RECHERCHE SUR LA STATIQUE DESCRIPTIVE

1. Statistiques descriptives:

- **La moyenne:** la somme de toutes les observations divisée par le nombre total d'observations ($\frac{\sum_{i=0}^{n} X_i}{n}$).

  Par exemple, la moyenne de 5, 7 et 10 est : $\frac{(5+7+10)}{3} = 7.33$.

- **Le mode:** la valeur qui apparaît le plus souvent dans un ensemble de données.

  Par exemple, le mode de 3, 3, 5, 7 et 9 est 3.

- **La médiane:** la valeur du milieu lorsque les données sont triées par ordre croissant ou décroissant. Par exemple, la médiane de 2, 5, 8, 10, 12 est 8.

2. Corrélation:

- **Coefficient de corrélation supérieur à zéro:** lorsque deux variables ont une relation positive, le coefficient de corrélation est supérieur à zéro. Par exemple, si l'augmentation de la température est associée à l'augmentation du nombre de ventes, le coefficient de corrélation est supérieur à zéro.
- **Coefficient de corrélation inférieur à zéro:** lorsque deux variables ont une relation négative, le coefficient de corrélation est inférieur à zéro. Par exemple, si l'augmentation du prix est associée à la diminution du nombre de ventes, le coefficient de corrélation est inférieur à zéro.
- **Coefficient de corrélation nul:** lorsque deux variables n'ont pas de relation, le coefficient de corrélation est nul.

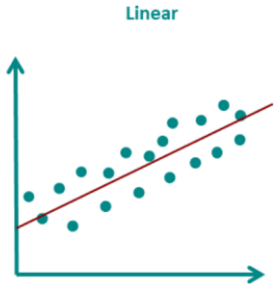
3. Variabilité:

- **Écart-type:** mesure de la dispersion des données par rapport à la moyenne. Plus l'écart-type est élevé, plus les données sont dispersées. La formule de l'écart-type est : $\sqrt{\frac{1}{n}\sum_{i=0}^{n}(X_i - \bar{X})^2}$
- **Variance:** carré de l'écart-type. La formule de la variance est : $\frac{1}{n}\sum_{i=0}^{n}(X_i - \bar{X})^2$
- **Plage:** différence entre la plus grande et la plus petite valeur. Par exemple, la plage de 2, 5, 8, 10, 12 est 10 (12-2).

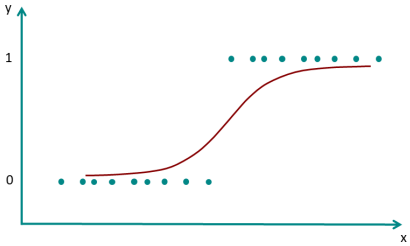
- **Percentile:** la valeur en dessous de laquelle un certain pourcentage des données se situe. Par exemple, le 25ème percentile est la valeur en dessous de laquelle 25% des données se situent.
- **Quartile:** les trois valeurs qui divisent un ensemble de données en quatre parties égales. Le premier quartile (Q1) est la médiane des données situées en dessous de la médiane. Le troisième quartile (Q3) est la médiane des données situées au-dessus de la médiane. Le deuxième quartile (Q2) est la médiane de l'ensemble de données.
- **Intervalle interquartile:** la différence entre le troisième et le premier quartile. C'est un indicateur de la dispersion des données autour de la médiane.

4. Régression:

- **La régression linéaire:** une méthode utilisée pour modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes. La formule de la régression linéaire est : $Y = a + b.X$, où y est la variable dépendante, x est la variable indépendante, a est l'ordonnée à l'origine et b est la pente de la ligne de régression.



- **La régression logistique**: La régression logistique est utilisée pour modéliser la relation entre une variable dépendante dichotomique (par exemple, oui / non) et une ou plusieurs variables indépendantes continues ou discrètes. La relation est modélisée par une équation en forme de courbe en S, appelée "courbe de régression logistique". Cette courbe est trouvée en maximisant la vraisemblance de l'échantillon observé.



5. Distribution de Probabilité :

- Les événements indépendants sont des événements qui n'ont pas d'impact l'un sur l'autre. Par exemple, si vous lancez une pièce de monnaie, le résultat de la première fois que vous la lancez n'affecte pas le résultat de la deuxième fois que vous la lancez.

  La probabilité conjointe de deux événements indépendants A et B est le produit de leur probabilité respective :

$$P(A \cap B) = P(A) \times P(B)$$

  Les événements dépendants sont des événements qui ont un impact l'un sur l'autre. Par exemple, si vous tirez une carte d'un jeu de cartes, la probabilité de tirer une deuxième carte de la même couleur dépend de la carte que vous avez tirée en premier.

  La probabilité conditionnelle de l'événement A sachant que l'événement B s'est produit est donnée par :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

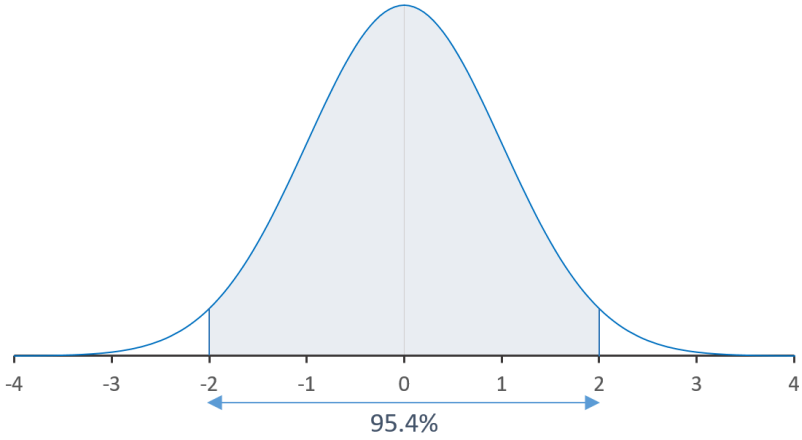
5. Distribution Normale:

- La distribution normale, également appelée distribution gaussienne, est une distribution de probabilité continue qui est souvent utilisée pour modéliser des phénomènes naturels. Elle a une forme en cloche symétrique et est définie par deux paramètres, la moyenne et l'écart type.

  La fonction de densité de probabilité de la distribution normale est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

  Voici une illustration graphique de la loi normale :



7. Biais

Le biais est une erreur systématique qui se produit dans une étude ou une expérience. Il existe plusieurs types de biais :

- **Biais de sélection** : se produit lorsque l'échantillon n'est pas représentatif de la population. Par exemple, si une étude sur l'hypertension est menée uniquement dans une région où la population est en meilleure santé que la moyenne nationale, cela peut conduire à une sous-estimation du taux d'hypertension.
- **Biais d'intervalle de temps :** se produit lorsque les données sont recueillies à différents moments dans le temps et que cela affecte les résultats. Par exemple, si une étude sur le tabagisme est menée à une époque où les cigarettes étaient moins nocives qu'aujourd'hui, cela peut conduire à une sous-estimation des effets néfastes du tabagisme.
- **Le théorème de la limite centrale :** indique que la moyenne d'un grand nombre d'échantillons aléatoires d'une population tend vers une distribution normale.
- **Compromis biais/variance :** il est souvent nécessaire de faire un compromis entre le biais et la variance d'une méthode statistique pour obtenir des résultats précis.
- **Test d'hypothèse :** permet de déterminer si une différence observée entre deux groupes est significative ou simplement due au hasard.
- **Relation entre les variables :** l'analyse de la relation entre deux variables peut être effectuée en calculant le coefficient de corrélation, qui mesure la force et la direction de la relation.

Afin de déterminer l'intensité de la corrélation linéaire entre deux ensembles de données, nous pouvons utiliser **le coefficient de corrélation de Pearson**, noté r.Plus précisément,

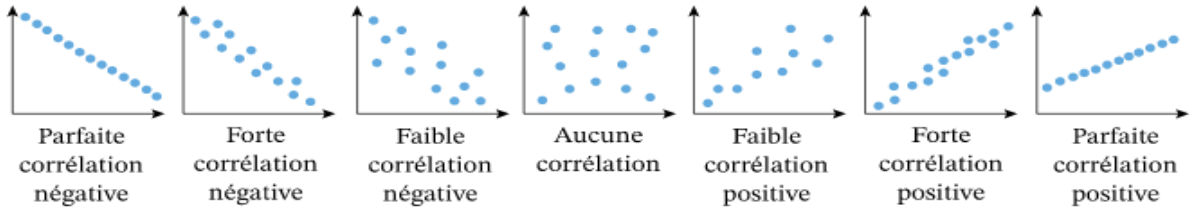
- si deux variables ont une **forte corrélation positive (directe)**, alors r **est proche de** 1 ;
- si deux variables ont une **faible corrélation positive (directe)**, alors r **est positif, mais plus proche de** 0 **que de** 1 ;
- si deux variables ont une **forte corrélation négative (inverse)**, alors r **est proche de** −1 ;
- si deux variables ont une **faible corrélation négative (inverse)**, alors r **est négatif, mais plus proche de** 0 **que de** −1 ;
- s'il n'y a **pas de corrélation**, alors r **est proche de** 0. Et on a : $\mathbf{r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}}$

  où $\boldsymbol{S_{xy}}$, $\boldsymbol{S_{xx}}$ et $\boldsymbol{S_{yy}}$ sont les statistiques sommaires définies comme

  $$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}, \quad S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}, \text{ et } S_{xy} = \sum xy - \frac{\sum x \sum y}{n};$$

  où x représente les valeurs d'une variable, y représente les valeurs de l'autre variable et n représente le nombre de points de données.

Cela peut également être vu sur la droite numérique ci-dessous :



- **Covariance :** mesure la corrélation entre deux variables en prenant en compte leur écart à leur moyenne respective. Une covariance positive indique une corrélation positive entre les deux variables, tandis qu'une covariance négative indique une corrélation négative. La covariance de deux variables aléatoires réelles X et Y ayant chacune une variance (finie), notée Cov(X, Y) et on a Variance de X, Var(X)=Cov(X,X). Pour calculer la covariance on a la formule suivante :

$$\mathrm{Cov(X,Y)} = S_{xy} = \frac{1}{\mathrm{n-1}} \sum_{i=0}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

Par Ibrahima Gabar Diop