Statistique Inférentielle
Chapitre 20 : Introduction

A propos du cours

Finalité du cours

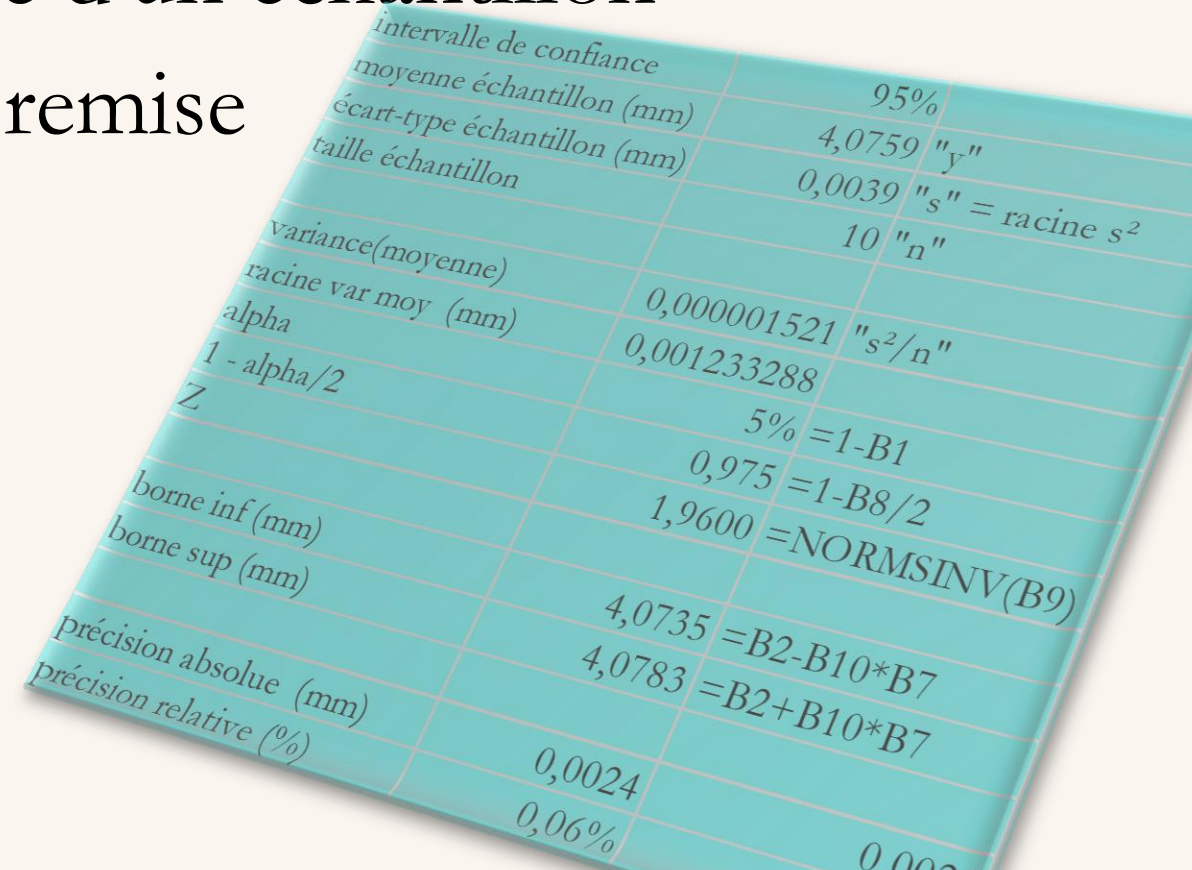
- s'approprier des connaissances générales en statistique en vue de traiter des données économiques, financières, commerciales, ….
- développer son sens critique dans l'interprétation de données ou documents statistiques
- utiliser des ressources logicielles existantes permettant le traitement statistique de données

Programme du cours

- Concepts de base de la statistique inférentielle (chap. 20)
- Estimation sur base d'un échantillon (chap. 21)
- Test d'hypothèse (chap. 22)
- Test de corrélation (chap. 23)

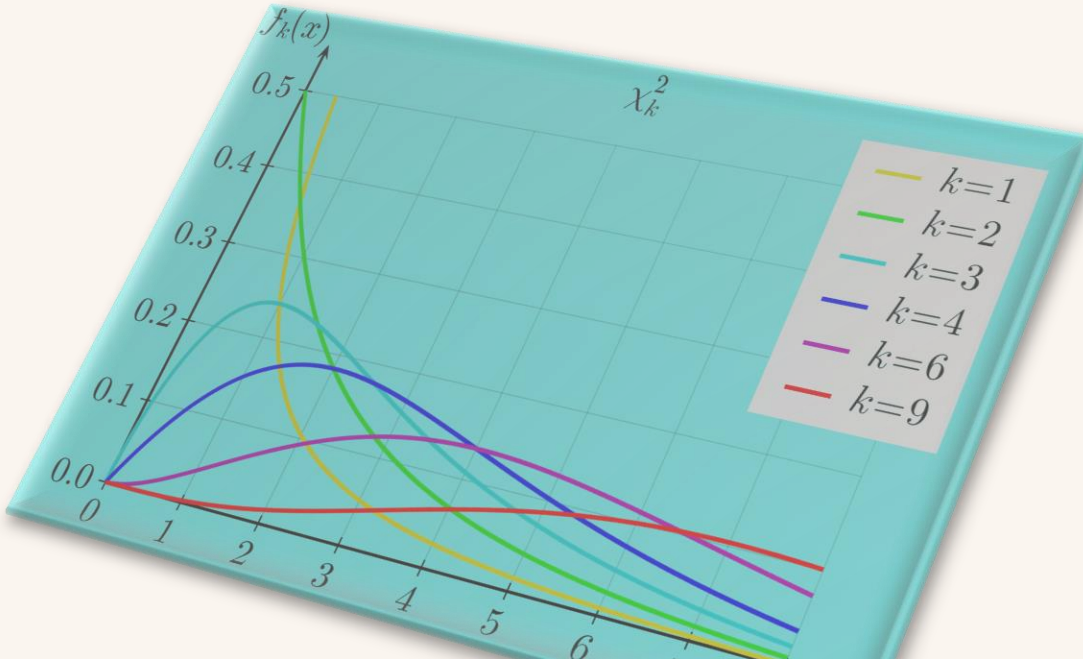
Chapitre 21 : Estimateur

- Quantile
- Moyenne et variance empirique d'un échantillon
- Probabilité égale avec ou sans remise
- Précision de l'estimation
- Intervalle de confiance
- Taille de l'échantillon
- Estimation d'une proportion



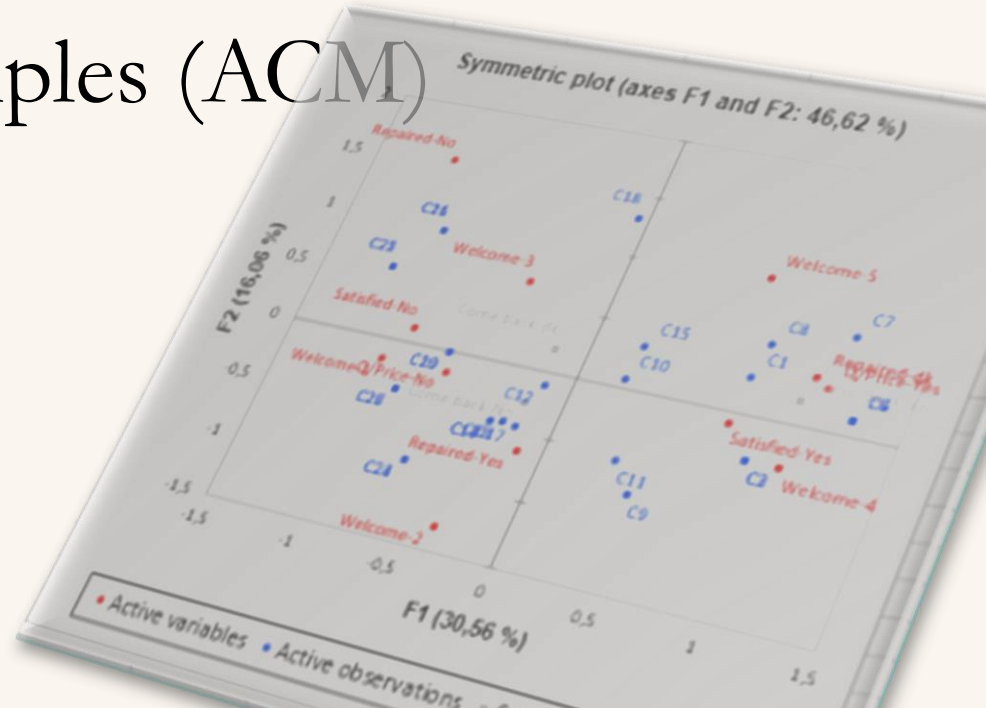
Chapitre 22 : Test d'hypothèse

- Définition d'un test
- Hypothèse, acceptée ou rejetée
- Niveau et puissance
- Comparaison à une valeur de référence
- Test du X^2 (khi-deux) d'ajustement
- Test du X^2 d'indépendance
- Test de Student
- Test de Fischer



Chapitre 23 : Test de corrélation

- Régression simple
- Analyse en Composantes Principales (ACP)
- Analyse Factorielle des Composants (AFC)
- Analyse des Correspondances Multiples (ACM)



Définitions - RAPPEL

un peu d'histoire

définition générale

paradigmes

Population et individus

Variables et modalités

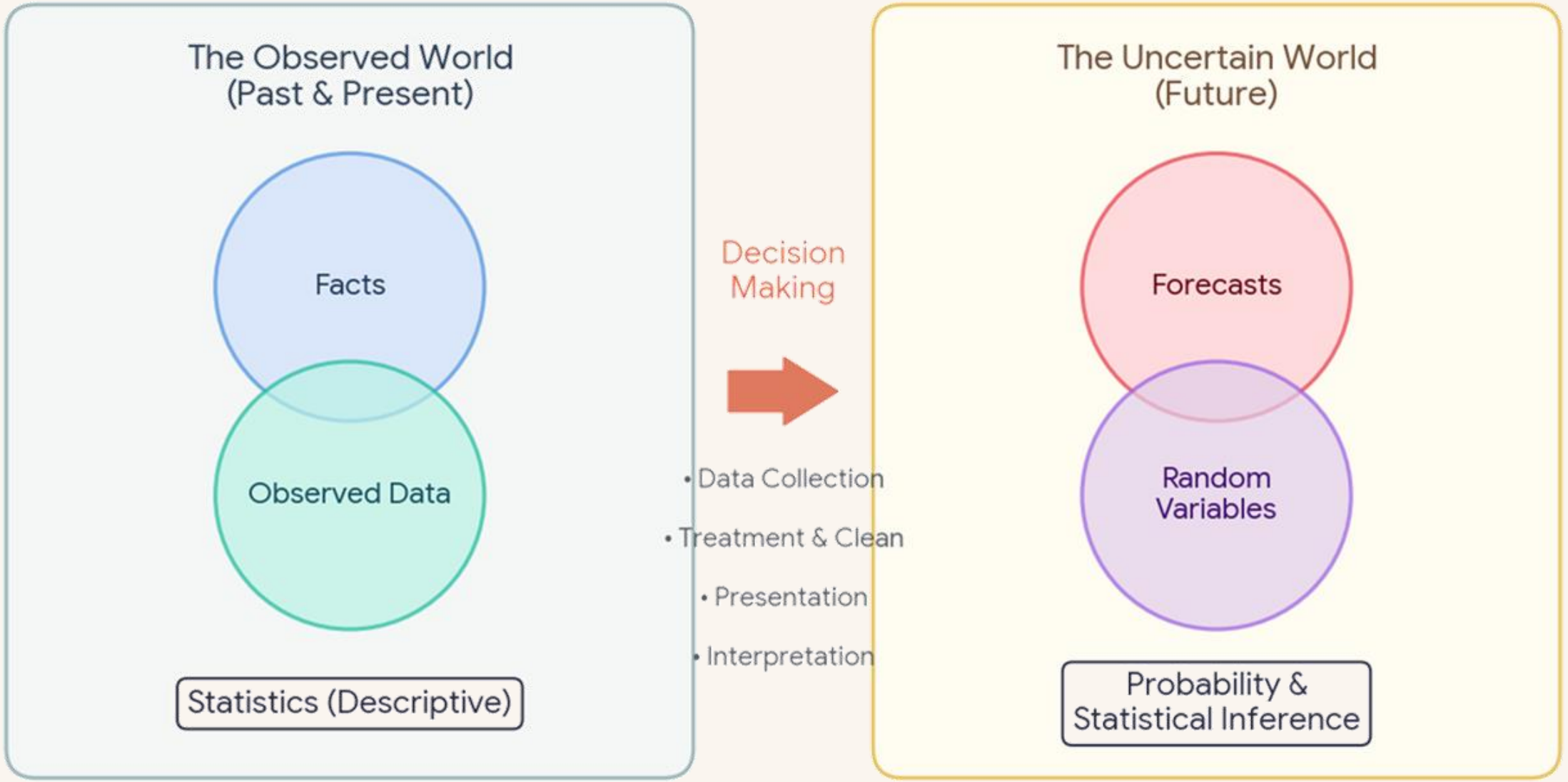
Types de variables

Échantillon

Statistique : Définition

- Ensemble des méthodes qui ont pour objet, sur base de données d'observation sur un groupe d'individus
  - la collecte des données à observer
  - leur traitement et leur analyse
  - leur présentation, visant à les rendre compréhensibles
  - l'interprétation des résultats en vue d'une prise de décision

Data Science Framework: Descriptive vs. Inferential Horizons



Unified Analytical Horizon

Statistique **Descriptive** : Définition

- Ensemble des techniques permettant de
  - Résumer et présenter les données
  - Faciliter leur compréhension
  - Représenter les données sous forme de tableaux, de graphiques et d'indicateurs numériques.

… quand les données disponibles sont nombreuses et complètes.

Statistique Inférentielle : Définition

Ensemble des techniques permettant de

- induire (prédire) les caractéristiques d'une population à partir de celles d'un échantillon,
  - en estimant la (in)certitude de la prédiction.
- tester si les données d'un échantillon corroborent bien une hypothèse de départ concernant une population.

Dit autrement,

- L'échantillon est traité d'une manière exacte (statistique descriptive).
- Puis, les résultats sont élargis à l'échelle de la population, en estimant l'incertitude (statistique inférentielle).

Statistique vs Probabilité : Définition

Statistiques	Probabilités
Données observées = valeurs mesurées ou collectées Exemple : taille des étudiants dans une classe	Données aléatoires = Résultat d'un processus incertain Exemple : résultat d'une enquête sur le port du chapeau
Analyse de données observées. Exemple : calculer la moyenne de la taille des étudiants de cette cette classe.	Étude de phénomènes aléatoires. Exemple : estimer la probabilité de porter un chapeau bleu
Descriptive (résumer les données) vs Inférentielle (généraliser à une population).	Théorique : Modéliser l'incertitude.

"Data Science" : Définition

- "Data Science"
  - en cas de données massives
  - statistiques = outils pour interpréter et analyser ces données massives
- Exemple
  - Une entreprise de vente en ligne souhaite analyser les données de ses ventes pour améliorer sa stratégie marketing.
  - Collecte des données
    - Historique des ventes
    - Données clients
  - Traitement et analyse
    - Identification des produits les plus vendus par catégorie.
    - Analyse des corrélations
      - âge des clients vs. type de produit acheté
- Intégration dans la Data Science
  - régression : prédiction des ventes du prochain trimestre en fonction des données historiques.
  - classification : segmentation des clients en groupes
    - "acheteurs réguliers" vs "occasionnels"
- Résultat final attendu de la Data Science
  - Le produit X est le plus vendu par les clients de 25-35 ans en été.
  - Recommandation : cibler cette tranche d'âge avec des promotions en juin.

Paradigmes statistiques : Trois façons d'interpréter la probabilité

La disponibilité et le coût des données orientent le choix de l'approche.

01 Fréquentisme
OBJECTIVISME

IDÉE-CLÉ

La répétition révèle la probabilité par la fréquence observée.

EXEMPLE

10 000 lancers

La fréquence de « pile » s'approche de 0,5.

→ **P(pile) = 0,5**

02 Subjectivisme
CONFIANCE

IDÉE-CLÉ

La probabilité traduit un degré de confiance révisable.

EXEMPLE

Une élection

« Quelle est la probabilité que ce candidat gagne l'élection ? »
Pas de répétition possible : le chiffre reflète notre confiance, à réviser avec l'expérience.

03 Bayésisme
APPRENTISSAGE

IDÉE-CLÉ

Chaque nouvelle donnée met à jour le poids des hypothèses.

EXEMPLE

Plusieurs diagnostics

En médecine, chaque résultat d'examen renforce ou affaiblit chacune des hypothèses.

Variables qualitatives et quantitatives

- Variable qualitative (ou "nominales"):
  - Elle exprime une qualité.
  - Elle se traduit par des noms:
    - origine sociale
    - groupe sanguin
    - préférence politique ou artistique
  - On ne peut pas la représenter par une quantité.
    - Pas de traitement mathématique

$$x = m_1, m_2, \ldots, m_q$$

Variables qualitatives et quantitatives

- Variable quantitative:
  - Elle exprime une quantité.
  - Elles sont mesurables.
    - taille
    - poids
    - notes d'un examen
  - Elles sont traitables numériquement.

$$x_i = 1, 2, 3 \ldots n$$

Variables qualitatives et quantitatives

- Deux types de variable quantitative :
  - la variable quantitative discrète
    - nombre fini de valeurs possibles
    - exemple
      - nombre de personnes vivant dans mon domicile
  - la variable quantitative continue
    - toutes les valeurs sont possibles dans un intervalle [a,b].
    - exemple
      - Poids, taille, revenu, …

Population et individus

- Individu ou unité statistique
  - Une unité distincte chez laquelle on peut observer une ou plusieurs caractéristiques données
- Population
  - Ensemble des unités ayant au-moins une propriété commune.
- Échantillon
  - Un sous-ensemble de la population.
  - Supposé être représentatif de la population.

- Exemple
  - On veut étudier le niveau d'études de la population bruxelloise adulte
  - unité
    = la personne adulte habitant Bruxelles
    = 1 personne
  - population
    = l'ensemble des Bruxellois adultes
    = 1 200 000 pers.
  - échantillon
    = sous-ensemble des Bruxellois adultes représentatifs de la population
    = 2000 pers. réparties dans toute la ville

Définitions propres à la statistique inférentielle

problématique

population

échantillon

hypothèse

test

corrélation

analyse factorielle

La problématique

L'inférence statistique permet d'obtenir de l'information sur une population à partir d'informations recueillies sur un échantillon.

- La population peut être
  - finie ou bien définie
    - ensemble des habitants d'un village
  - finie mais mal définie
    - ensemble des insectes d'Amazonie
  - infinie
    - ensemble de prélèvements d'eau dans l'océan



Échantillon

- Échantillon
  - partie de la population qui est réellement observée
- Échantillon aléatoire
  - tous les individus de la population ont la même chance (probabilité) d'appartenir à l'échantillon
- Echantillon simple
  - tous les individus de l'échantillon sont prélevés de manière indépendante les uns des autres



- On considère toujours ici des échantillons simples et aléatoires
  - sauf si précisé autrement.

Hypothèse statistique

- On étudie une population dont ses unités possèdent une propriété représentée par une valeur encore inconnue
- Une hypothèse est formulée sur la valeur du paramètre.
- On prélève un échantillon.
- On vérifie si l'hypothèse doit être rejetée ou acceptée.

- Exemple
  - Une usine produit des ampoules dont la durée de vie moyenne théorique est de 10 000 heures.
  - Le responsable qualité soupçonne que la durée de vie a diminué après un changement de fournisseur d'un composant électronique.
  - Il souhaite vérifier cette hypothèse.

Test d'hypothèse

- On veut vérifier si l'hypothèse formulée est supportée ou non par les observations.
- On procéde alors à un test d'hypothèse.
- Une hypothèse HO est formulée sur la valeur du paramètre.
  - Et aussi l'hypothèse alternative HA.
- Méthodes appliquées
  - Pour conclure si l'écart observé entre la valeur obtenue par l'échantillon et celle provenant de l'hypothèse est trop important pour être uniquement imputable au hasard de l'échantillonnage.
  - Pour déterminer entre quelles bornes peut varier la valeur obtenue par l'échantillon

Test d'hypothèse : exemple



- Évaluation de l'efficacité d'un nouveau médicament
- Expérience
  - L'objectif est de déterminer si le médicament est plus efficace qu'un placebo.
  - Cent patients sont répartis aléatoirement en deux groupes de 50 :
    - l'un reçoit le médicament ;
    - l'autre reçoit un placebo.
  - Après une période définie, l'intensité de la douleur est évaluée sur une échelle de 0 à 10.
- Hypothèses
  - Hypothèse nulle : Le médicament n'est pas plus efficace que le placebo.
  - Hypothèse alternative : Le médicament est plus efficace que le placebo
- Test
  - Les résultats des deux groupes sont recueillis.
  - Le paramètre est calculé dans chaque groupe, puis comparé.
  - Si l'écart entre les groupes dépasse le seuil fixé, l'hypothèse nulle est rejetée et le médicament est considéré comme plus efficace que le placebo.
  - Si l'écart n'atteint pas ce seuil, l'hypothèse nulle n'est pas rejetée et les données ne permettent pas de conclure à une efficacité supérieure du médicament.

Test de corrélation

- corrélation
  - mesure statistique indiquant dans quelle mesure deux variables sont liées, au niveau de l'échantillon
- test de corrélation
  - test calculant si deux variables sont significativement corrélées, au niveau de la population
  - principe du test d'hypothèse

Test de corrélation : exemple



- Contexte
  - Existe-t-il une relation entre le temps d'étude et les notes obtenues à un examen ?
- Étapes du test
  - Un échantillon aléatoire de 30 étudiants est tiré parmi les 500 de la classe.
  - Pour chaque étudiant, on relève :
    - le nombre d'heures d'étude ;
    - la note obtenue à l'examen.
  - Une corrélation de 85 % indique une forte association dans l'échantillon.
- Test d'hypothèse sur la corrélation
  - Inférence à la population
  - Hypothèse nulle (H0) : il n'existe pas de corrélation entre le temps d'étude et les notes.
  - Hypothèse alternative (H1) : il existe une corrélation entre le temps d'étude et les notes.
  - On calcule la valeur p.
  - Si $p < 5\ \%$, le résultat est statistiquement significatif. L'hypothèse nulle est rejetée.
  - Si $p \geq 5\ \%$, H0 n'est pas rejetée : les données ne permettent pas de conclure à une corrélation.
- Interprétation du test
  - p calculé = $4\ \% < 5\ \%$
  - L'hypothèse nulle est rejetée.
  - Les données montrent une relation significative entre le temps d'étude et les notes.