De la probabilité au test d'hypothèse

Ahmed Fouad EL HADDAD

IEP de Fontainebleau

October 28, 2025

Corrélation et causalité

La séance précédente portait sur la corrélation.

La corrélation mesure la **force** et la **direction** d'une association entre deux variables. Une corrélation élevée suggère une relation, mais ne signifie pas qu'il existe un lien de cause à effet.

Pourquoi la corrélation n'implique-t-elle pas la causalité ?

- Deux variables peuvent être corrélées en raison d'un **facteur tiers** (variable confondante).
- Parfois, la corrélation est simplement due au **hasard**.
- Même une forte corrélation peut être **spécieuse**, comme entre la consommation de glaces et le taux de noyades.

Une corrélation observable ne suffit donc pas à démontrer une relation causale. C'est là que la probabilité intervient : elle nous aide à tester nos observations.

De l'ajustement à la vérification : le rôle de la probabilité

La statistique consiste à ajuster un modèle aux données. Jusqu'ici, nous avons ajusté des fonctions linéaires pour détecter des relations. Mais comment savoir si ces relations sont réelles ou dues au hasard ?

Rôle des distributions de probabilité :

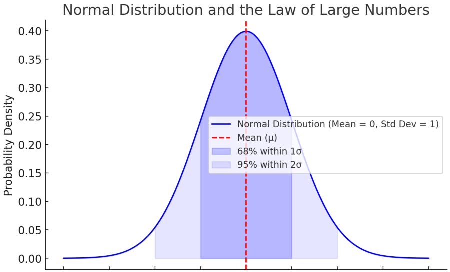
- Elles permettent d'évaluer si une relation observée est significative ou aléatoire.
- Elles modélisent le hasard et quantifient l'incertitude.

Comprendre le hasard est probablement la plus grande invention humaine après les frites... et Bourdieu.

La distribution normale et le hasard

Que signifie réellement dire qu'un phénomène est « aléatoire » ? Souvent, nous parlons de chance ou de malchance, mais en statistique, il s'agit d'une **régularité du hasard**.

Exemple : Imaginez un enseignant qui décide de tout faire à la main sans logiciel. Vous pouvez dire : « Pas de chance ! ». Mais la probabilité permet justement de **modéliser cette chance** et de savoir quand elle peut se produire.



Pourquoi la loi normale est-elle centrale ?

La distribution normale est au cœur de la statistique :

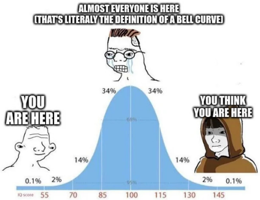
- Elle décrit de nombreux phénomènes naturels et sociaux (taille, revenus, erreurs de mesure…).
- Par le **théorème central limite**, la somme de nombreuses variables aléatoires indépendantes suit une loi normale.
- Elle sert de base à l'inférence statistique : beaucoup de tests reposent sur cette hypothèse.

Comprendre la normalité permet de mesurer les écarts par rapport aux attentes — c'est la base du **test d'hypothèse**.

De la normalité au test d'hypothèse

La loi normale permet de mesurer à quel point une observation est inhabituelle. Si un résultat est très improbable dans un monde « normal », c'est peut-être qu'un effet réel se cache derrière.

C'est le principe des **tests d'hypothèse** : distinguer le hasard d'un effet significatif.



Le test d'hypothèse : mesurer notre degré d'erreur

Idée essentielle : on ne prouve pas qu'une hypothèse est vraie, on mesure **à quel point on peut se tromper**.

- On part d'une hypothèse nulle H_0 : pas d'effet réel.
- On collecte des données et on se demande : « Si H_0 était vraie, quelle est la probabilité d'observer ce résultat ? »
- Si cette probabilité est très faible, on rejette H_0 en faveur d'une hypothèse alternative H_A.

Exemple : Un nouvel enseignement est testé.

- H_0 : la méthode n'a aucun effet.
- H_A : la méthode améliore les résultats.

Si l'augmentation observée est très improbable sous H_0, on conclut que la méthode fonctionne.

Tester la significativité d'une corrélation

Le **coefficient de corrélation** r mesure la force et la direction d'une relation linéaire. Mais une corrélation observée peut être due au hasard.

Objectif : déterminer si cette corrélation est **statistiquement significative**.

- H_0 : pas de corrélation dans la population ($\rho = 0$).
- H_1 : corrélation réelle ($\rho \neq 0$).

Pourquoi une statistique de test ?

Même une corrélation élevée peut être due à l'échantillon. Les petits échantillons amplifient les illusions statistiques.

Solution : transformer la corrélation en une statistique standardisée t, que l'on peut comparer à une distribution de référence.

La formule du test t

Pour évaluer la significativité d'une corrélation, on calcule une statistique appelée **t de Student** :

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- r : corrélation observée dans l'échantillon ;
- n : nombre d'observations ;
- Le dénominateur mesure la part de la relation qui reste inexpliquée.

Cette formule transforme une simple corrélation en une mesure **standardisée** que l'on peut comparer à une distribution théorique pour juger si elle est due au hasard ou à un effet réel.

Une petite histoire : William Gosset et la "Student's t"

La loi t n'a pas été inventée dans une université, mais dans une brasserie.

Au début du XX^{e} siècle, un chimiste anglais nommé **William Sealy Gosset** travaille pour la brasserie Guinness, à Dublin. Il cherche un moyen de tirer des conclusions fiables à partir de petits échantillons de bière (eh oui, les tests statistiques aussi peuvent naître dans les pubs).

Les lois normales existantes supposaient des échantillons très grands — or Gosset devait travailler sur des séries de dix ou vingt mesures. Pour corriger cette limite, il met au point une nouvelle distribution qu'il publie en 1908 sous le pseudonyme *"Student"* (pour éviter les contraintes de confidentialité imposées par son employeur).

C'est ainsi qu'est née la **distribution t de Student** — un outil qui permet encore aujourd'hui de raisonner rigoureusement avec de petits échantillons.

Comprendre le dénominateur $\sqrt{1 - r^2}$

Le dénominateur traduit la part de hasard dans la corrélation :

$$1 - r^2 = \text{la proportion de variation non expliquée.}$$

En termes simples :

- Si r est faible, beaucoup de choses échappent à la corrélation : le test devient plus prudent.
- Si r est fort (proche de ± 1), presque toute la variation est expliquée : le test devient plus affirmatif.

Ainsi, le test t ajuste automatiquement notre degré de confiance selon la force de la corrélation observée.

Le rôle de la taille d'échantillon

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Pourquoi $n - 2$?

- On retire deux **degrés de liberté**, car deux paramètres (les moyennes) sont estimés avant de mesurer la corrélation.
- Cela revient à dire : moins il y a d'observations, plus l'incertitude est grande.
- Quand n augmente, le terme $\sqrt{n-2}$ croît : on gagne en précision et la statistique t devient plus stable.

En somme, **plus l'échantillon est grand, plus le test est capable de détecter de petites corrélations réelles**.

Pourquoi la distribution t ?

La statistique t ne suit pas la loi normale, mais une loi **t de Student**.

Pourquoi ?

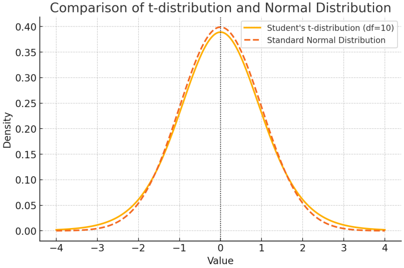
- Lorsque les échantillons sont petits, les estimations de la moyenne et de l'écart-type sont elles-mêmes incertaines.
- La loi t corrige cette incertitude : ses **queues sont plus épaisses**, ce qui signifie que les valeurs extrêmes sont un peu plus probables.
- Quand la taille d'échantillon augmente, cette incertitude diminue et la loi t **se rapproche progressivement de la loi normale**.

Sous H_0, la plupart des valeurs de t sont proches de 0 ; une valeur très grande (positive ou négative) indique un effet difficilement attribuable au hasard.

Comparer la loi normale et la loi t

Différences principales :

- La loi t a des **queues plus épaisses** : elle reflète davantage d'incertitude quand n est petit.
- La loi normale est une approximation valable uniquement pour des échantillons très grands.
- À mesure que n augmente, la loi t se resserre et finit par devenir pratiquement identique à la loi normale.



Moralité : la loi t, c'est la loi normale adaptée à la vie réelle — celle des

Que signifie la p-value ?

La **p-value** indique la probabilité d'obtenir une corrélation aussi forte que celle observée si, en réalité, il n'y en avait aucune (H_0 vraie).

Interprétation :

- Si $p < 0{,}05$: la probabilité d'un tel résultat sous H_0 est très faible $\rightarrow$ on rejette H_0.
- Si $p \geq 0{,}05$: le résultat reste compatible avec le hasard $\rightarrow$ on ne rejette pas H_0.

La p-value ne mesure pas la *force* du lien, mais la *crédibilité statistique* du résultat. Elle répond à la question : « Si le hasard seul gouvernait le monde, verrait-on souvent ce que j'observe ? »

Exemple : revenu et usage des transports publics

Hypothèses :

- H_0 : pas de corrélation entre revenu et usage du transport public.
- H_A : corrélation négative significative.

Résultats :

$$r = -0{,}45, \quad p = 0{,}03.$$

Interprétation : Puisque $p < 0{,}05$, on rejette H_0 : le revenu influence significativement l'usage du transport public.

Ce qu'il faut retenir

En résumé :

1. Calculer la corrélation.
2. Transformer en statistique t.
3. Lire la p-value et en déduire la significativité.

Idée clé : Les distributions de probabilité nous permettent d'estimer **à quel point nos observations peuvent être attribuées au hasard**. Elles sont le pont entre le **monde des données observées** et le **monde des hypothèses**.