Chapitre V. Liaisons et indépendance

La **liaison entre deux variables X et Y** exprime l'information que donne la connaissance de l'une pour la connaissance de l'autre. On distingue 3 types de liaisons:

- Indépendance (information nulle): la connaissance de la valeur x_i mesurée sur l'individu numéro i ne donne aucune information sur la valeur y_i.
- Liaison fonctionnelle (information totale): la connaissance de x_i permet de déterminer sans ambiguïté la valeur y_i (grâce à une fonction).
- Liaison partielle (information partielle): la connaissance de x_i donne une information incomplète sur y_i.

1) Notion d'indépendance entre deux variables

Définition: Les variables X et Y sont **indépendantes** si l'une des propriétés ci-dessous est vérifiée (les propriétés sont équivalentes donc si l'une est vérifiée, les autres le sont automatiquement).

1. Les distributions conditionnelles de X en fréquences sont égales entre elles, c'est-à-dire $n_{i1}/n_{\bullet 1} = n_{i2}/n_{\bullet 2} = \ldots = n_{iK'}/n_{\bullet K'}$, pour tout $i = 1, \cdots, K$.

2. Les distributions conditionnelles de Y en fréquences sont égales entre elles, c'est-à-dire $n_{1j}/n_{1\bullet} = n_{2j}/n_{2\bullet} = \ldots = n_{Kj}/n_{K\bullet}$, pour tout $j = 1, \cdots, K'$.

3. Les distributions conditionnelles de X (resp. Y) en effectifs sont proportionnelles entre elles.

4. Pour tout i et tout j, $n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$.

Remarques:

- On peut montrer que la propriété 1 (resp. propriété 2) implique que les distributions conditionnelles en fréquences de X (resp. de Y) sont toutes égales à la distribution marginale en fréquences de X (resp. de Y).
- La propriété 3 indique que les variables X et Y sont indépendantes si les distributions conditionnelles de X (ou de Y) en effectifs sont proportionnelles; cela revient à dire que les colonnes (ou les lignes) de la distribution conjointe en effectifs sont proportionnelles, puisque les distributions conditionnelles en effectifs sont effectivement les colonnes (pour X) ou les lignes (pour Y).
- La propriété 4 peut aussi s'exprimer au moyen des fréquences: pour tout i et tout j, $f_{ij} = f_{i\bullet} \times f_{\bullet j}$.

D'après la propriété 4, les variables X et Y sont indépendantes si, pour tous les couples de modalités (m_i, m'_j) ou pour toutes les cellules du tableau de contingence (distribution conjointe), les **effectifs observés** n_{ij} sont égaux aux quantités $\frac{n_{i\bullet} \times n_{\bullet j}}{n}$; il suffit que cette égalité ne soit pas vérifiée dans une seule cellule pour que les deux variables ne soient pas indépendantes.
La quantité $\frac{n_{i\bullet} \times n_{\bullet j}}{n}$ est donc l'effectif qu'on devrait observer pour que X et Y soient indépendantes; on l'appelle l'effectif d'indépendance ou **effectif théorique (d'indépendance)** de la modalité (m_i, m'_j), et on le note $\tilde{n}_{ij}$ ("n tilde i j").

Le tableau de contigence contenant les effectifs théoriques $\tilde{n}_{ij}$ s'appelle **tableau de contigence sous hypothèse d'indépendance**.

2) Test d'indépendance du Khi2

Distribution théorique d'indépendance

Notons D la distribution conjointe (observée) de X et Y.

Définition: La **distribution théorique d'indépendance d'une distribution conjointe** D est la distribution conjointe notée $\tilde{D}$ dont les effectifs sont les effectifs théoriques $\tilde{n}_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$; c'est la distribution qu'on devrait observer si X et Y étaient indépendantes. $\tilde{D}$ a les mêmes marges que D, et ses distributions conditionnelles en fréquence sont égales.

D'après la propriété 4, X et Y sont donc indépendantes si la distribution conjointe D est égale à sa distribution théorique d'indépendance $\tilde{D}$.

Le **taux de liaison d'un couple de modalités** (m_i, m'_j) mesure l'écart (relatif) entre l'effectif observé et l'effectif qu'on devrait observer sous hypothèse d'indépendance (l'effectif théorique). Sa valeur est:

$$t_{ij} = \frac{n_{ij} - \tilde{n}_{ij}}{\sqrt{\tilde{n}_{ij}}}.$$

Statistique du χ^2 : distance entre distribution observée et distribution théorique (d'indépendance).

Dans la pratique, la distribution observée D n'est presque jamais identique à la distribution d'indépendance $\tilde{D}$, même quand on sait que X et Y sont indépendantes. Ceci est dû aux "fluctuations d'échantillonnage": les "effets du hasard" font que, même pour des variables en théorie indépendantes dans une population, les observations issues de ces variables (qui sont mesurées sur un échantillon pris au hasard) ont une distribution D qui n'est pas exactement la même que la distribution d'indépendance $\tilde{D}$. Pour étudier l'indépendance de X et Y, nous allons donc être conduits à juger de la proximité de D avec $\tilde{D}$. La statistique du χ^2 est une mesure de l'écart entre une distribution conjointe observée et sa distribution théorique d'indépendance. Sa valeur est la somme des carrés des taux de liaisons:

$$\chi^2(D) = \sum_{i=1}^{K}\sum_{j=1}^{K'} t_{ij}^2 = \sum_{i=1}^{K}\sum_{j=1}^{K'} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}$$

Les $K \times K'$ termes positifs ou nuls $t_{ij}^2 = \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}$ s'appellent **contributions (des couples de modalités (m_i, m'_j)) au χ^2** .

$\chi^2(D)$ est nul si et seulement si $D = \tilde{D}$. En effet, $\chi^2(D)$ étant une somme de nombres positifs ou nuls, il ne peut s'annuler que si tous les termes sont nuls, autrement dit si les effectifs observés n_{ij} sont égaux aux effectifs théoriques $\tilde{n}_{ij}$. Cette dernière remarque permet d'énoncer une cinquième définition équivalente de l'indépendance: X et Y sont indépendantes si $\chi^2(D) = 0$.

Comme on l'a évoqué précédemment, en pratique il est très rare d'avoir $\chi^2(D)$ nul pour une distribution D observée sur un échantillon pris au hasard, même si les variables X et Y sont réellement indépendantes dans la population. On va donc introduire une notion d'indépendance moins stricte, l'indépendance statistique.

X et Y sont dites **statistiquement indépendantes** si les écarts entre les n_{ij} et $\tilde{n}_{ij}$ sont "petits" et peuvent être considérés comme "l'effet du hasard induit par l'échantillonnage". Ceci équivaut à dire que X et Y sont statistiquement indépendantes si et seulement si le χ^2 de la distribution est "petit" et peut être considéré comme une conséquence des "fluctuations de l'échantillonnage".

Le **test du χ^2 d'indépendance** vise à décider si X et Y peuvent être considérées comme statistiquement indépendantes dans une population, à partir de leur mesure conjointe sur un échantillon.

On souhaite donc tester l'hypothèse:

H: Les variables X et Y sont statistiquement indépendantes.

On calcule le χ^2 de la distribution conjointe observée D, et on voit si cette valeur est "suffisament petite".

Si X et Y sont indépendantes, la distribution des distances du χ^2 d'échantillons choisis au hasard est une loi connue: la distribution (théorique) du χ^2; elle dépend du nombre $(K-1)\times(K'-1)$, son degré de liberté (ddl). Cette distribution est généralement donnée par des quantiles dans une table (voir table du χ^2 sur le site internet).

Par exemple, si $K = K' = 2$, on a un ddl de $(2-1)\times(2-1) = 1$ et le 95ème centile est égal à 3.8415, le 99ème à 6.6349; cela signifie que si on mesurait (X, Y) sur un grand nombre d'échantillons, sous les hypothèses que X et Y sont indépendantes dans la population et les échantillons sont "choisis au hasard", alors 5% des échantillons auraient une valeur du χ^2 (dont la formule est donnée en bas de la page 6) supérieure à 3.8415, 1% supérieure à 6.6349

Procédure pour appliquer le test d'indépendance du χ^2

1. Calcul de $\chi^2(D)$, la distance du χ^2 de la distribution conjointe observée D (calcul du tableau des effectifs théoriques $\tilde{n}_{ij}$, calculs et addition des contributions), puis de son degré de liberté $ddl = (K-1) \times (K'-1)$.

2. Choix d'un quantile d'ordre $1-\alpha$, noté $q_{1-\alpha}$, de la distribution théorique du χ^2 comme seuil de la décision. On prendra généralement le 95ème ($q_{.95}$) ou le 99ème ($q_{.99}$) centile ($1-\alpha = 0.95$ ou $1-\alpha = 0.99$).

3. Calcul de la valeur $l_{ddl}(1-\alpha)$ du quantile $q_{1-\alpha}$ à partir de la table du χ^2 et du degré de liberté ddl.

4. Prise de décision en comparant $\chi^2(D)$ à $l_{ddl}(1-\alpha)$:
   - Si $\chi^2(D) \geq l_{ddl}(1-\alpha)$, on rejette l'hypothèse **H** d'indépendance de X et Y dans la population, en les considérant comme liées.
   - Si $\chi^2(D) < l_{ddl}(1-\alpha)$, on ne rejette pas l'hypothèse d'indépendance en considérant comme plausible qu'elles le soient.

Remarque: Pour le choix du quantile de la distribution du χ^2, $q_{1-\alpha} = l_{ddl}(1-\alpha)$, la quantité α est appelée l'erreur de 1ère espèce et représente la probabilité de se tromper lorsqu'on rejette l'hypothèse d'indépendance **H** (alors que celle-ci est en fait vraie).

Exemple: Parmi un groupe de 200 malades qui se plaignent de ne pas bien dormir, certains ont pris un somnifère sous forme de cachet, d'autres ont pris un cachet de sucre; tous pensaient prendre un somnifère. Après la nuit, on leur a demandé si le cachet avait été efficace. Le tableau suivant donne la répartition des réponses (on suppose que tous les malades ont dit la vérité) :

Réponse	**Ont bien dormi**	**N'ont pas bien dormi**
Somnifère	52	12
Sucre	96	40

On calcule tout d'abord les effectifs marginaux:

X\Y	**Ont bien dormi**	**N'ont pas bien dormi**	Total X
Somnifère	52	12	64
Sucre	96	40	136
Total Y	148	52	n=200

Les effectifs théoriques $\tilde{n}_{ij}$ sont donnés dans le tableau ci-dessous:

X\Y	**Ont bien dormi**	**N'ont pas bien dormi**
Somnifère	47.36	16.64
Sucre	100.64	35.36

D'où la valeur de la distance du χ^2:

$$\begin{aligned}\chi^2(D) &= \frac{(52-47.36)^2}{47.36} + \frac{(96-100.64)^2}{100.64} + \frac{(12-16.64)^2}{16.64} + \frac{(40-35.36)^2}{35.36} \\ &= 2.57\end{aligned}$$

Ici le degré de liberté (ddl) est: ddl $= (2-1) * (2-1) = 1$. Pour $\alpha = 5\%$ par exemple, d'après la table du χ^2, le quantile est $l_1(95\%) = 3.8415$. On ne rejette donc pas l'hypothèse d'indépendance **H** et on considère qu'il est possible que les variables soient indépendantes.

3) Cas de deux variables quantitatives: le coefficient de corrélation linéaire

Dans ce paragraphe nous étudions la liaison entre deux variables quantitatives. On peut, bien entendu, appliquer aux deux variables les procédures développées dans le paragraphe précedent. La particularité est que chaque observation étant un couple de nombres (x_i, y_i), elle peut être représentée graphiquement par un point d'un plan; on peut alors faire appel à des procédures géométriques pour visualiser l'ensemble des observations, et analyser la forme du nuage ainsi formé. Pour étudier l'existence d'une liaison entre les deux variables, nous introduisons la notion de **covariance** (d'un couple de variables) puis le **coefficient de corrélation linéaire**, un indice de covariation linéaire des deux variables.

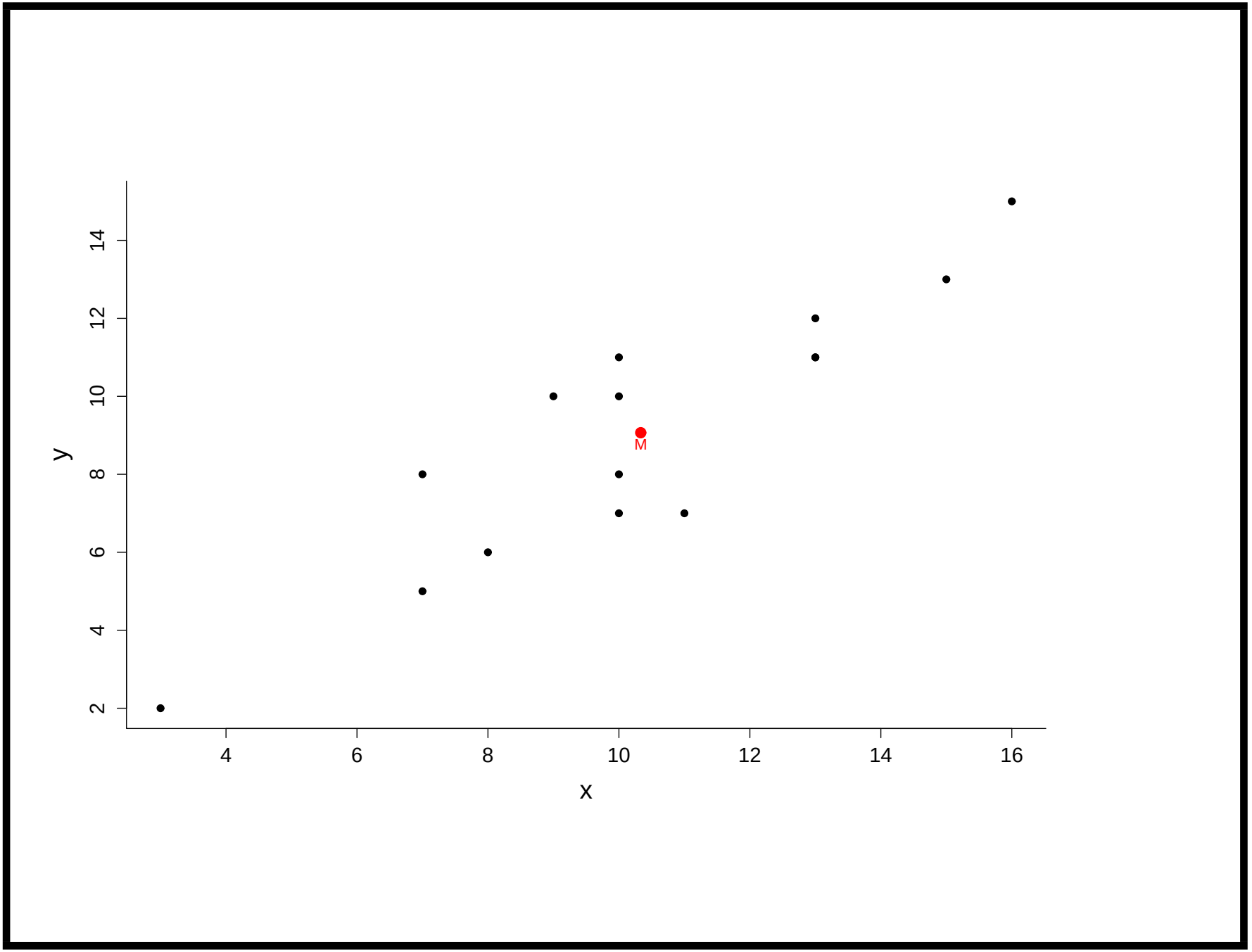
Nuage des observations

Sur un plan, on place un repère orthonormé: l'axe horizontal gradué des abscisses qui identifie les valeurs de la variable X, et l'axe vertical gradué des ordonnées qui identifie celles de la variable Y; l'observation de l'individu numéro i est représentée par le point d'abscisse x_i et d'ordonnée y_i. L'ensemble des points est le nuage des observations. Le point "moyen" M de coordonnées $\bar{x}$ et $\bar{y}$ est le centre de gravité du nuage.

Exemple: Notes de partiel et de TD

Dans le tableau ci-dessous, on a relevé les notes de TD (**X**) et les notes de partiel (**Y**) obtenues en statistique par un groupe de 15 étudiants.

X	8	13	10	10	9	15	16	13	10	13	7	10	11	3	7
Y	6	11	11	7	10	13	15	12	10	11	8	8	7	2	5


y
2
4
6
8
10
12
14
x
4
6
8
10
12
14
16
M

Covariance d'un couple de variables (X, Y)

La **covariance** d'une série de n couples d'observations (x_i, y_i) de deux variables quantitatives X et Y est donnée par la formule suivante :

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{n}$$

On peut montrer qu'elle est égale à "la moyenne des produits moins le produit des moyennes":

$$cov(x, y) = \frac{\sum_{i=1}^{n} x_i \times y_i}{n} - \bar{x} \times \bar{y}$$

Lorsque les résultats sont fournis sous forme de tableau de contingence et non de données brutes, la première formule peut s'écrire comme ceci:

$$cov(x, y) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}(m_i - \bar{x}) \times (m'_j - \bar{y})}{n}$$

Si la variable X (resp. Y) est continue, on remplace la modalité m_i (resp. m'_j) par le centre de la $ième$ (resp. $jème$) classe pour tout $i = 1, ..., K$ (resp. tout $j = 1, ..., K'$).

Coefficient de corrélation linéaire d'un couple de variables (X, Y)

Le **coefficient de corrélation linéaire** d'une série de n couples d'observations de deux variables X et Y est le rapport de leur covariance par le produit de leur écart-type; on le note r et sa formule est:

$$r(x, y) = \frac{cov(x, y)}{\sigma(x)\sigma(y)}$$

r est un nombre sans dimension (il ne dépend pas des unités de mesure), et on peut montrer qu'il est compris entre -1 et 1. Lorsque r est proche 1 ou -1, on dit que les variables sont linéairement corrélées.

Si $r > 0$, il peut exister une liaison linéaire (positive) entre X et Y, d'autant plus forte que r est proche de 1. Si $r < 0$, il peut exister une liaison linéaire (négative) entre X et Y, d'autant plus forte que r est proche de -1.

Remarque: Si les variables X et Y sont indépendantes, la covariance est proche de 0, ce qui entraîne que r est lui aussi proche de 0. Dans ce cas, on dit que les variables sont non corrélées linéairement: l'indépendance entraîne la non corrélation linéaire. La réciproque est fausse: une valeur de r proche de 0 n'entraîne pas toujours l'indépendance. En conclusion, la non corrélation linéaire (un r proche de 0) équivaut à dire que X et Y ne sont pas liées par une relation linéaire, et non que X et Y ne sont pas liées.

Exemple: Notes de partiel et de TD.

Pour calculer la covariance (et ensuite le coefficient de corrélation linéaire), nous avons besoin des moyennes de X et Y: $\bar{x} = 10.333$ et $\bar{y} = 9.067$.
La covariance est alors égale à: $cov(x, y) = 9.578$.
Les variances sont calculées ci-dessous:

$$\text{var}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = 10.622 \quad \text{et} \quad \text{var}(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2 = 10.596.$$

D'où les valeurs suivantes pour les écart-types (racine carrée de la variance):
$\sigma(x) = 3.259$ et $\sigma(y) = 3.255$.

A partir de la covariance et des écart-types, on en déduit finalement le coefficient de corrélation linéaire:

$$r(x, y) = 0.903.$$

Le coefficient étant proche de 1, on a donc une corrélation linéaire assez forte entre X et Y. On peut également dire qu'en moyenne les notes de partiel et de TD évoluent dans le même sens.

4) Cas d'une variable quantitative et d'une variable qualitative: Analyse de la variance

Dans ce paragraphe, nous supposons que X est une variable qualitative et Y une variable quantitative. Nous étudions alors comment l'analyse de la variance de Y permet de tester l'égalité des moyennes conditionnelles de cette variable dans les sous-populations induites par X; dans cette problématique, X est appelée la variable explicative, ou le facteur explicatif, et Y la variable expliquée.

Formule de décomposition de la variance: la variance de Y (variance globale de Y) peut se calculer à partir des variances et des moyennes conditionnelles de Y par la formule de décomposition de la variance. En effet la variance globale des observations de l'échantillon (ou variance de Y) est la somme de deux quantités:

- La variance intra: la moyenne des variances conditionnelles pondérée par la taille des K sous-populations. Elle quantifie la part de la variabilité intrinsèque de Y dans la variance globale.
- La variance inter: la variance des moyennes conditionnelles également pondérée par la taille des sous-populations. Elle mesure l'hétérogénéité des sous-populations.

On a donc: $\text{Var}(Y) := \sigma^2(y) = \text{Var intra} + \text{Var inter}$, où

$$\text{Variance intra} = \frac{1}{n}\sum_{i=1}^{K} n_{i\bullet}\sigma^2(y)_i$$

et

$$\text{Variance inter} = \frac{1}{n}\sum_{i=1}^{K} n_{i\bullet}(\bar{y}_i - \bar{y})^2$$

En termes synthétiques la décomposition de la variance s'énonce "Variance totale = Variance intra + Variance inter", ou encore "Variance totale = moyenne des variances + variance des moyennes".

Test de l'égalité des moyennes par l'analyse de la variance

Soit

$$T = (n - K) \times \frac{\text{Var inter}}{\text{Var intra}}$$

calculée à partir de l'observation d'un échantillon de taille n, K étant le nombre de sous-populations (le nombre de modalités de X).
On considère l'hypothèse suivante:
H: Les moyennes de Y sont identiques dans les K sous-populations (c'est-à-dire que le facteur X n'a pas d'effet global sur Y).

Si **H** est vraie, on sait (par des travaux mathématiques) que les valeurs de T varient approximativement comme la distribution d'une χ^2 à $K - 1$ degrés de liberté (ddl).

Si **H** est vraie, les valeurs de T calculées sur différents échantillons vont varier au voisinage de 0, puisque les variances inter seront proches de 0 (mais pas exactement égales à 0 à cause des fluctuations d'échantillonnage). Le principe du test est identique à celui du χ^2 pour l'indépendance statistique :

1. Calcul de la valeur t de T pour l'échantillon observé.

2. Choix d'un seuil s au-delà duquel les valeurs de T sont considérées comme improbables si l'hypothèse $\mathbf{H}$ est vraie. Pour ce faire on utilise une table de la distribution théorique du χ^2 qui est la distribution de T dans ce cas (ddl=$K-1$) en choisissant l'ordre $1-\alpha$ du quantile (comme pour le test d'indépendance du χ^2).

3. Si t fait partie des valeurs "improbables", c'est à dire $t \geq s$, de deux choses l'une: ou bien $\mathbf{H}$ est fausse (les moyennes de Y sont différentes dans les sous-populations), et il n'y a rien d'étonnant à ce que t soit une "grande" valeur; ou bien $\mathbf{H}$ est vraie et l'échantillon observé est un des rares échantillons atypiques qui donnent un t élevé sous l'hypothèse $\mathbf{H}$; on choisira donc de rejeter l'hypothèse $\mathbf{H}$, en considérant que le facteur X a un effet global sur Y.

4. Si t ne fait pas partie des valeurs "improbables", c'est à dire $t < s$, on ne rejette pas l'hypothèse $\mathbf{H}$, et on considére qu'il est possible que le facteur X n'ait pas d'effet global sur Y.