, la fourchette sera de 90 euros.

# Regression

La régression est une méthode d'analyse statistique qui permet de modéliser la relation entre une variable dépendante, que nous cherchons à décrire, à expliquer, à prédire et une ou plusieurs variables indépendantes ou explicatives.

## Regression Linéaire

Elle modélise la relation linéaire entre une variable dépendante et une ou plusieurs variables indépendantes. Elle peut être utilisée pour prédire la valeur d'une variable dépendante à partir de la valeur d'une variable indépendante.

Formule: y = a + bx

## Regression Logistique

Elle modélise la relation non-linéaire entre une variable dépendante binaire et une ou plusieurs variables indépendantes

$$f(x) = \frac{1}{1 + e^{-x}}$$

# Distribution des Probabilités

permet de calculer les probabilités d'événements indépendants ou dépendants

## Événement indépendant

probabilité d'un événement ne dépend pas de la réalisation d'un autre événement.

## Événement dépendant

probabilité d'un événement dépend de la réalisation d'un autre événement.

## Distribution Normale

Distribution de probabilité continue en forme de cloche, souvent utilisée pour modéliser des phénomènes naturels et sociaux.

# BIAIS

Le biais est une erreur systématique qui peut affecter les résultats statistiques.

## Biais de Selection

un biais de sélection se produit lorsque les participants d'une étude ne sont pas choisis au hasard, ce qui peut entraîner une distorsion des résultats

## Biais d'intervalle de temps

Biais d'intervalle de temps : ce biais se produit lorsqu'il y a une différence dans le moment où les données sont collectées ou mesurées pour différents groupes

## Théorème de la Limite Centrale

le théorème de la limite centrale est un concept important en statistique qui indique que la moyenne d'un grand nombre d'échantillons aléatoires d'une population suivra une distribution normale.  Cela signifie que même si la distribution de la population d'origine n'est pas normale, la moyenne des échantillons sera normale.

## Convariance

la covariance entre deux variables aléatoires est un nombre permettant de quantifier leurs écarts conjoints par rapport à leurs espérances respectives. Elle s’utilise également pour deux séries de données numériques (écarts par rapport aux moyennes)

## Relation entre les variables

Relation entre les variables : en statistique, la relation entre deux variables peut être exprimée en termes de corrélation (linéaire) ou de covariance (linéaire et non linéaire). La corrélation mesure la force et la direction de la relation linéaire entre deux variables, tandis que la covariance mesure la force de la relation linéaire et non linéaire entre deux variables.

## Test d'hypotjhèse

un test d'hypothèse est une méthode statistique qui permet de déterminer si les résultats d'une étude sont statistiquement significatifs ou s'ils pourraient être dus au hasard. Il est basé sur la comparaison entre les données observées et les données attendues sous l'hypothèse nulle

# Covariance

En statistiques, la covariance est une méthode mathématique permettant d'évaluer le sens de variation de deux variables et, par là, de qualifier l'indépendance de ces variables.

## Relation entre les variables

Cette relation peut être vérifiée par plusieurs grandeurs telle la covariance et le coefficient de corrélation

## Test d’hypothèse

Un test d’hypothèse (ou test statistique) est une démarche qui a pour but de
fournir une règle de décision permettant, sur la base de résultats d’échantillon, de faire
un choix entre deux hypothèses statistiques.


## Compromis biais/variance

Le dilemme (ou compromis) biais-variance est le problème de minimiser simultanément deux sources d'erreurs qui empêchent les algorithmes d'apprentissage supervisé de généraliser au-delà de leur échantillon d'apprentissage

THANK YOU

# JEREJEF

Avez-vous des questions ?