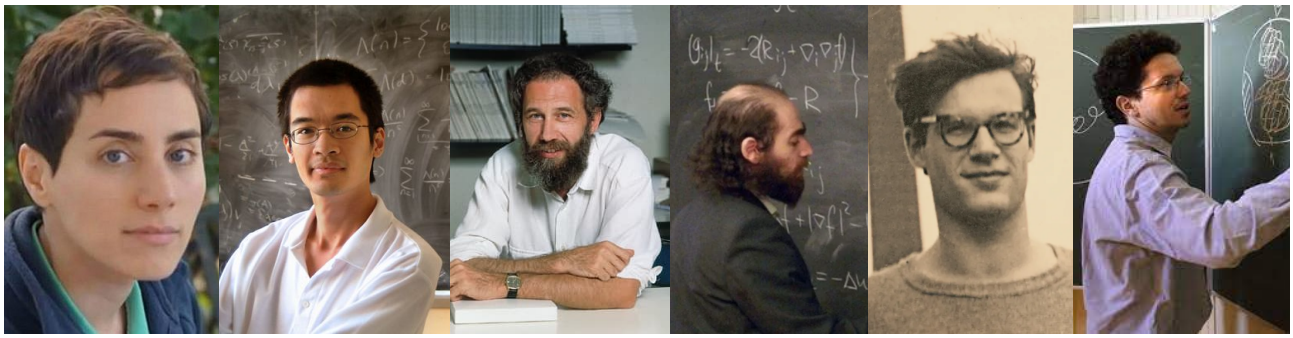
Statistiques et analyses de données

Antoine Géré

Année 2025 - 2026[1]



Ces notes sont en cours d'élaboration. Si vous avez la moindre question ou remarque ne pas hésiter à contacter par mail : a.gere@istom.fr.

Table des matières

1 Le degré de liberté pour les tests du χ^2 d'ajustement

Dans un test du χ^2 d'ajustement, le **degré de liberté** représente le nombre de modalités dont l'effectif peut varier librement avant que l'ensemble du système ne soit mathématiquement bloqué par les contraintes de l'expérimentation ou de l'enquête.

[1] version du 29 mai 2026

1.1 Situation usuelle

Imaginons un agronome travaillant avec une coopérative de producteurs au Burkina Faso pour s'adapter à la baisse des précipitations. Il met en place un essai comparatif avec *3 variétés de sorgho* ($k = 3$) :

- *Variété A* : Locale traditionnelle (cycle long).
- *Variété B* : Améliorée précoce (cycle court).
- *Variété C* : Hybride résistante au stress hydrique.

L'enveloppe budgétaire allouée par le projet permet de financer l'installation de parcelles de démonstration chez 30 paysans au total ($N = 30$). L'agronome doit répartir ces parcelles :

- Pour la variété A, il a un choix total : il décide de l'installer chez 12 paysans.
- Pour la variété B, il conserve une marge de manœuvre : il choisit de l'attribuer à 10 paysans.
- Pour la variété C ... **il n'a plus aucune liberté de choix**. Le total étant fixé à 30, la variété C est obligatoirement affectée aux $30 - (12 + 10) = 8$ paysans restants.

Règle du $k - 1$

Sur les k modalités initiales, l'expérimentateur n'a pu choisir librement l'effectif que pour $k - 1$ d'entre elles. La dernière est mathématiquement verrouillée par l'effectif total N.

$$ddl = k - 1$$

1.2 Situation générale

Dans un test d'ajustement du χ^2, la formule standard des degrés de liberté est

$$ddl = k - 1$$

où k est le nombre de modalitées comparées et 1 représente la contrainte du total général fixe. Cependant, cette relation suppose que la loi théorique de référence est parfaitement connue à l'avance.

Dans des situations plus complexes, l'agronome veut tester si ses données de terrain s'ajustent à une loi mathématique théorique (loi de Poisson, loi Normale, loi Binomiale) dont **il ne connaît pas les paramètres exacts au départ**.

Pour construire son tableau d'effectifs théoriques, il est alors obligé d'utiliser ses propres données observées pour **estimer** ces paramètres manquants. Chaque paramètre estimé (m) ajoute une contrainte mathématique au système et supprime une « liberté » supplémentaire. La relarion générale devient alors :

$$ddl = k - 1 - m \tag{1}$$

Exemple concret

Un entomologiste compte le nombre de larves d'insectes par pied de canne à sucre sur un échantillon de 200 plants. Il classe ses observations en $k = 5$ catégories : 0 larve, 1 larve, 2 larves, 3 larves, et 4 larves ou plus. Il souhaite vérifier si la répartition de ces parasites suit une **loi de Poisson**.

Le problème : Pour calculer les effectifs théoriques de la loi de Poisson, la formule mathématique exige de connaître la moyenne théorique (λ) du nombre de larves par pied. L'entomologiste ne la connaît pas.

L'estimation : Il utilise donc son échantillon de terrain pour calculer la moyenne observée. Imaginons qu'il trouve une moyenne de 1,2 larve par pied. Il utilise cette valeur unique comme estimation de λ.

- Il y a $k = 5$ catégories d'infestation.
- On retire 1 degré de liberté pour la contrainte du total des 200 plants.
- On retire 1 degré de liberté car on a extrait la moyenne des données pour utiliser le modèle théorique.

Le calcul final des degrés de liberté pour ce test est donc :

$$ddl = k - 1 - m = 5 - 1 - 1 = 3$$

2 Le degré de liberté pour les tests du χ^2 d'indépendance

Contrairement au test d'ajustement qui étudie une seule variable, le **test du χ^2 d'indépendance** analyse la relation entre **deux variables qualitatives**. Les données sont alors croisées dans un tableau à double entrée appelé **tableau de contingence**.

2.1 La règle mathématique

Pour un tableau de contingence comportant L lignes et C colonnes, la relation mathématique des degrés de liberté est :

$$ddl = (L - 1) \times (C - 1) \tag{2}$$

Cette relation représente le nombre de cases intérieures du tableau que l'on peut remplir de manière totalement indépendante avant que les totaux des lignes et des colonnes ne verrouillent automatiquement le reste des valeurs.

2.2 Illustration

Contexte

Un agronome réalise une enquête auprès de **100 caféiculteurs** au Honduras ($N = 100$). Il cherche à savoir si le **type de fertilisation** appliqué aux caféiers (Variable 1, en lignes) influence la **qualité commerciale du café** récolté (Variable 2, en colonnes).
Le tableau est structuré avec :

- **3 lignes ($L = 3$)** : Compost seul, Engrais chimique, Pratique mixte.
- **3 colonnes ($C = 3$)** : Standard, Premium, Spécialité.

Avant de commencer le calcul du χ^2, l'agronome calcule les totaux pour chaque ligne et chaque colonne à partir des données de son enquête. Ces totaux (appelés **les marges**) deviennent des **contraintes fixes et inviolables** pour la suite du calcul.

Remplissons le tableau ligne par ligne pour observer le moment exact où le système se bloque mathématiquement :

- **Sur la ligne « Compost » (Ligne 1) :** L'agronome commence à remplir les cases.
  - Il est **libre** de saisir une valeur pour la qualité *Standard* (ex : 10),
  - puis pour la qualité *Premium* (ex : 15).

  - En revanche, la troisième case (*Spécialité*) est **bloquée** : le total de la ligne devant impérativement faire 40, sa valeur est obligatoirement $40 - (10 + 15) = 15$.

- **Sur la ligne « Chimique » (Ligne 2) :**
  - Il est à nouveau **libre** d'attribuer des valeurs pour les deux premières colonnes (ex : 20 et 5).
  - La troisième case est immédiatement **bloquée** par la contrainte de ligne $(35 - (20 + 5) = 10)$.
- **Sur la ligne « Mixte » (Ligne 3) : Toutes les cases sont déjà bloquées**. En effet, les totaux de chaque colonne ont été fixés au départ.
  - Pour la colonne *Standard*, le total visé est 45. Ayant déjà posé 10 et 20 sur les lignes précédentes, la dernière case vaut obligatoirement $45 - (10 + 20) = 15$.
  - Il en va de même pour toutes les autres colonnes de cette ligne.

Voici la cartographie du tableau de contingence de l'enquête. Les cases vertes représentent les zones d'expression libre du hasard ou de l'agronome. Les cases rouges représentent les cellules mathématiquement contraintes par les marges.

Fertilisation \ Qualité	Standard	Premium	Spécialité	Total Lignes (Fixe)
Compost	10 (Libre)	15 (Libre)	15 (Bloqué)	40
Chimique	20 (Libre)	5 (Libre)	10 (Bloqué)	35
Pratique mixte	15 (Bloqué)	10 (Bloqué)	0 (Bloqué)	25
Total Colonnes (Fixe)	45	30	25	100 (N)

La zone de liberté forme un rectangle parfait de $2 \times 2 = 4$ cases en haut à gauche du tableau. C'est exactement le résultat fourni par notre relation :

$$ddl = (L - 1) \times (C - 1) = (3 - 1) \times (3 - 1) = 2 \times 2 = \mathbf{4}$$

3 Le degré de liberté pour les tests du χ^2 d'homogénéité

Le **test du χ^2 d'homogénéité** utilise la même configuration technique (tableau de contingence à double entrée) que le test d'indépendance. La seule différence réside dans le protocole de collecte : au lieu de tirer un seul grand échantillon, l'expérimentateur fixe à l'avance la taille de plusieurs sous-échantillons (groupes) pour vérifier si leurs comportements sont similaires.

Pour un tableau comportant L lignes (les groupes comparés) et C colonnes (les variables observées), la formule des degrés de liberté reste :

$$ddl = (L - 1) \times (C - 1) \tag{3}$$

4 Le degré de liberté pour l'ANOVA à 1 Facteur

Dans le cadre d'une **ANOVA à 1 facteur** (Analyse de Variance), la notion de degré de liberté (ddl) s'applique différemment par rapport au test du χ^2. L'ANOVA n'analyse pas des fréquences, mais cherche à comparer deux types de variabilités (variances) qui s'affrontent :

1. La variabilité **Inter-groupes** (due à l'effet du facteur).
2. La variabilité **Intra-groupes** (due à l'erreur résiduelle).

Pour effectuer cette comparaison, le système sépare mathématiquement les degrés de liberté totaux en deux sous-catégories distinctes.

Contexte pour l'illustration

Un agronome teste l'effet de la taille des branches sur le rendement de cacaoyers au Guatemala. Il dispose d'un échantillon total de $N = 12$ **parcelles homogènes**, réparties équitablement en $k = 3$ **groupes** expérimentaux (Taille Mensuelle, Taille Trimestrielle, et un Témoin sans taille). Chaque groupe possède donc un effectif de $n = 4$ parcelles de répétition.

4.1 Les degrés de liberté Inter-groupes, ddl_{Inter}

Les ddl_{Inter} sont directement associés à la **variabilité entre les groupes**. Ils mesurent combien de moyennes de traitements peuvent varier librement avant de verrouiller le système.

La relation est définie par :

$$ddl_{\text{Inter}} = k - 1 \tag{4}$$

La contrainte réside dans le fait que la moyenne des k groupes doit obligatoirement être ancrée autour de la **moyenne générale** globale de l'essai.

- La moyenne observée du Groupe A (Taille mensuelle) est libre (ex : 10 kg).
- La moyenne observée du Groupe B (Taille trimestrielle) est également libre (ex : 14 kg).
- La moyenne du Groupe C (Témoin)... **est automatiquement bloquée**. Sa valeur est mathématiquement forcée pour respecter la valeur de la moyenne générale.

Avec 3 groupes à comparer, le système ne dispose que de $3 - 1 = \mathbf{2}$ informations indépendantes.

4.2 Les degrés de liberté Intra-groupes, ddl_{Intra}

Les ddl_{Intra} (ou degrés de liberté de l'erreur résiduels) mesurent la **variabilité au sein de chaque groupe**. Ils comptabilisent la flexibilité des répétitions individuelles soumises à un *même* traitement.

La relation s'écrit :

$$ddl_{\text{Intra}} = N - k \tag{5}$$

Chaque groupe de l'essai fonctionne comme un sous-système fermé ayant sa propre contrainte : sa moyenne interne.

- Dans le Groupe A (comprenant $n = 4$ parcelles), la moyenne interne est fixée à 10 kg. Les 3 premières parcelles peuvent afficher des rendements aléatoires dictés par la nature (ex : 11, 9 et 12 kg). La 4$^{\text{e}}$ parcelle perd toute liberté : elle est mathématiquement forcée (à 8 kg) pour que la moyenne du groupe reste stable. Chaque sous-groupe perd ainsi exactement 1 degré de liberté.
- Liberté résiduelle dans le Groupe A : $4 - 1 = 3$ parcelles libres.
- Liberté résiduelle dans le Groupe B : $4 - 1 = 3$ parcelles libres.
- Liberté résiduelle dans le Groupe C : $4 - 1 = 3$ parcelles libres.

Sur l'ensemble du terrain, le hasard intra-groupe a pu s'exprimer sur $3 + 3 + 3 = \mathbf{9}$ parcelles indépendantes. L'application confirme le modèle : $N - k = 12 - 3 = \mathbf{9}$.

4.3 Bilan

L'ANOVA réalise une décomposition de la liberté globale du système :

$$ddl_{\text{Total}} = ddl_{\text{Inter}} + ddl_{\text{Intra}} \implies 11 = 2 + 9$$

Tableau de structure d'une ANOVA à 1 facteur

Source de variation	**Somme des Carrés**	**Degrés de liberté (ddl)**	**Carré Moyen (CM)**	**Statistique F**
Inter-groupes (Facteur)	SCE_{Inter}	$k-1=\mathbf{2}$	$CM_{\text{Inter}}=\frac{SCE_{\text{Inter}}}{2}$	$F_{\text{obs}}=\frac{CM_{\text{Inter}}}{CM_{\text{Intra}}}$
Intra-groupes (Erreur)	SCE_{Intra}	$N-k=\mathbf{9}$	$CM_{\text{Intra}}=\frac{SCE_{\text{Intra}}}{9}$	
Totale	SCE_{Totale}	$N-1=\mathbf{11}$		

Une Somme des Carrés (SCE) augmente artificiellement avec le nombre d'observations. Pour obtenir une variance pure et comparable (appelée **Carré Moyen**), le statisticien doit impérativement diviser chaque SCE par son propre nombre de degrés de liberté.

La statistique finale F_{obs} est le rapport de ces deux variances épurées.