

Modélisation

(première partie)

Expériences aléatoires

Expériences aléatoires

Une expérience aléatoire est une expérience qu'on peut réaliser de multiples fois, et dont le résultat peut varier d'une fois à l'autre.

Exemples :

- faire rouler un dé
- prélever un parisien « au hasard » (le mesurer...);
- lancer une flèche sur cible (de suffisamment loin)...

Les jeux de hasard ont joué un rôle historique dans le développement de la théorie des probabilités.

Le premier problème du Chevalier de Méré

« Pourquoi est-il avantageux de parier qu'on va « sortir » un six en lançant quatre fois le dé, alors qu'il ne l'est pas de parier qu'on va sortir un double six en lançant vingt-quatre fois deux dés ? »

👉 Une définition possible de la **probabilité** de gagner est la proportion de parties gagnantes « à très long terme » (à l'infini...)

Avant le début « officiel » du développement de la théorie, Méré raisonnait sur la base du fait : la probabilité de faire six avec un dé est $\frac{1}{6}$.

Description d'expériences aléatoires répétées

Distinguer

- les mesures quantitatives discrètes (nombre d'enfants)
- les mesures quantitatives continues (stature ou poids d'un sujet)
- les mesures qualitatives (sexe, lieu de naissance, couleur des yeux)

Certaines mesures peuvent être mixtes : p. ex. un taux d'anticorps mesurés chez des individus exposés à un agent infectieux : chez certains patients cette mesure sera nulle, chez d'autres, elle sera positive.

👉 rapporter la proportion d'individus chez lesquels la mesure est nulle ;
décrire la distribution des mesures strictement positives.

Une mesure qualitative

Tables d'effectifs, ou de proportions empiriques (en précisant la taille de l'échantillon), ou les deux à la fois :

Marrons	Verts	Bleus	Total
90 (60%)	40 (27%)	20 (13%)	150

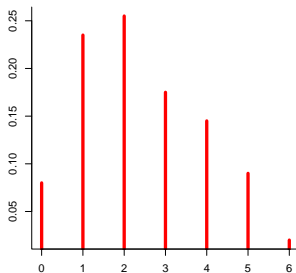
Effectifs (proportions) des différentes couleurs d'yeux

Une mesure quantitative discrète

Table ou diagramme « en bâton ».

0	1	2	3	4	5	6
16 (8%)	47 (23.5%)	51 (25.5%)	35 (17.5%)	29 (14.5%)	18 (9%)	4 (2%)

Nombre d'enfants par couple (total 200 couples)



Une mesure quantitative discrète

Mesures de localisation

La moyenne est

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n).$$

On utilisera également la médiane (définie plus loin).

Mesures de dispersion

L'écart absolu moyen est la moyenne des écarts absolus

$$e_a = \frac{1}{n} (|x_1 - \bar{x}| + \cdots + |x_n - \bar{x}|).$$

Une mesure quantitative discrète

Mesures de localisation

La moyenne est

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n).$$

On utilisera également la médiane (définie plus loin).

Mesures de dispersion

On utilise plus souvent la **variance** et l'**écart-type**.

La variance est la moyenne des carrés des écarts quadratiques

$$\tilde{S}^2 = \frac{1}{n} \left((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \right) = \frac{1}{n} \left(x_1^2 + \cdots + x_n^2 \right) - \bar{x}^2.$$

L'écart-type est la racine carrée de la variance.

Une mesure quantitative continue

Dès qu'il y a trop de mesures, impossible de les tabuler toutes.

👉 discrétiser les mesures, en créant des classes

[145, 150]	(150, 155]	(155, 160]	(160, 165]	(165, 170]
4	19	33	34	32
(170, 175]	(175, 180]	(180, 185]	(185, 190]	
29	25	17	7	

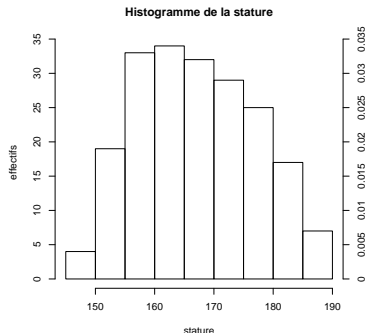
répartition de 200 statures en 9 classes de largeur 5 centimètres

Une mesure quantitative continue

Dès qu'il y a trop de mesures, impossible de les tabuler toutes.

Cette table se représente naturellement par un histogramme.

La hauteur des rectangles est (proportionnelle à) l'effectif dans chaque classe. Si l'aire totale est 1, c'est une approximation de la densité de la variable.

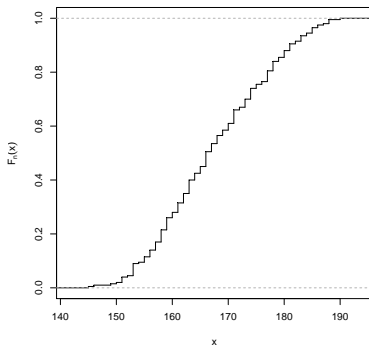


Une mesure quantitative continue

Fonction de répartition empirique

C'est la fonction $F_n(x) =$ proportion de mesures inférieures ou égales à x .

C'est une fonction « en escalier » dont les marches ont une hauteur multiple de $\frac{1}{n}$



Une mesure quantitative continue

Quantiles

En plus de la moyenne et la variance (ou l'écart-type), on rapportera certains quantiles de l'échantillon.

La **médiane** $x_{0,5}$ est le quantile de niveau 0,5, c-à-d que $F_n(x_{0,5}) = 0,5$: on a la moitié des mesures $\leq x_{0,5}$.

Les autres quantiles sont définis de façon analogue : x_α vérifie $F_n(x_\alpha) = \alpha$: une proportion α des mesures est $\leq x_\alpha$.

On rapporte fréquemment le **premier** et le **troisième quartiles**, qui sont respectivement $x_{0,25}$ et $x_{0,75}$.

Une mesure quantitative continue

Quantiles et boîtes à moustaches

L'**écart inter-quartile** est une mesure de dispersion définie par :

$$\text{IQR} = x_{0,75} - x_{0,25}.$$

La **boîte à moustaches** ou **boîte de Tukey** est une façon compacte de représenter une distribution.

Les bords de la boîte vont du premier au troisième quartile ; dans la boîte, un trait montre la position de la médiane.



Une mesure quantitative continue

Quantiles et boîtes à moustaches

L'**écart inter-quartile** est une mesure de dispersion définie par :

$$\text{IQR} = x_{0,75} - x_{0,25}.$$

La **boîte à moustaches** ou **boîte de Tukey** est une façon compacte de représenter une distribution.

La convention la plus fréquente est que les moustaches finissent aux dernières mesures à une distance $< 1,5 \times \text{IQR}$ du bord de la boîte. Les mesures qui tombent en dehors des moustaches sont des mesures « exceptionnelles » (**outliers**), on les représente par des points.



Coefficients d'asymétrie et d'aplatissement

Le k^{e} moment centré est

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k.$$

Le 2^e moment centré m_2 est donc la variance \tilde{S}^2 ,

Le coefficient d'asymétrie (ou *skewness*) est

$$\gamma_1 = \frac{m_3}{m_2^{3/2}} = \frac{1}{n} \sum_i \left(\frac{x_i - \bar{x}}{\sqrt{m_2}} \right)^3.$$

Si $\gamma_1 > 0$, asymétrie à droite.

Coefficients d'asymétrie et d'aplatissement

Le k^{e} moment centré est

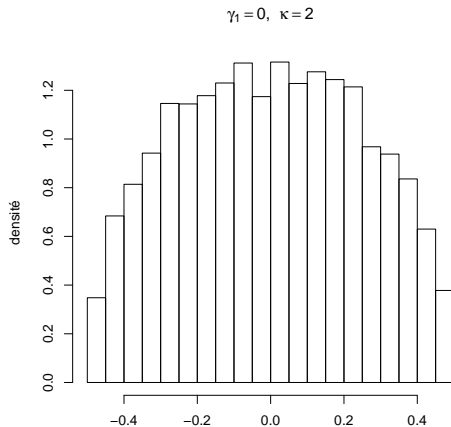
$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k.$$

Le 2^e moment centré m_2 est donc la variance \tilde{S}^2 ,

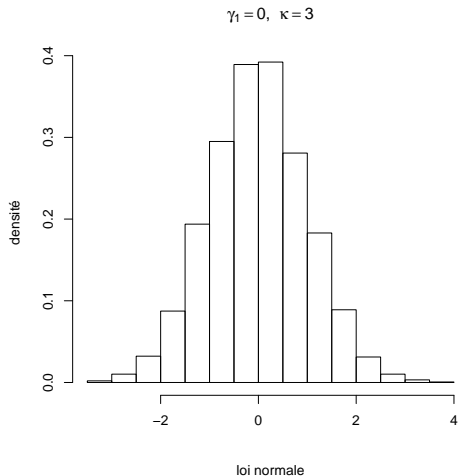
Le coefficient d'aplatissement (ou *kurtosis*) est

$$\kappa = \frac{m_4}{m_2^2} = \frac{1}{n} \sum_i \left(\frac{x_i - \bar{x}}{\sqrt{m_2}} \right)^4.$$

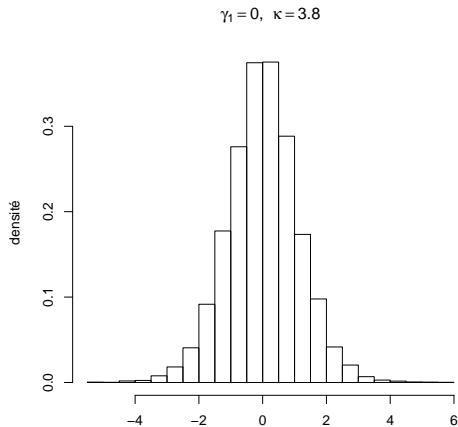
Coefficients d'asymétrie et d'aplatissement



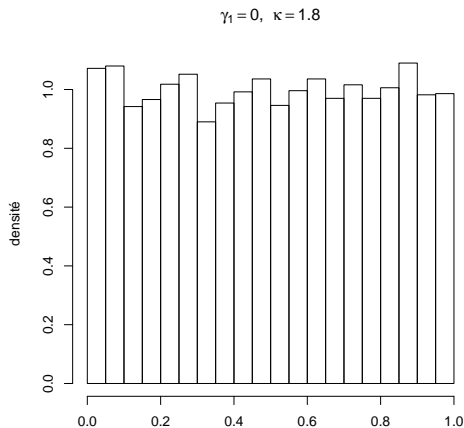
Coefficients d'asymétrie et d'aplatissement



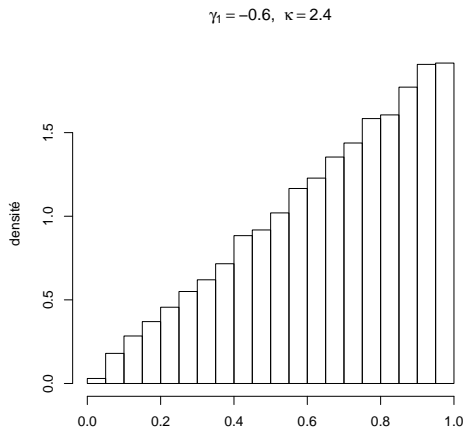
Coefficients d'asymétrie et d'aplatissement



Coefficients d'asymétrie et d'aplatissement

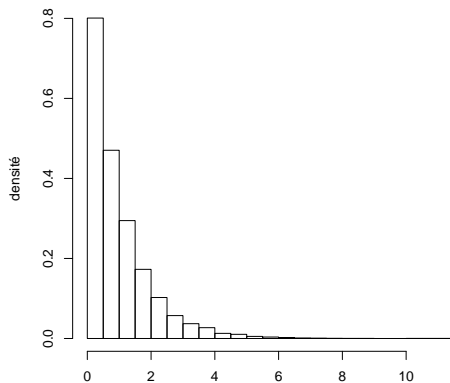


Coefficients d'asymétrie et d'aplatissement



Coefficients d'asymétrie et d'aplatissement

$\gamma_1 = 2.2$, $\kappa = 10.7$



Deux mesures qualitatives ou discrètes

Quand on effectue deux mesures sur une même expérience aléatoire (par exemple sur un même individu tiré au hasard), on peut dresser une **table de contingence** :

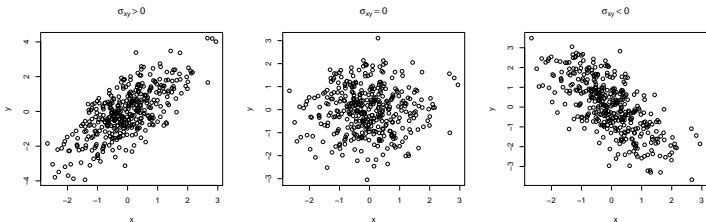
	Marrons	Verts	Bleus
Hommes	44	24	10
Femmes	46	16	10

Les effectifs observés pour chacune des deux variables peut être retrouvée en faisant la somme des colonnes ou des lignes de la table (on parle d'effectifs marginaux).

Deux mesures continues

Covariance, corrélation

Un **nuage de points** permet de se faire une idée de la distribution des mesures.



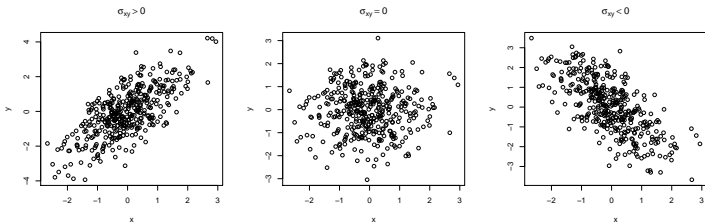
L'allure du nuage de points est reliée au signe de la **covariance** entre les x_i et les y_i

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})$$

Deux mesures continues

Covariance, corrélation

Un **nuage de points** permet de se faire une idée de la distribution des mesures.



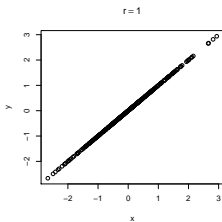
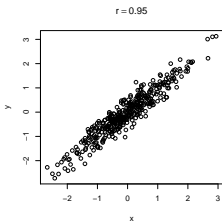
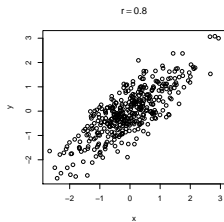
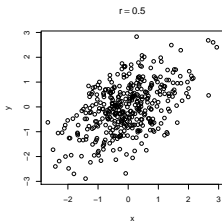
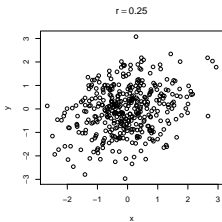
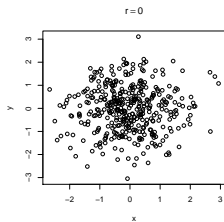
La **corrélation** est la covariance divisée par le produit des écart-types

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Elle est sans unité et est comprise entre -1 et 1.

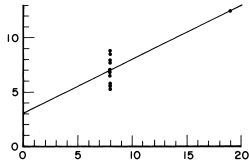
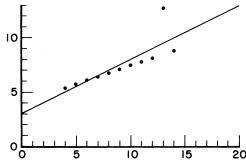
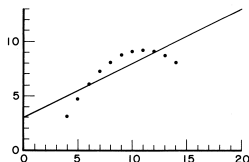
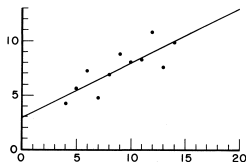
Deux mesures continues

Covariance, corrélation



Deux mesures continues

Attention !

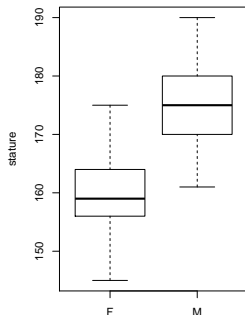


Quatre jeux de données avec $\bar{x} = 9$, $\bar{y} = 7,5$, $S_x^2 = 10$, $S_y^2 = 3,75$, et $\sigma_{xy} = 5$.

Anscombe 1973, Graphs in Statistical Analysis

Une mesure qualitative ou discrète, une mesure continue

On décrit la distribution de la mesure continue pour chacun des niveaux de la mesure qualitative (ou chaque valeur de la mesure discrète).



Rappels de probabilité

Espaces probabilisés

On note Ω l'ensemble des résultats possibles de l'expérience aléatoire.

Les parties de Ω sont appelés des **événements**.

La probabilité d'un événement A est la probabilité qu'une expérience « tombe » dans A , on écrit $\mathbb{P}(A)$ au lieu de $\mathbb{P}(\omega \in A)$.

Deux événements A et B sont **incompatibles** si $A \cap B = \emptyset$.

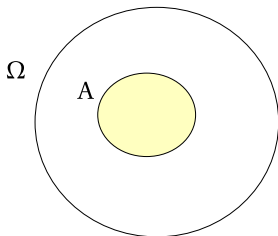
Une **probabilité** sur Ω est une fonction \mathbb{P} de l'ensemble des événements à valeurs dans $[0, 1]$, qui vérifie

- $\mathbb{P}(\Omega) = 1$
- Si A_1, A_2, \dots , sont deux à deux incompatibles, alors

$$\mathbb{P}(A_1 \cup A_2 \cup \dots) = \sum_{k=1}^{+\infty} \mathbb{P}(A_k).$$

Exemple : la cible

La cible est un bon modèle pour raisonner sur les probabilités. On suppose que tous les points sont également susceptibles d'être atteints par le tireur.

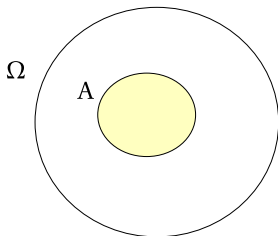


La probabilité de l'événement A est le rapport des surfaces de A et de Ω :

$$\mathbb{P}(A) = \frac{S(A)}{S(\Omega)}$$

Exemple : la cible

La cible est un bon modèle pour raisonner sur les probabilités. On suppose que tous les points sont également susceptibles d'être atteints par le tireur.

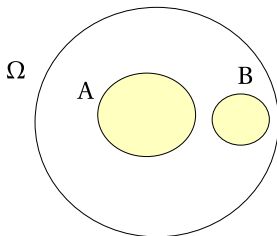


On a bien

$$\mathbb{P}(\Omega) = \frac{S(\Omega)}{S(\Omega)} = 1$$

Exemple : la cible

Deux événements A et B sont incompatibles si le résultat d'une expérience ne peut pas être à la fois dans A et dans B :

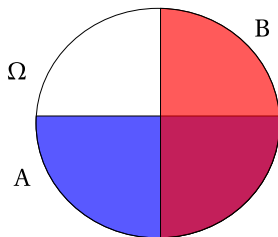


Dans ce cas, on a bien

$$\mathbb{P}(A \cup B) = \frac{S(A \cup B)}{S(\Omega)} = \frac{S(A) + S(B)}{S(\Omega)} = \frac{S(A)}{S(\Omega)} + \frac{S(B)}{S(\Omega)} = \mathbb{P}(A) + \mathbb{P}(B)$$

Événements indépendants

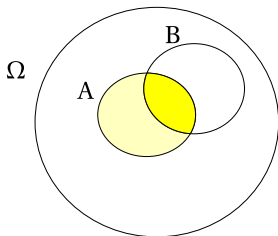
Deux événements A et B sont **indépendants** si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.



Ici, A = moitié inférieure, et B = moitié droite ;
on a $\mathbb{P}(A) = 0,5$, $\mathbb{P}(B) = 0,5$, et $\mathbb{P}(A \cap B) = 0,25$.

Probabilités conditionnelles

On note $\mathbb{P}(B|A)$ la probabilité qu'une expérience réalisée soit dans B , sachant qu'elle est dans A .

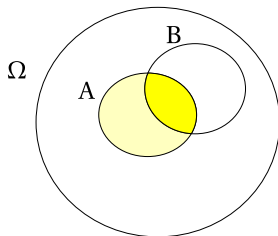


On fait le rapport des surfaces de $A \cap B$ et de A :

$$\mathbb{P}(B|A) = \frac{S(A \cap B)}{S(A)} = \frac{S(A \cap B)/S(\Omega)}{S(A)/S(\Omega)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

Probabilités conditionnelles

On note $\mathbb{P}(B|A)$ la probabilité qu'une expérience réalisée soit dans B , sachant qu'elle est dans A .

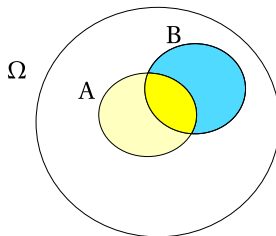


A et B sont indépendants, ou $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, ssi $\mathbb{P}(B|A) = \mathbb{P}(B)$

👉 savoir que $\omega \in A$ n'apporte pas d'information sur $\mathbb{P}(\omega \in B)$.

Formule des probabilités totales

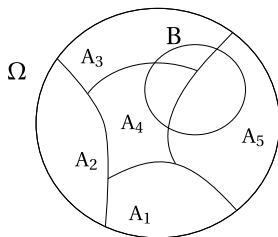
Une expérience dans B peut être dans A ou dans son complémentaire \bar{A} .



$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap \bar{A}) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\bar{A})\mathbb{P}(\bar{A})$$

Formule des probabilités totales

Plus généralement, on peut considérer A_1, \dots, A_n , 2 à 2 incompatibles avec $\Omega = A_1 \cup \dots \cup A_n$.



$$\mathbb{P}(B) = \mathbb{P}(B|A_1)\mathbb{P}(A_1) + \dots + \mathbb{P}(B|A_n)\mathbb{P}(A_n)$$

Formule de Bayes

On a $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$; on en déduit la formule de Bayes :

$$\begin{aligned}\mathbb{P}(A|B) &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\bar{A})\mathbb{P}(\bar{A})}\end{aligned}$$

Variables aléatoires

Variables aléatoires

Une **variable aléatoire** X est une mesure réalisée sur une expérience aléatoire. On peut dire que X est une fonction de Ω dans \mathbb{R} .

On décrit les variables aléatoires par leur loi, c'est-à-dire en donnant le moyen de calculer les valeurs

$$\mathbb{P}(a < X \leq b)$$

pour tous $a, b \in \mathbb{R}$.

On distinguera deux types de variables aléatoires : les variables aléatoires **discrètes** et les variables aléatoires **continues à densité**.

Variables aléatoires discrètes

Une variable aléatoire X est discrète si on peut énumérer les valeurs que X peut prendre : x_1, x_2, \dots

Leur loi est donnée par la **fonction de masse** $\mathbb{P}(X = x)$.

Alors pour tout A

$$\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x).$$

Exemple : lancer d'un dé

Si X est le résultat d'un lancer de dé à 6 faces équilibré, la loi de X est donnée par

$$\mathbb{P}(X = k) = \frac{1}{6}$$

pour k entier entre 1 et 6.

Si $A = \{2, 4, 6\}$ (tirage pair), on a

$$\mathbb{P}(X \in A) = \mathbb{P}(X = 2) + \mathbb{P}(X = 4) + \mathbb{P}(X = 6) = \frac{1}{2}.$$

Exemple : loi binomiale

Loi du tirage avec remise : on tire n boules dans une urne qui contient une proportion p de boules rouges ; X est le nombre de boules rouges tirées.

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

pour k entier entre 0 et n .

Le **coefficient binomial** $\binom{n}{k}$ est

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Exemple : loi binomiale

Loi du tirage avec remise : on tire n boules dans une urne qui contient une proportion p de boules rouges ; X est le nombre de boules rouges tirées.

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

pour k entier entre 0 et n .

Exemple de calcul pratique d'un coefficient binomial :

$$\begin{aligned} \binom{11}{5} &= \frac{11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2}{(5 \times 4 \times 3 \times 2) \times (6 \times 5 \times 4 \times 3 \times 2)} \\ &= \frac{11 \times 10 \times 9 \times 8 \times 7}{5 \times 4 \times 3 \times 2} \end{aligned}$$

Exemple : loi binomiale

Loi du tirage avec remise : on tire n boules dans une urne qui contient une proportion p de boules rouges ; X est le nombre de boules rouges tirées.

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

pour k entier entre 0 et n .

Exemple de calcul pratique d'un coefficient binomial :

$$\begin{aligned} \binom{11}{5} &= \frac{11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2}{(5 \times 4 \times 3 \times 2) \times (6 \times 5 \times 4 \times 3 \times 2)} \\ &= \frac{11 \times 9 \times 8 \times 7}{4 \times 3} \end{aligned}$$

Exemple : loi binomiale

Loi du tirage avec remise : on tire n boules dans une urne qui contient une proportion p de boules rouges ; X est le nombre de boules rouges tirées.

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

pour k entier entre 0 et n .

Exemple de calcul pratique d'un coefficient binomial :

$$\begin{aligned} \binom{11}{5} &= \frac{11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2}{(5 \times 4 \times 3 \times 2) \times (6 \times 5 \times 4 \times 3 \times 2)} \\ &= \frac{11 \times 3 \times 8 \times 7}{4} \end{aligned}$$

Exemple : loi binomiale

Loi du tirage avec remise : on tire n boules dans une urne qui contient une proportion p de boules rouges ; X est le nombre de boules rouges tirées.

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

pour k entier entre 0 et n .

Exemple de calcul pratique d'un coefficient binomial :

$$\begin{aligned}\binom{11}{5} &= \frac{11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2}{(5 \times 4 \times 3 \times 2) \times (6 \times 5 \times 4 \times 3 \times 2)} \\ &= 11 \times 3 \times 2 \times 7 \\ &= 462\end{aligned}$$

Variables aléatoires continues à densité

Une variable aléatoire X sera dite **continue de densité $f(x)$** si pour tout $a \leq b$ on a

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx.$$

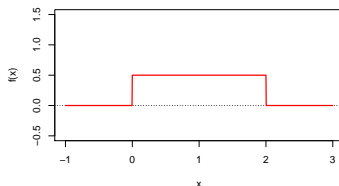
Pour que $f(x)$ soit une densité, il faut $f(x) \geq 0$ pour tout x , et

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Exemple : loi uniforme

La loi uniforme sur $[0, 2]$ est la loi de densité

$$f(x) = \begin{cases} \frac{1}{2} & \text{si } x \in [0, 2] \\ 0 & \text{sinon} \end{cases}$$



Espérance d'une variable aléatoire

L'**espérance** d'une variable aléatoire discrète est

$$E(X) = \sum_x x \mathbb{P}(X = x).$$

L'espérance d'une variable aléatoire continue à densité est

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx.$$

Loi des grands nombres : si X_1, X_2, \dots sont des variables aléatoires indépendantes et de même loi d'espérance μ

$$\frac{1}{n}(X_1 + \dots + X_n) \xrightarrow[n \rightarrow \infty]{} \mu.$$

Exemples

Espérance d'un jet de dés :

$$E(X) = \frac{1}{6} \times 1 + \cdots + \frac{1}{6} \times 6 = \frac{21}{6} = 3,5.$$

Espérance d'une variable uniforme sur $[0, 2]$:

$$E(X) = \int_0^2 \frac{1}{2} x dx = \left[\frac{1}{4} x^2 \right]_0^2 = 1 - 0 = 1.$$

Propriétés de l'espérance

Linéarité de l'espérance : si X et Y sont deux v.a. et a et b sont des constantes, on a

$$\begin{aligned}E(aX + b) &= aE(X) + b \\E(aX + bY) &= aE(X) + bE(Y).\end{aligned}$$

Espérance d'une fonction de X : Si Φ est une fonction, $\Phi(X)$ est une variable aléatoire ; son espérance est

$$E(\Phi(X)) = \sum_x \Phi(x) \mathbb{P}(X = x)$$

ou

$$E(\Phi(X)) = \int_{-\infty}^{+\infty} \Phi(x) f(x) dx.$$

Variance d'une variable aléatoire

Notons $\mu = E(X)$. La **variance** de X est

$$\text{var}(X) = E\left((X - \mu)^2\right).$$

La variance est donc la « moyenne » des carrés des écarts de X à μ , sa « moyenne ». Elle mesure la dispersion de X autour de son espérance μ .

Si X est une mesure avec une unité : par exemple une taille en mètres ; alors $\text{var}(X)$ est en mètres carrés. L'**écart-type** de X , qui est $\sqrt{\text{var}(X)}$ est en mètres. C'est l'ordre de grandeur des écarts qu'on peut attendre entre X et μ .

Propriétés de la variance

Si X est une variable aléatoire, on a

$$\text{var}(X) = E(X^2) - E(X)^2.$$

Si a et b sont des constantes,

$$\text{var}(aX + b) = a^2 \text{var}(X).$$

Notons que l'écart-type de $aX + b$ est égal à a fois l'écart-type de X .

Exemple

Variance d'une variable uniforme sur $[0, 2]$:

$$E(X^2) = \int_0^2 \frac{1}{2}x^2 dx = \left[\frac{1}{6}x^3 \right]_0^2 = \frac{8}{6} = \frac{4}{3},$$

donc

$$\text{var}(X) = E(X^2) - E(X)^2 = \frac{4}{3} - 1 = \frac{1}{3}.$$

Coefficients d'asymétrie et d'aplatissement

Le k^{e} moment de X est

$$\mu_k = E(X^k).$$

Le premier moment est donc l'espérance $\mu_1 = E(X)$.

Coefficients d'asymétrie et d'aplatissement

Le k^{e} moment centré est

$$m_k = E \left((X - \mu_1)^k \right).$$

Le 2^e moment centré m_2 est donc la variance σ^2 .

Coefficients d'asymétrie et d'aplatissement

Le k^e moment centré est

$$m_k = E \left((X - \mu_1)^k \right).$$

Le 2^e moment centré m_2 est donc la variance σ^2 .

Le coefficient d'asymétrie (ou *skewness*) est

$$\gamma_1 = \frac{m_3}{\sigma^3} = E \left(\left(\frac{X - \mu_1}{\sigma} \right)^3 \right).$$

Si $\gamma_1 > 0$, asymétrie à droite.

Coefficients d'asymétrie et d'aplatissement

Le k^e moment centré est

$$m_k = E \left((X - \mu_1)^k \right).$$

Le 2^e moment centré m_2 est donc la variance σ^2 .

Le coefficient d'aplatissement (ou *kurtosis*) est

$$\kappa = \frac{m_4}{\sigma^4} = E \left(\left(\frac{X - \mu_1}{\sigma} \right)^4 \right).$$

On définit aussi l'« *excess kurtosis* » = $\kappa - 3$.

Fonction de répartition et quantiles

La **fonction de répartition** d'une variable aléatoire X est

$$F(x) = \mathbb{P}(X \leq x).$$

Elle se calcule à partir de la loi de X .

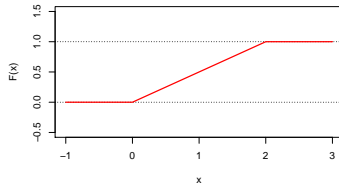
Le **quantile** de niveau α de X est le nombre x_α tel que $\mathbb{P}(X \leq x_\alpha) = \alpha$, c'est-à-dire

$$F(x_\alpha) = \alpha.$$

Exemple : loi uniforme

La fonction de répartition de la loi uniforme sur $[0, 2]$ est

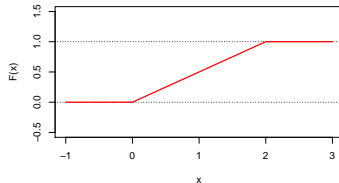
$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ \int_0^x \frac{1}{2} dx = \frac{1}{2}x & \text{si } x \in [0, 2] \\ 1 & \text{si } x > 2 \end{cases}$$



Exemple : loi uniforme

Donc le quantile de niveau α est $x_\alpha = 2\alpha$ pour $\alpha \in [0, 1]$.

Par exemple la médiane est $x_{0,5} = 1$.



Intervalle de pari

L'intervalle $[u, v]$ est un **intervalle de pari** de niveau $\gamma = 1 - \alpha$ pour X si

$$\mathbb{P}(u \leq X \leq v) = \gamma.$$

Lien avec les quantiles : Si x_a et x_b sont les quantiles de niveau a et b pour X , on a

$$\mathbb{P}(x_a \leq X \leq x_b) = b - a.$$

Pour un intervalle de niveau prescrit $\gamma = 1 - \alpha$ on prendra souvent $a = \alpha/2$ et $b = 1 - \alpha/2$:

$$\mathbb{P}(x_{\alpha/2} \leq X \leq x_{1-\alpha/2}) = 1 - \alpha.$$

Variables aléatoires simultanées

Couples de variables aléatoires

Si on fait plusieurs mesures sur une même expérience aléatoire, les valeurs obtenues ne sont pas nécessairement indépendantes (ex : poids/taille).

On définit naturellement l'indépendance de X et Y par

$$\mathbb{P}(X \in A \text{ et } Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

Dans le cas général on ne peut pas se contenter de décrire la loi de X et Y , il faut décrire leur **loi jointe**.

Description par les lois marginales et conditionnelles

On peut donner la loi de X , et, pour toutes les valeurs x possibles pour X , la loi de Y conditionnellement à $X = x$.

On pourra avoir deux lois discrètes ($\mathbb{P}(X = x)$ et $\mathbb{P}(Y = y|X = x)$), deux lois continues (X de densité $f(x)$, et Y de densité conditionnelle à $X = x$: $g(y|X = x)$), ou un mélange des deux.

On peut calculer les probabilités $\mathbb{P}((X, Y) \in A)$ à l'aide de la formule des probabilités totales, par exemple

$$\mathbb{P}((X, Y) \in A) = \sum_{(x,y) \in A} \mathbb{P}(X = x) \mathbb{P}(Y = y|X = x)$$

$$\mathbb{P}((X, Y) \in A) = \iint_{(x,y) \in A} f(x) g(y|X = x) dx dy$$

Description par la loi jointe

Dans le cas discret on donne une fonction de masse $\mathbb{P}(X = x, Y = y)$;

$$\mathbb{P}((X, Y) \in A) = \sum_{(x,y) \in A} \mathbb{P}(X = x, Y = y).$$

Dans le cas continu on donne une densité $f(x, y)$;

$$\mathbb{P}((X, Y) \in A) = \iint_A f(x, y) dx dy.$$

Si X et Y sont indépendantes, dans le cas discret, on a

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y).$$

Description par la loi jointe

Dans le cas discret on donne une fonction de masse $\mathbb{P}(X = x, Y = y)$;

$$\mathbb{P}((X, Y) \in A) = \sum_{(x,y) \in A} \mathbb{P}(X = x, Y = y).$$

Dans le cas continu on donne une densité $f(x, y)$;

$$\mathbb{P}((X, Y) \in A) = \iint_A f(x, y) dx dy.$$

Si X et Y sont indépendantes, dans le cas continu, on a

$$f(x, y) = \phi(x)\psi(y),$$

où ϕ est la densité de X et ψ la densité de y .

Lien entre les deux descriptions

La loi de marginale de X peut se déduire de la loi de (X, Y) .

Dans le cas discret, sa fonction de masse est :

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y).$$

Dans le cas continu, sa densité $\phi(x)$ est

$$\phi(x) = \int_y f(x, y) dy.$$

Exemple

Fonction de masse $\mathbb{P}(X = x, Y = y)$

$x =$	0	1	2
$y = 0$	0,10	0,06	0,04
1	0,05	0,04	0,01
2	0	0,35	0,35

Exemple

Fonction de masse $\mathbb{P}(X = x, Y = y)$

👉 lois marginales

$x =$	0	1	2	
$y = 0$	0,10	0,06	0,04	0,2
1	0,05	0,04	0,01	0,1
2	0	0,35	0,35	0,7
	0,15	0,45	0,40	

Lien entre les deux descriptions

La loi de Y conditionnellement à $X = x$ peut également être retrouvée à partir de la loi jointe.

Dans le cas discret, on obtient la fonction de masse

$$\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)}.$$

Dans le cas continu, on obtient la densité

$$\psi(y|X = x) = \frac{f(x, y)}{\phi(x)}.$$

Exemple (suite)

Fonction de masse $\mathbb{P}(X = x, Y = y)$

👉 lois marginales et conditionnelles

$x =$	0	1	2	
$y = 0$	0,10	0,06	0,04	0,2
1	0,05	0,04	0,01	0,1
2	0	0,35	0,35	0,7

$\mathbb{P}(X = x y = 0)$	0,50	0,30	0,20
---------------------------	------	------	------

Exemple (suite)

Fonction de masse $\mathbb{P}(X = x, Y = y)$

👉 lois marginales et conditionnelles

$x =$	0	1	2	
$y = 0$	0,10	0,06	0,04	0,2
1	0,05	0,04	0,01	0,1
2	0	0,35	0,35	0,7

$\mathbb{P}(X = x y = 1)$	0,50	0,40	0,10
---------------------------	------	------	------

Exemple (suite)

Fonction de masse $\mathbb{P}(X = x, Y = y)$

👉 lois marginales et conditionnelles

$x =$	0	1	2	
$y = 0$	0,10	0,06	0,04	0,2
1	0,05	0,04	0,01	0,1
2	0	0,35	0,35	0,7

$\mathbb{P}(X = x y = 2)$	0	0,50	0,50
---------------------------	---	------	------

Covariance de deux variables aléatoires

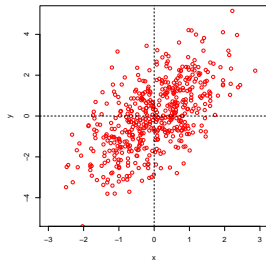
On note $E(X) = \mu_X$ et $E(Y) = \mu_Y$. La **covariance** de X et Y est

$$\text{cov}(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right).$$

Signe de la covariance (heuristique)

Si $\text{cov}(X, Y)$ est positif, $X - \mu_X$ et $Y - \mu_Y$ tendent à prendre des valeurs de même signe

👉 quand $X > \mu_X$, Y tend à être $> \mu_Y$



Covariance de deux variables aléatoires

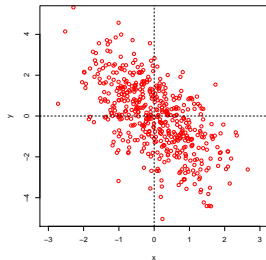
On note $E(X) = \mu_X$ et $E(Y) = \mu_Y$. La **covariance** de X et Y est

$$\text{cov}(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right).$$

Signe de la covariance (heuristique)

Inversement si $\text{cov}(X, Y)$ est négatif, quand $X > \mu_X$, Y tend à être $< \mu_Y$.

Attention la taille des écarts à μ_X et à μ_Y est aussi prise en compte dans la covariance.



Propriétés de la covariance

On a

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y).$$

Si a et b sont des constantes, on a les règles de calcul suivantes :

$$\begin{aligned}\text{cov}(X, X) &= \text{var}(X) \\ \text{cov}(X, Y) &= \text{cov}(Y, X) \\ \text{cov}(aX + b, Y) &= a \text{cov}(X, Y) \\ \text{cov}(X_1 + X_2, Y) &= \text{cov}(X_1, Y) + \text{cov}(X_2, Y).\end{aligned}$$

On a aussi

$$\text{var}(X + Y) = \text{var}(X) + 2\text{cov}(X, Y) + \text{var}(Y)$$

Corrélation de deux variables aléatoires

Le **coefficient de corrélation linéaire** ou plus simplement **coefficient de corrélation** de X et Y est

$$r = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

r est toujours entre -1 et 1 .

$$\text{cor}(X, X) = 1$$

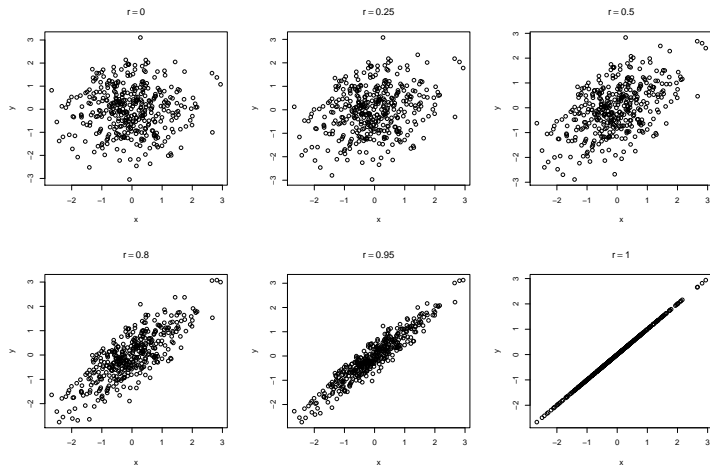
$$\text{cor}(X, Y) = \text{cor}(Y, X)$$

$$\text{cor}(X, -Y) = -\text{cor}(X, Y)$$

$$\text{cor}(aX + b, cY + d) = \text{cor}(X, Y) \text{ (si } a, c > 0 \text{)}.$$

(Invariance par changement d'échelle)

Corrélation de deux variables aléatoires



Somme et moyenne de variables aléatoires indépendantes

Si X et Y sont indépendantes alors $\text{cov}(X, Y) = 0$. On a donc $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$. Plus généralement si X_1, \dots, X_n sont indépendantes,

$$\text{var}(X_1 + \dots + X_n) = \text{var}(X_1) + \dots + \text{var}(X_n).$$

👉 Si X_1, \dots, X_n sont indépendantes de même espérance μ et de même variance σ^2 , la moyenne empirique $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$ est une variable aléatoire d'espérance μ et de variance $\frac{1}{n}\sigma^2$.

Loi faible des grands nombres

On fixe $\varepsilon \in \mathbb{R}$. La probabilité que $|\bar{X}_n - \mu| < \varepsilon$ tend vers 1 quand $n \rightarrow +\infty$.

Loi de la somme de variables aléatoires indépendantes

Soit $Z = X + Y$ avec X et Y indépendantes. Dans le cas discret :

$$\begin{aligned}P(Z = z) &= \sum_{x,y \text{ tq } x+y=z} \mathbb{P}(X = x)\mathbb{P}(Y = y) \\&= \sum_x \mathbb{P}(X = x)\mathbb{P}(Y = z - x).\end{aligned}$$

Dans le cas continu, on a la formule analogue pour la densité $h(z)$ de Z :

$$h(z) = \int_x f(x)g(z - x)dx.$$

Processus de Bernoulli et de Poisson

Expérience de Bernoulli

Une **expérience de Bernoulli** est une expérience aléatoire ayant deux résultats possibles, le succès ou l'échec.

On note p la probabilité de succès ; $1 - p$ est la probabilité d'échec.

👉 Variable X de loi $\mathcal{B}(p)$, avec

$$\mathbb{P}(X = 1) = p \quad (\text{succès}),$$

$$\mathbb{P}(X = 0) = 1 - p \quad (\text{échec}).$$

On a $E(X) = p$ et $\text{var}(X) = p(1 - p)$.

Processus de Bernoulli

Le **processus de Bernoulli** consiste à renouveler une expérience de Bernoulli un nombre (potentiellement) infini de fois.

On suppose que les expériences successives sont indépendantes.

Ceci se modélise comme une suite de variables aléatoires indépendantes X_1, X_2, \dots , de loi $\mathcal{B}(p)$, qui donnent les résultats des expériences successives.

Exemple : séries de parties de pile ou face ($p = 0,5$), etc.

Nombre de succès : la loi binomiale

Soit X le nombre de succès après n expériences :

$$X = X_1 + \cdots + X_n.$$

Alors X suit une **loi binomiale** de paramètres n et p : $X \sim \text{Bin}(n, p)$.

Pour k entre 0 et n , on a

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

On a $E(X) = np$ et $\text{var}(X) = np(1 - p)$.

Rang du premier succès : la loi géométrique

On renouvelle une expérience de Bernoulli de loi $\mathcal{B}(p)$. On s'intéresse à X : le rang du premier succès.

Alors X suit une **loi géométrique** de paramètre p ; on a

$$\mathbb{P}(X = k) = (1 - p)^{k-1}p$$

pour $k \geq 1$.

On a $E(X) = \frac{1}{p}$ et $\text{var}(X) = \frac{1-p}{p^2}$.

On a également $\mathbb{P}(X > k) = (1 - p)^k$.

Propriété d'oubli

On a $\mathbb{P}(X > k + \ell) = \mathbb{P}(X > k)\mathbb{P}(X > \ell)$:

$$\begin{aligned}\mathbb{P}(X > k + \ell) &= (1 - p)^{k+\ell} \\ &= (1 - p)^k (1 - p)^\ell \\ &= \mathbb{P}(X > k)\mathbb{P}(X > \ell).\end{aligned}$$

Donc

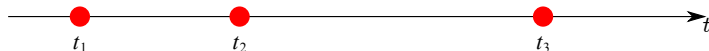
$$\begin{aligned}\mathbb{P}(X > k + \ell | X > k) &= \frac{\mathbb{P}(X > k + \ell \text{ et } X > k)}{\mathbb{P}(X > k)} \\ &= \frac{\mathbb{P}(X > k + \ell)}{\mathbb{P}(X > k)} = \mathbb{P}(X > \ell).\end{aligned}$$

Les expériences successives étant indépendantes, le fait d'avoir déjà échoué à k expériences ne change pas la probabilité d'échouer à ℓ nouvelles expériences ou plus. Le processus est dit **sans mémoire**.

Processus de Poisson

Le processus de Poisson est un processus de comptage en temps continu. On s'intéresse à la survenue d'événements indépendants, qui arrivent à des instants $t_1, t_2, \dots \in \mathbb{R}^{>0}$.

Exemple : observation de voitures au bord d'une route. On observe en moyenne 20 voitures par heure (par exemple), et ceci ne varie pas... les observations (nombre de voitures, intervalle de temps entre deux voitures) sont des variables aléatoires.



Nombre d'événements : la loi de Poisson

On appelle **taux** du processus de Poisson le nombre moyen λ_0 d'événements par unité de temps (ex : $\lambda_0 = 20$ voitures à l'heure).

Le nombre X d'événements pendant une durée Δt suit une **loi de Poisson** de paramètre $\lambda = \lambda_0 \times \Delta t$.

Si $X \sim \mathcal{P}(\lambda)$, pour $k \geq 0$, on a

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda).$$

On a $E(X) = \lambda$ et $\text{var}(X) = \lambda$.

Temps d'attente : la loi exponentielle

Le temps T d'attente avant le premier événement suit une **loi exponentielle** de paramètre $\lambda = \lambda_0$ (le taux du processus) : $T \sim \mathcal{E}(\lambda)$.

C'est une loi continue de densité

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{si } t \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

On a $E(T) = \frac{1}{\lambda}$ et $\text{var}(T) = \frac{1}{\lambda^2}$.

On a également $\mathbb{P}(T > t) = e^{-\lambda t}$.

Propriété d'oubli

On a $\mathbb{P}(T > t + s) = \mathbb{P}(T > t)\mathbb{P}(T > s)$:

$$\begin{aligned}\mathbb{P}(T > t + s) &= e^{-\lambda(t+s)} \\ &= e^{-\lambda t} e^{-\lambda s} \\ &= \mathbb{P}(T > t)\mathbb{P}(T > s)\end{aligned}$$

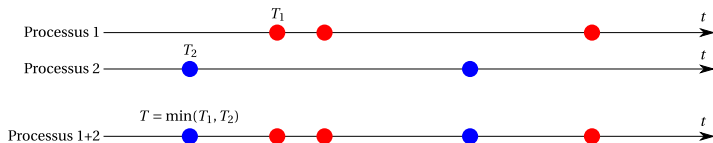
Donc

$$\mathbb{P}(T > t + s | T > t) = \frac{\mathbb{P}(T > t + s)}{\mathbb{P}(T > t)} = \mathbb{P}(T > s).$$

Le taux du processus est constant : le fait d'avoir déjà attendu un temps t ne change pas la probabilité de devoir attendre un temps s ou supérieur avant un événement. Le processus est sans mémoire.

Superposition de processus de Poisson

- un observateur compte les voitures : processus de Poisson de paramètre λ_1
- un autre observateur compte les camions : processus de Poisson de paramètre λ_2 .
- un troisième observateur compte tous les véhicules : processus de Poisson de paramètre $\lambda = \lambda_1 + \lambda_2$.



En moyenne λ_1 voitures par unité de temps, λ_2 camions : $\lambda_1 + \lambda_2$ véhicules.

Superposition de processus de Poisson

Si X_1 et X_2 sont deux variables aléatoires indépendantes, $X_1 \sim \mathcal{P}(\lambda_1)$ et $X_2 \sim \mathcal{P}(\lambda_2)$, alors

$$X_1 + X_2 \sim \mathcal{P}(\lambda = \lambda_1 + \lambda_2).$$

Si T_1 et T_2 sont deux variables aléatoires indépendantes, $T_1 \sim \mathcal{E}(\lambda_1)$ et $T_2 \sim \mathcal{E}(\lambda_2)$, alors

$$\min(T_1, T_2) \sim \mathcal{E}(\lambda = \lambda_1 + \lambda_2).$$

Processus de Poisson limite d'un processus de Bernoulli

Considérons un processus de Poisson de taux λ_0 ; en découpant le temps en petits intervalles δ_t , assez courts pour négliger la probabilité que deux événements se produisent dans un même intervalle de temps, on peut le voir comme une succession d'expériences de Bernoulli avec $p = \lambda_0 \delta_t$. Comme δ_t est petit, p est petit également.

Exemple : si $\lambda_0 = 20$ voitures par heures, chaque seconde ($\delta_t = 1/3600$) est une expérience de Bernoulli avec probabilité $p = \frac{20}{3600}$ de voir apparaître une voiture.

Le nombre de voitures observées après n expériences suit une loi $\text{Bin}(n, p)$; ou, n expériences correspondant à un temps $n\delta_t$, $\mathcal{P}(\lambda = \lambda_0 \cdot n\delta_t = np)$.

De façon générale, si p est petit, la loi $\text{Bin}(n, p)$ peut être approchée par une loi $\mathcal{P}(\lambda = np)$.

Loi de Gauss, du χ^2 , etc.

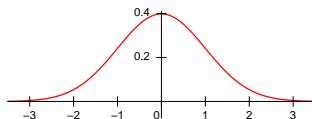
Loi de Gauss

...ou de Laplace-Gauss, ou loi normale.

La loi $\mathcal{N}(\mu, \sigma^2)$ a pour densité

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$

Son espérance est μ et sa variance σ^2 .



La loi $\mathcal{N}(0, 1)$ est dite loi normale standard ou loi normale centrée réduite.

- Si $X \sim \mathcal{N}(\mu, \sigma^2)$ alors $(aX + b) \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
- En particulier, si $Z \sim \mathcal{N}(0, 1)$, alors $X = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$,
et si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $Z = \frac{1}{\sigma}(X - \mu) \sim \mathcal{N}(0, 1)$.
- Si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ sont indépendantes, alors

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Loi de Gauss bivariée

La loi de (X, Y) est une loi de Gauss bivariée si

- La loi marginale de X est gaussienne : $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$
- La loi de Y conditionnellement à $X = x$ est gaussienne :

$$Y|X = x \sim \mathcal{N}(\alpha + \beta x, \tau).$$

Alors la loi marginale de Y est gaussienne, d'espérance $\mu_Y = \alpha + \beta\mu_X$, de variance $\sigma_Y^2 = \beta^2\sigma_X^2 + \tau$.

On a de plus $\sigma_{XY} = \beta\sigma_X^2$.

Loi de Gauss bivariée

La loi de (X, Y) est une loi de Gauss bivariée si

- La loi marginale de X est gaussienne : $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$
- La loi de Y conditionnellement à $X = x$ est gaussienne :

$$Y|X = x \sim \mathcal{N}(\alpha + \beta x, \tau).$$

Inversement, si (X, Y) suit une loi de Gauss bivariée, si on connaît μ_X , μ_Y , σ_X^2 , σ_Y^2 et $\sigma_{XY} = r\sigma_X\sigma_Y$, on peut calculer α , β , et τ .

👉 On peut décrire la loi de (X, Y) par μ_X , μ_Y , σ_X^2 , σ_Y^2 et σ_{XY} ou r .

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2} = r \frac{\sigma_Y}{\sigma_X}, \quad \alpha = \mu_Y - \beta\mu_X = \mu_Y - r \frac{\sigma_Y}{\sigma_X} \mu_X \quad \text{et} \quad \tau = (1 - r^2)\sigma_Y^2$$

Loi de Gauss bivariée

Dans le modèle Gaussien, ceci donne une interprétation du coefficient de corrélation r : l'espérance de Y conditionnellement à $X = x$ est

$$E(Y|X = x) = \alpha + \beta x = \mu_Y + r \left(\frac{x - \mu_X}{\sigma_X} \right) \sigma_Y$$

👉 l'écart attendu entre Y et μ_Y , sachant X , exprimé en nombre d'écart-types σ_Y , est le nombre d'écart-types observés pour X .

Exemple Soit (X, Y) la taille de deux frères. On donne $\mu_X = \mu_Y = 175$, $\sigma_X = \sigma_Y = 7$ et $r = 0,4$.

Si le premier frère mesure $X = 189$, soit un écart à la moyenne de $\frac{x - \mu_X}{\sigma_X} = 2$ écart-types, en espérance la taille du deuxième frère Y vaut $\mu_Y + 0,4 \times 2 \times \sigma_Y = 180,6$.

On a les propriétés suivantes :

- Le vecteur (X, Y) suit une loi bivariée ssi toutes les combinaisons linéaires $aX + bY$ sont gaussiennes.
- Si (X, Y) suit une loi de Gauss bivariée, avec $\text{cov}(X, Y) = 0$, alors X et Y sont indépendantes.

Attention, il ne suffit pas que X et Y soient toutes deux gaussiennes pour que leur loi jointe soit gaussienne !

On peut définir de la même façon les lois multivariées de dimension > 2 .
La loi de (X_1, X_2, X_3) est gaussienne si

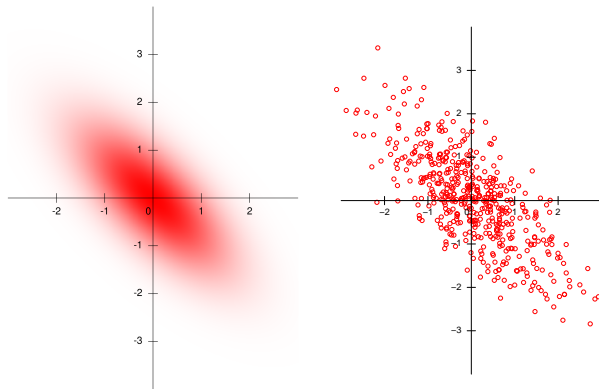
- La loi marginale de (X_1, X_2) est gaussienne ;
- La loi de X_3 conditionnellement à $X_1 = x_1, X_2 = x_2$ est gaussienne :

$$X_3 | X_1 = x_1, X_2 = x_2 \sim \mathcal{N}(\alpha + \beta_1 x_1 + \beta_2 x_2, \tau).$$

Etc.

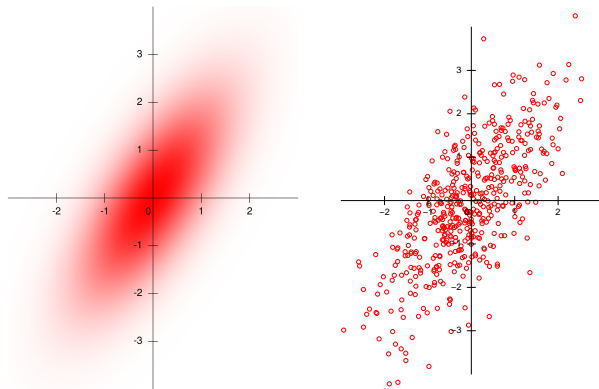
Il y a comme auparavant correspondance entre les coefficients α, β_1, β_2 et τ , et les variances / covariances des variables X_1, X_2, X_3 , mais les relations sont plus complexes à écrire.

Lois de Gauss bivariées : exemples et contre-exemples



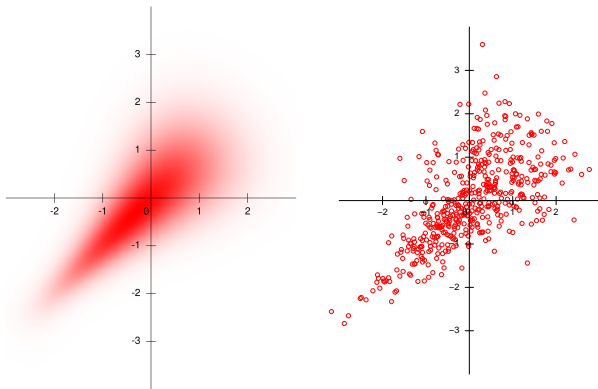
Densité et 500 points au hasard, pour un vecteur gaussien avec $\sigma_x^2 = \sigma_y^2 = 1$ et $r = -0,7$.

Lois de Gauss bivariées : exemples et contre-exemples



Densité et 500 points au hasard, pour un vecteur gaussien avec $\sigma_x^2 = 1$, $\sigma_y^2 = 2$ et $r = 0,7$.

Lois de Gauss bivariées : exemples et contre-exemples



Densité et 500 points au hasard, pour un vecteur (X, Y) non gaussien avec X, Y tous deux gaussiens

Loi du χ^2

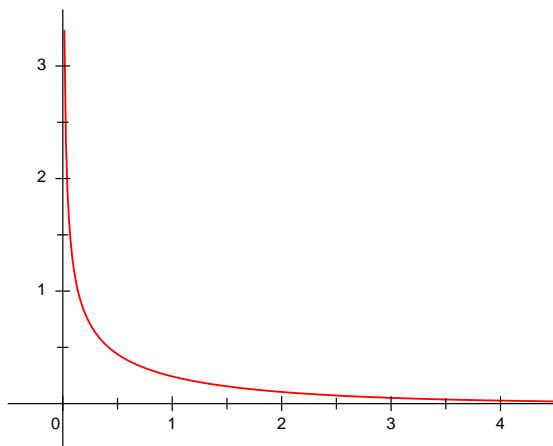
Soient Z_1, \dots, Z_d des variables aléatoires indépendantes de loi $\mathcal{N}(0, 1)$. La variable aléatoire $Y = Z_1^2 + \dots + Z_d^2$ suit une loi continue à densité, appelée **loi du χ^2 à d degrés de libertés**, notée $\chi^2(d)$.

L'espérance de Y est $E(Y) = d$ et sa variance $\text{var}(Y) = 2d$.

Soient deux variables aléatoires indépendantes $Y_1 \sim \chi^2(d_1)$ et $Y_2 \sim \chi^2(d_2)$. Alors $Y = Y_1 + Y_2$ suit une loi $\chi^2(d = d_1 + d_2)$.

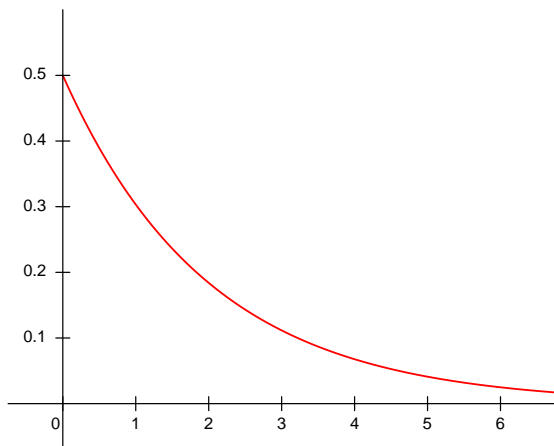
Si Y_1 et Y_2 sont indépendantes, si $Y_1 \sim \chi^2(d_1)$ et $Y = Y_1 + Y_2$ suit une loi $\chi^2(d)$ alors $Y_2 \sim \chi^2(d_2 = d - d_1)$.

Loi du χ^2



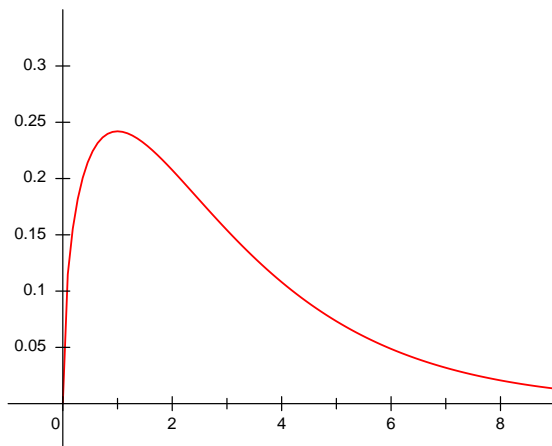
Densité du $\chi^2(1)$

Loi du χ^2



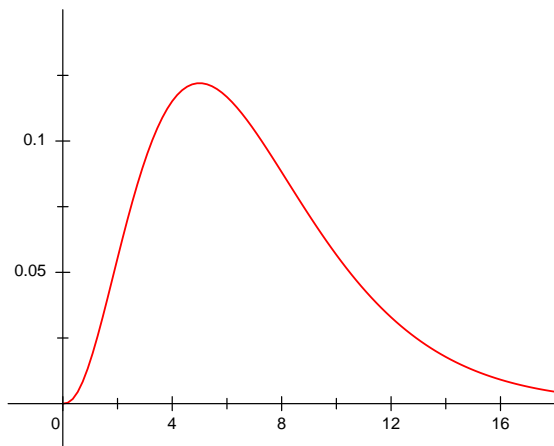
Densité du $\chi^2(2)$

Loi du χ^2



Densité du $\chi^2(3)$

Loi du χ^2



Densité du $\chi^2(7)$

Théorème central limite

Soient X_1, X_2, \dots des variables aléatoires indépendantes et de même loi, d'espérance μ et une variance σ^2 .

On pose $S_n = X_1 + \dots + X_n$. Quand n tend vers l'infini, la loi de

$$Y_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$$

s'approche de la loi normale centrée réduite $\mathcal{N}(0, 1)$.

En pratique, cela signifie que pour n « assez grand » on peut approcher la loi de S_n par la loi normale $\mathcal{N}(n\mu, n\sigma^2)$.

Théorème central limite

Soient X_1, X_2, \dots des variables aléatoires indépendantes et de même loi, d'espérance μ et une variance σ^2 .

On pose $S_n = X_1 + \dots + X_n$. Quand n tend vers l'infini, la loi de

$$Y_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$$

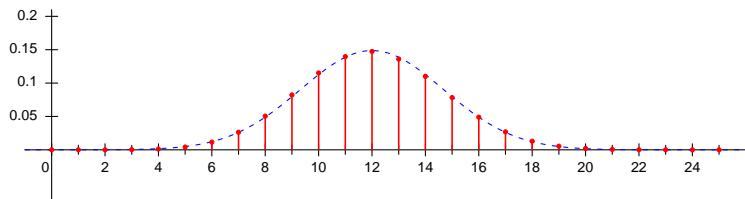
s'approche de la loi normale centrée réduite $\mathcal{N}(0, 1)$.

En pratique, cela signifie que pour n « assez grand » on peut approcher la loi de S_n par la loi normale $\mathcal{N}(n\mu, n\sigma^2)$.

👉 Pour n assez grand, on peut approcher la loi de $\bar{X}_n = \frac{1}{n}S_n$ par la loi normale $\mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$.

Théorème central limite : application

La loi binomiale $\mathcal{Bin}(n, p)$ est la somme de n variables de Bernoulli indépendantes de paramètre p . On peut donc approcher cette loi binomiale par une loi normale $\mathcal{N}(np, np(1 - p))$ « dès que n est assez grand ».



La fonction de masse de la loi $\mathcal{Bin}(n = 30, p = 0,4)$
et la densité de $\mathcal{N}(np = 12, np(1 - p) = 7,2)$

Théorème central limite : quand n est-il assez grand ?

Théorème de Berry-Esséen

Soient X_1, X_2, \dots indépendantes de même loi, d'espérance μ , de variance σ^2 .

Soit $\rho = E(|X - \mu|^3)$.

On pose $S_n = X_1 + \dots + X_n$ et $Y_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$.

Le théorème central limite affirme qu'on peut approcher $\mathbb{P}(Y_n \leq a)$ par $\mathbb{P}(Z \leq a)$ avec $Z \sim \mathcal{N}(0, 1)$; le théorème de Berry-Esséen ajoute que l'erreur commise est inférieure à

$$0,5 \frac{\rho}{\sigma^3 \sqrt{n}}.$$

Théorème central limite : quand n est-il assez grand ?

Exemples

- Pour une loi de Poisson de paramètre $\lambda = 1/10\,000$, on a $\rho/\sigma^3 \simeq 1000$.
- Pour une loi de Bernoulli de paramètre p , on a

$$\frac{\rho}{\sigma^3} = \frac{p^2 + (1-p)^2}{\sqrt{p(1-p)}} \leq \frac{1}{\sqrt{p(1-p)}},$$

on peut donc faire l'approximation normale dès que $np(1-p)$ assez grand.

Méthode du Delta

Transformation d'une variable aléatoire

Soit X une variable aléatoire d'espérance μ et de variance σ^2 . On considère la variable aléatoire $Y = \Phi(X)$.

Espérance et variance : On peut calculer $E(Y)$ par

$$E(Y) = E(\Phi(X)) = \int_x \phi(x)f(x)dx$$

(pour X de densité $f(x)$). Le calcul de $\text{var}(Y)$ via $E(Y^2) = E(\Phi(X)^2)$ peut se faire de la même façon.

On peut même **calculer la densité** $g(y)$ de Y en calculant d'abord la fonction de répartition $G(y) = \mathbb{P}(Y \leq y)$; la densité g est la dérivée de G .

Exemple : loi exponentielle

On rappelle que la densité de $T \sim \mathcal{E}(\lambda)$ est $f(t) = \lambda e^{-\lambda t}$, et sa fonction de répartition $F(t) = 1 - e^{-\lambda t}$.

Soit $U = aT$ (avec $a > 0$). On a

$$\mathbb{P}(U \leq u) = \mathbb{P}(aT \leq u) = \mathbb{P}\left(T \leq \frac{u}{a}\right) = 1 - e^{-\frac{\lambda u}{a}}.$$

On reconnaît la fonction de répartition d'une loi $\mathcal{E}(\lambda/a)$. La densité de U est sa dérivée

$$g(u) = \frac{\lambda}{a} e^{-\frac{\lambda}{a} u},$$

et c'est bien la densité de $\mathcal{E}(\lambda/a)$.

Méthode du Delta

De tels calculs ne sont pas toujours faciles à mener.

La **méthode du Delta** permet d'approcher l'espérance et la variance de Y quand σ^2 est assez petit :

$$E(Y) \simeq \Phi(\mu) + \frac{1}{2}\Phi''(\mu)\sigma^2 \simeq \Phi(\mu)$$

$$\text{var}(Y) \simeq \Phi'(\mu)^2\sigma^2$$

Plus σ^2 est petit, meilleure est l'approximation.

Méthode du Delta

Si de plus $X \sim \mathcal{N}(\mu, \sigma^2)$, on peut approcher la loi de $Y = \Phi(X)$ par une loi

$$\mathcal{N}\left(\Phi(\mu), \Phi'(\mu)^2 \sigma^2\right).$$

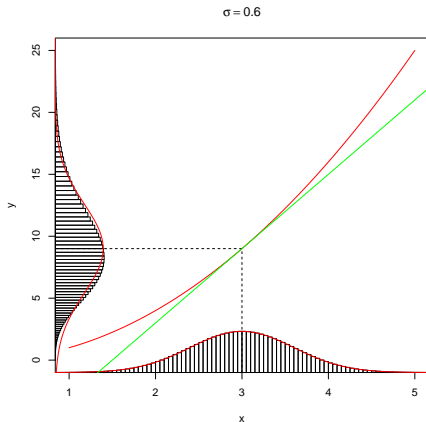
Plus σ^2 est petit, meilleure est l'approximation.

Application Si X_1, X_2, \dots sont indépendantes et de même loi d'espérance μ et de variance σ^2 , $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$ est approximativement normale, d'espérance μ et de variance $\frac{1}{n}\sigma^2$.

👉 Si n assez grand, $\Phi\left(\bar{X}_n\right)$ est approximativement normale, d'espérance $\Phi(\mu)$ et de variance $\frac{1}{n}\Phi'(\mu)^2\sigma^2$.

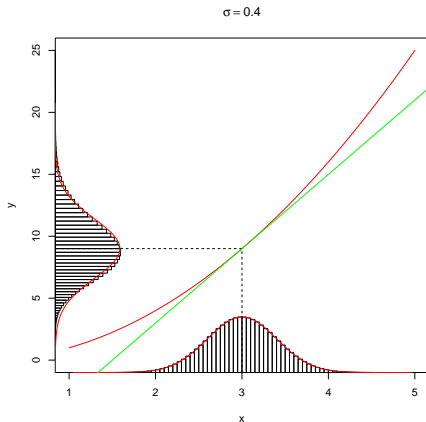
Méthode du Delta

Si $X \sim \mathcal{N}(3, \sigma^2)$, $Y = X^2$ est approximativement $\mathcal{N}(9, 36\sigma^2)$.



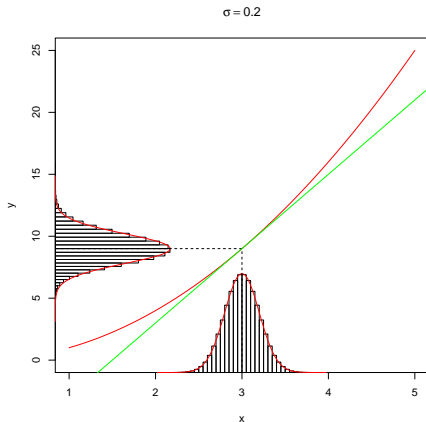
Méthode du Delta

Si $X \sim \mathcal{N}(3, \sigma^2)$, $Y = X^2$ est approximativement $\mathcal{N}(9, 36\sigma^2)$.



Méthode du Delta

Si $X \sim \mathcal{N}(3, \sigma^2)$, $Y = X^2$ est approximativement $\mathcal{N}(9, 36\sigma^2)$.



Application : loi de Poisson

Soit $U \sim \mathcal{P}(\lambda)$. Montrons qu'approximativement (pour λ assez grand)

$$\sqrt{U} \sim \mathcal{N}\left(\sqrt{\lambda}, \frac{1}{4}\right)$$

Quand λ est assez grand, la loi de U s'approche d'une loi normale :

$$U \sim \mathcal{N}(\lambda, \lambda)$$

(on pourra penser au fait qu'une variable de loi $\mathbb{P}(100)$ est la somme de 100 variables indépendantes de loi $\mathbb{P}(1)$).

La loi de $V = \frac{1}{\lambda} U$ s'approche donc d'une loi normale :

$$V \sim \mathcal{N}\left(1, \frac{1}{\lambda}\right).$$

Application : loi de Poisson

On pose $\Phi(x) = \sqrt{x}$, de sorte que $\Phi'(x) = \frac{1}{2\sqrt{x}}$.

Par la méthode du Delta, la loi de \sqrt{V} est approximativement normale d'espérance $\Phi(1) = 1$ et de variance $\Phi'(1)^2 \times \text{var}(V) = \frac{1}{4} \times \frac{1}{\lambda} = \frac{1}{4\lambda}$, donc

$$\sqrt{V} \sim \mathcal{N}\left(1, \frac{1}{4\lambda}\right).$$

On en déduit que la loi de $\sqrt{U} = \sqrt{\lambda}V$ s'approche d'une loi normale,

$$\sqrt{U} \sim \mathcal{N}\left(\sqrt{\lambda}, \frac{1}{4}\right).$$

Modélisation

(deuxième partie)

Estimations

- On effectue des expériences aléatoires indépendantes ; sur chacune d'elle on effectue une mesure : X_1, \dots, X_n sont des v.a. indépendantes de même loi
 - On veut estimer θ , un paramètre de cette loi, ou une quantité qui dépend de cette loi
 - Exemples : l'espérance, la variance, un quantile de la loi...
- 👉 On en calcule un **estimateur** $T = t(X_1, \dots, X_n)$
la valeur de T dépend de X_1, \dots, X_n ; c'est donc une variable aléatoire.
- 👉 Son espérance, sa variance, sa loi nous intéressent !

Premier exemple : estimer l'espérance

- Soient X_1, \dots, X_n des v.a. indépendantes de même loi, et μ l'espérance de cette loi
- L'estimateur naturel de μ est $\hat{\mu} = \bar{X} = \frac{1}{n} (X_1 + \dots + X_n)$
- Son espérance est $E(\bar{X}) = \mu$ (estimateur **sans biais**)
- Si la variance commune des X_i est σ^2 , on a

$$\text{var}(\bar{X}) = \frac{1}{n^2} \text{var}(X_1 + \dots + X_n) = \frac{1}{n} \sigma^2$$

- (Loi « asymptotique ») Si n « assez grand », le Th. Central Limite assure que la loi de \bar{X} est approximativement normale :

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{1}{n} \sigma^2\right)$$

Deuxième exemple : estimer la variance

Espérance μ connue

- Soient X_1, \dots, X_n indépendantes de même loi, d'espérance μ et de variance σ^2
- Si μ est connu, on peut estimer σ^2 par

$$\widehat{\sigma^2} = \frac{1}{n} \left((X_1 - \mu)^2 + \dots + (X_n - \mu)^2 \right)$$

- C'est un estimateur sans biais :

$$\begin{aligned} E(\widehat{\sigma^2}) &= \frac{1}{n} \left(E((X_1 - \mu)^2) + \dots + E((X_n - \mu)^2) \right) \\ &= \frac{1}{n} \left(\sigma^2 + \dots + \sigma^2 \right) \\ &= \sigma^2 \end{aligned}$$

👉 Que faire si μ inconnu ?

Deuxième exemple : estimer la variance

Espérance μ inconnue

- Il est naturel de remplacer μ par son estimateur \bar{X} : on obtient l'estimateur

$$\widetilde{S}^2 = \frac{1}{n} \left((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right)$$

- On a

$$\begin{aligned} \widetilde{S}^2 &= \frac{1}{n} (X_1^2 + \dots + X_n^2) - (\bar{X})^2 \\ &= \frac{1}{n} \left((X_1^2 + \dots + X_n^2) - \frac{1}{n} (X_1 + \dots + X_n)^2 \right) \end{aligned}$$

- Est-ce que cet estimateur est sans biais ?

Deuxième exemple : estimer la variance

Espérance μ inconnue

On a

$$E\left(\widetilde{S}^2\right) = \frac{1}{n} \left(E\left(X_1^2\right) + \cdots + E\left(X_n^2\right) \right) - E\left(\left(\overline{X}\right)^2\right)$$

Pour chacun des X_i , on a $E\left(X_i^2\right) = E\left(X_i\right)^2 + \text{var}\left(X_i\right) = \mu^2 + \sigma^2$. D'autre part

$$E\left(\left(\overline{X}\right)^2\right) = E\left(\overline{X}\right)^2 + \text{var}\left(\overline{X}\right) = \mu^2 + \frac{1}{n}\sigma^2.$$

Pour finir, on a

$$E\left(\widetilde{S}^2\right) = \left(\mu^2 + \sigma^2\right) - \left(\mu^2 + \frac{1}{n}\sigma^2\right) = \frac{n-1}{n}\sigma^2$$

Cet estimateur est **biaisé**.

Deuxième exemple : estimer la variance

Espérance μ inconnue

On obtient un estimateur sans biais en posant

$$\begin{aligned} S^2 &= \frac{n}{n-1} \widetilde{S}^2 \\ &= \frac{1}{n-1} \left((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right) \\ &= \frac{1}{n-1} \left((X_1^2 + \dots + X_n^2) - \frac{1}{n} (X_1 + \dots + X_n)^2 \right). \end{aligned}$$

Qualité d'un estimateur

Soient X_1, \dots, X_n sont des v.a. indépendantes de même loi.

On estime θ par $T = t(X_1, \dots, X_n)$.

- Le **biais** de T est $\text{biais}(T) = E(T - \theta) = E(T) - \theta$.

On préfère les estimateurs sans biais.

- La variance de T est (naturellement) $\text{var}(T)$.

On préfère les estimateurs de petite variance.

- L'**erreur quadratique moyenne** de T est

$$\text{eqm}(T) = E((T - \theta)^2).$$

On a la relation suivante : $\text{eqm}(T) = \text{biais}(T)^2 + \text{var}(T)$.

C'est un (bon) compromis entre biais et variance...

Le cas gaussien

- Soient X_1, \dots, X_n des v.a. indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$.
- L'estimateur de μ est $\hat{\mu} = \bar{X}$, de loi $\mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$.
- Si μ est connu, on estime σ^2 sans biais par

$$\widehat{\sigma^2} = \frac{1}{n} \left((X_1 - \mu)^2 + \dots + (X_n - \mu)^2 \right).$$

En écrivant

$$\frac{n}{\sigma^2} \widehat{\sigma^2} = \left(\frac{X_1 - \mu}{\sigma} \right)^2 + \dots + \left(\frac{X_n - \mu}{\sigma} \right)^2$$

on voit que $\frac{n}{\sigma^2} \widehat{\sigma^2}$ suit une loi $\chi^2(n)$.

Le cas gaussien

- En pratique on ne connaît pas μ , donc on utilise

$$S^2 = \frac{1}{n-1} \left((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right)$$

On peut montrer que

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$$

(on a « perdu » un degré de liberté en estimant μ ...)

- On utilisera souvent la notation un peu abusive

$$S^2 \sim \frac{\sigma^2}{n-1} \chi^2(n-1)$$

qu'on peut lire comme : S^2 est égale à $\frac{\sigma^2}{n-1}$ multiplié par une v.a. qui suit un $\chi^2(n-1)$.

- On a la propriété importante suivante : \bar{X} et S^2 sont indépendantes.

Estimation d'une proportion p

On a fait n expériences de Bernoulli de paramètre p , on veut estimer p (la probabilité du succès). Soit X le nombre de succès : on estime naturellement p par la proportion de succès observée,

$$\hat{p} = \frac{1}{n}X.$$

On sait que $X \sim \mathcal{Bin}(n, p)$, donc $E(\hat{p}) = p$ et $\text{var}(\hat{p}) = \frac{1}{n}p(1 - p)$.
Si n « assez grand », \hat{p} est approximativement normale,

$$\hat{p} \sim \mathcal{N}\left(p, \frac{1}{n}p(1 - p)\right).$$

Estimation d'une proportion p

Le retour du Delta

Rappel : la méthode du Delta assure que, si n assez grand, la loi de $\Phi(\hat{p})$ est approx. normale, d'espérance $\Phi(p)$ et de variance $\Phi'(p)^2 \times \frac{1}{n}p(1-p)$.

Si on choisit $\Phi(x) = \arcsin(\sqrt{x})$, on a $\Phi'(x) = \frac{1}{2\sqrt{x(1-x)}}$.

La variance de $\Phi(\hat{p})$ est approximativement

$$\left(\frac{1}{2\sqrt{p(1-p)}} \right)^2 \times \frac{1}{n}p(1-p) = \frac{1}{4n}.$$

Donc $\Phi(\hat{p}) \sim \mathcal{N}\left(\Phi(p), \frac{1}{4n}\right)$.

Avantages : « stabilisation » de la variance ; amélioration de la qualité de l'approximation normale.

Intervalles de confiance

Procédure d'intervalle de confiance

Soient X_1, \dots, X_n sont des v.a. indépendantes de même loi. On s'intéresse à θ qui dépend de cette loi.

On dit que $T_1 = t_1(X_1, \dots, X_n)$ et $T_2 = t_2(X_1, \dots, X_n)$ fournissent un intervalle de confiance de niveau $\gamma = (1 - \alpha)$ pour θ si

$$\mathbb{P}(T_1 \leq \theta \leq T_2) = \gamma.$$

Usuellement on prend $\gamma = 0,95$ ($\alpha = 0,05$) ou $\gamma = 0,90$ ($\alpha = 0,10$).

Attention à l'interprétation : θ est un paramètre fixé, ce sont bien les bornes T_1 et T_2 qui varient au fil des expériences.

Espérance d'une variable normale de variance connue

- Soient X_1, \dots, X_n des v.a. indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$. On suppose σ^2 connue, on veut un intervalle de confiance pour μ .
- La loi de \bar{X} est normale, d'espérance μ et de variance $\frac{1}{n}\sigma^2$.
- La loi de $\frac{\mu - \bar{X}}{\sigma/\sqrt{n}}$ est donc $\mathcal{N}(0, 1)$, et on a l'intervalle de pari

$$\mathbb{P}\left(-1,96 \leq \frac{\mu - \bar{X}}{\sigma/\sqrt{n}} \leq 1,96\right) = 0,95$$

d'où

$$\mathbb{P}\left(-1,96 \frac{\sigma}{\sqrt{n}} \leq \mu - \bar{X} \leq 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

et pour finir

$$\mathbb{P}\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

Espérance d'une variable normale de variance connue

Application à l'estimation d'une proportion

- n expériences de Bernoulli : $X \sim \text{Bin}(n, p)$ succès ; estimateur $\hat{p} = \frac{1}{n}X$
- Avec $\Phi(x) = \arcsin \sqrt{x}$, on a approx. $\Phi(\hat{p}) \sim \mathcal{N}\left(\Phi(p), \frac{1}{4n}\right)$

👉 Intervalle de confiance à 95% pour $\Phi(p)$:

$$\Phi(\hat{p}) - 1,96 \frac{1}{2\sqrt{n}} \leq \Phi(p) \leq \Phi(\hat{p}) + 1,96 \frac{1}{2\sqrt{n}}$$

- On en déduit un intervalle de confiance pour p en appliquant $\Phi^{-1}(x) = (\sin x)^2$ aux bornes de l'intervalle

- Soit $Z \sim \mathcal{N}(0, 1)$ et $Y \sim \chi^2(d)$, indépendantes. Par définition,

$$X = \frac{Z}{\sqrt{Y/d}}$$

suit une loi de Student à d de degrés de libertés, notée $t(d)$.

- 🔊 Soient X_1, \dots, X_n des v.a. indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$; on considère les estimateurs usuels de l'espérance et de la variance, \bar{X} et S^2 . On a

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t(n-1)$$

Loi de Student

Preuve

En effet $\bar{X} \sim \mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$ et $S^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$ sont indépendantes.

On écrit

$$\bar{X} = \mu + \sqrt{\frac{\sigma^2}{n}}Z \quad \text{et} \quad S^2 = \frac{\sigma^2}{n-1}Y,$$

avec $Z \sim \mathcal{N}(0, 1)$, $Y \sim \chi^2(n-1)$, indépendantes.

Alors

$$\begin{aligned} \frac{\bar{X} - \mu}{\sqrt{S^2/n}} &= \frac{\sqrt{\frac{\sigma^2}{n}}Z}{\sqrt{\frac{\sigma^2}{n} \times \frac{Y}{n-1}}} \\ &= \frac{Z}{\sqrt{\frac{Y}{n-1}}} \sim t(n-1) \end{aligned}$$

Espérance d'une variable normale de variance inconnue

Soient X_1, \dots, X_n des v.a. indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$; on considère les estimateurs usuels de l'espérance et de la variance, \bar{X} et S^2 . On a

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t(n-1)$$

et donc, si $t_{1-\alpha/2}^{n-1}$ est le quantile $1 - \frac{\alpha}{2}$ de la loi $t(n-1)$,

$$\mathbb{P} \left(-t_{1-\alpha/2}^{n-1} \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \leq t_{1-\alpha/2}^{n-1} \right) = 1 - \alpha$$

On en déduit l'intervalle de confiance

$$\mathbb{P} \left(\bar{X} - t_{1-\alpha/2}^{n-1} \sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{X} + t_{1-\alpha/2}^{n-1} \sqrt{\frac{S^2}{n}} \right) = 1 - \alpha.$$

Variance d'une variable normale

Soient X_1, \dots, X_n des v.a. indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$. On estime σ^2 par son estimateur usuel S^2 , qui suit une loi $\frac{\sigma^2}{n-1} \chi^2(n-1)$. On pose $Y = \frac{n-1}{\sigma^2} S^2$, donc $Y \sim \chi^2(n-1)$. Alors

$$\mathbb{P}\left(x_{\alpha/2}^{n-1} \leq Y \leq x_{1-\alpha/2}^{n-1}\right) = 1 - \alpha$$

$$\mathbb{P}\left(\frac{1}{x_{1-\alpha/2}^{n-1}} \leq \frac{1}{Y} \leq \frac{1}{x_{\alpha/2}^{n-1}}\right) = 1 - \alpha$$

et donc


$$\mathbb{P}\left(\frac{n-1}{x_{1-\alpha/2}^{n-1}} S^2 \leq \sigma^2 \leq \frac{n-1}{x_{\alpha/2}^{n-1}} S^2\right) = 1 - \alpha.$$

Loi F (loi de Fisher-Snedecor)

- Soient $Y_1 \sim \chi^2(d_1)$ et $Y_2 \sim \chi^2(d_2)$, indépendantes. La loi $F(d_1, d_2)$ est la loi de

$$X = \frac{Y_1/d_1}{Y_2/d_2}.$$

- Si $X \sim F(d_1, d_2)$, $\frac{1}{X} \sim F(d_2, d_1)$. On en déduit que le quantile de niveau α de $F(d_1, d_2)$ est l'inverse du quantile de niveau $1 - \alpha$ de $F(d_2, d_1)$: $F_{\alpha}^{d_1, d_2} = \left(F_{1-\alpha}^{d_2, d_1}\right)^{-1}$.

 Soient deux estimations indépendantes d'une même variance : $S_1^2 \sim \frac{\sigma^2}{n_1-1} \chi^2(n_1 - 1)$, $S_2^2 \sim \frac{\sigma^2}{n_2-1} \chi^2(n_2 - 1)$. Alors

$$\frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Loi F : quotient de deux variances

👉 Soient deux estimations indépendantes de deux variances σ_1^2 et σ_2^2 :
 $S_1^2 \sim \frac{\sigma_1^2}{n_1-1} \chi^2(n_1 - 1)$, $S_2^2 \sim \frac{\sigma_2^2}{n_2-1} \chi^2(n_2 - 1)$. Alors

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

👉 On en déduit l'IC suivant sur σ_2^2/σ_1^2 :

$$\mathbb{P} \left(\frac{S_2^2}{S_1^2} F_{\alpha/2}^{n_1-1, n_2-1} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2} F_{1-\alpha/2}^{n_1-1, n_2-1} \right) = 1 - \alpha.$$

Tests

- On effectue des expériences aléatoires indépendantes ; sur chacune d'elle on effectue une mesure : X_1, \dots, X_n sont des v.a. indépendantes de même loi
- On veut décider entre deux hypothèses portant sur θ , un paramètre de cette loi, ou une quantité qui dépend de cette loi
- 👉 On calcule une statistique $T = t(X_1, \dots, X_n)$, et on prend une décision selon la valeur de T
Généralement : si $T \in A$ (un intervalle) on retient une des hypothèses, sinon on retient l'autre.

Forme des hypothèses testés

- Une des deux hypothèses (l'« hypothèse nulle ») sera de la forme $H_0 : \theta = \theta_0$
- L'autre (l'« hypothèse alternative ») pourra prendre plusieurs formes, selon les informations dont on dispose a priori sur les valeurs possibles de θ
 - $H_1 : \theta \neq \theta_0$ (test bilatéral)
 - $H_1 : \theta > \theta_0$ (test unilatéral)
 - $H_1 : \theta = \theta_1$ (il n'y a que deux valeurs possibles)
 - etc.
- Si $T \in A$ on choisit H_0 , si $T \notin A$ on choisit H_1

Erreurs et risques

Il y a deux façons de se tromper.

- Erreur de type 1 (ou de première espèce) : retenir H_1 alors que H_0 est vrai.

👉 risque associé : $\alpha = \mathbb{P}(T \notin A | H_0)$

- Erreur de type 2 (ou de seconde espèce) : retenir H_0 alors que H_1 est vrai.

👉 risque associé : $\beta = \mathbb{P}(T \in A | H_1)$

On aimerait que les valeurs de α et β soient petites toutes les deux.

Mais dans beaucoup de situations on ne peut calculer que α !!

👉 On fixe α à une petite valeur ($\alpha = 0.05$ en cours de biostats...)

Asymétrie entre H_0 et H_1

Le vocabulaire usuel prend acte de cette asymétrie :

- A est la « zone d'acceptation », son complémentaire « la zone de rejet »
- si $T \notin A$, « on rejette l'hypothèse nulle »
- si $T \in A$, « on ne rejette pas l'hypothèse nulle » ou « on accepte l'hypothèse nulle »
- certains enseignants interdisent cette dernière formulation !
- on met en garde : ne pas rejeter l'hypothèse nulle est provisoire
- ...mais en pratique, rejeter l'hypothèse nulle est tout aussi provisoire. C'est la reproductibilité des résultats qui importe.

Exemple

On a n variables X_1, \dots, X_n indépendantes, de loi $\mathcal{N}(\mu, \sigma^2)$.

La valeur de σ^2 est supposée connue.

On veut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.

On considère la statistique de test

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{1}{n}\sigma^2}}.$$

Sous H_0 , $T \sim \mathcal{N}(0, 1)$.

👉 On obtient un test de risque $\alpha = 0,05$ en rejetant H_0 si $|T| > 1,96$.

On ne peut calculer le risque β que si la « vraie » valeur de μ est connue.

Exemple

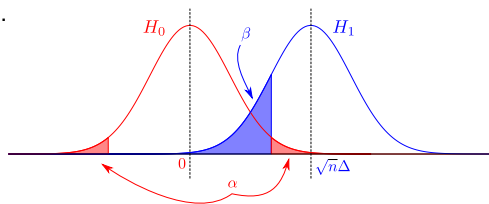
Si $\mu = \mu_1$, T suit toujours une loi normale, d'espérance

$$E(T) = \frac{\mu_1 - \mu_0}{\sqrt{\frac{1}{n}\sigma^2}} = \sqrt{n} \times \frac{\mu_1 - \mu_0}{\sigma} = \sqrt{n} \times \Delta$$

et de variance 1. On appelle Δ la taille de l'effet : c'est l'écart entre les moyennes, exprimé en « nombre d'écart-type ». Si on fait le test avec $\alpha = 0,05$, on a

$$\beta = \mathbb{P}(|T| < 1,96),$$

qui dépend de Δ .



Degré de signification

Pour un test où la règle de rejet est de la forme « on rejette H_0 quand $T > s$ », on définit le **degré de signification** (ou p -valeur, en anglais *p-value*) : c'est la probabilité d'observer, si H_0 est vraie, une valeur de T supérieure à la valeur observée.

- En pratique on observe une valeur $t = t(x_1, \dots, x_n)$, on calcule $p = \mathbb{P}(T > t)$.
- La règle devient « on rejette H_0 si $p \leq \alpha$ »
- La valeur de p est devenue un standard de publication... plus p est petit, plus on emporte la conviction
- p **n'est pas** la probabilité que l'hypothèse nulle soit vraie
- p ne reflète pas non plus l'importance de la découverte : sa valeur dépend de la taille de l'échantillon

Lien avec tests diagnostics

Dans un contexte médical, on utilise les termes de **sensibilité** et de **spécificité** d'un test diagnostic : la sensibilité est la probabilité de correctement diagnostiquer un individu atteint (diagnostic positif), la spécificité est la probabilité de correctement diagnostiquer un individu sain (diagnostic négatif).

En prenant H_0 : « l'individu testé est sain », le rejet de H_0 correspond à un diagnostic positif, le non rejet de H_0 correspond à un diagnostic négatif. On a alors les relations suivantes entre α , β , et sensibilité, spécificité :

$$\begin{aligned}\text{Spécificité} &= 1 - \alpha \\ \text{Sensibilité} &= 1 - \beta\end{aligned}$$

Valeurs prédictives

- **valeur prédictive positive** (VPP) : prob. d'être atteint si le test est positif
- **valeur prédictive négative** (VPN) : prob. d'être sain si le test est négatif

En notant \mathcal{P} la prévalence de la maladie dans la population testée, c'est-à-dire la probabilité a priori d'être sous H_1 : $\mathcal{P} = \mathbb{P}(H_1)$, on a :

$$\begin{aligned} \text{VPP} &= \mathbb{P}(H_1 | T \notin A) \\ &= \frac{\mathbb{P}(T \notin A | H_1) \mathbb{P}(H_1)}{\mathbb{P}(T \notin A | H_0) \mathbb{P}(H_0) + \mathbb{P}(T \notin A | H_1) \mathbb{P}(H_1)} \\ &= \frac{(1 - \beta) \mathcal{P}}{\alpha(1 - \mathcal{P}) + (1 - \beta) \mathcal{P}} \\ &= \frac{\text{Se} \cdot \mathcal{P}}{(1 - \text{Sp})(1 - \mathcal{P}) + \text{Se} \cdot \mathcal{P}} \end{aligned}$$

Valeurs prédictives

- **valeur prédictive positive** (VPP) : prob. d'être atteint si le test est positif
- **valeur prédictive négative** (VPN) : prob. d'être sain si le test est négatif

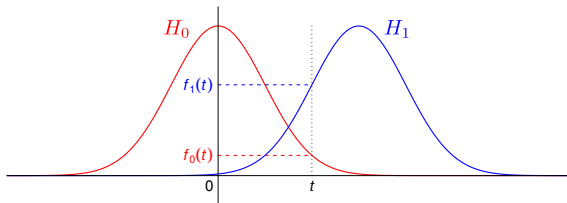
En notant \mathcal{P} la prévalence de la maladie dans la population testée, c'est-à-dire la probabilité a priori d'être sous H_1 : $\mathcal{P} = \mathbb{P}(H_1)$, on a :

$$\begin{aligned} \text{VPN} &= \mathbb{P}(H_0 | T \in A) \\ &= \frac{\mathbb{P}(T \in A | H_0) \mathbb{P}(H_0)}{\mathbb{P}(T \in A | H_0) \mathbb{P}(H_0) + \mathbb{P}(T \in A | H_1) \mathbb{P}(H_1)} \\ &= \frac{(1 - \alpha)(1 - \mathcal{P})}{(1 - \alpha)(1 - \mathcal{P}) + \beta \mathcal{P}} \\ &= \frac{\text{Sp} \cdot (1 - \mathcal{P})}{\text{Sp}(1 - \mathcal{P}) + (1 - \text{Se}) \cdot \mathcal{P}} \end{aligned}$$

Risque a posteriori

On observe $T = t$. Si on connaît :

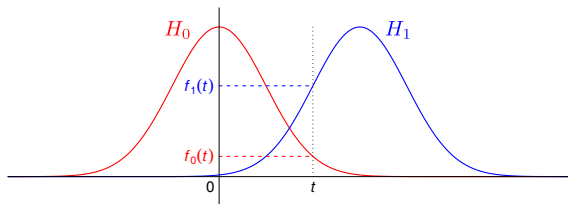
- la densité f_0 (resp. f_1) de T sous H_0 (resp. sous H_1)
- la probabilité a priori de H_0 (sain) et de H_1 (atteint) ($1 - \mathcal{P}$ et \mathcal{P})



Risque a posteriori

On peut calculer les probabilités a posteriori de H_0 et de H_1 , sachant $T = t$:

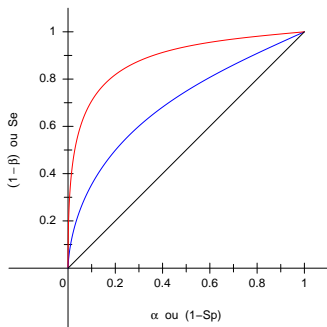
$$\frac{\mathbb{P}(H_1|T=t)}{\mathbb{P}(H_0|T=t)} = \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \times \frac{f_1(t)}{f_0(t)} = \frac{\mathcal{P}}{1-\mathcal{P}} \times \frac{f_1(t)}{f_0(t)}$$



Courbe ROC

Si la règle de rejet est de la forme : rejeter H_0 si $T > s$ (on peut se ramener à ce cas pour la plupart des test usuels), le risque α (ou $1 - Sp$) et le risque β (ou $1 - Se$) dépendent de la valeur du seuil s .

La courbe décrite par $(\alpha, 1 - \beta) = (1 - Sp, Se)$ quand on fait varier s est la courbe ROC du test.



Quelques tests usuels

Test sur une moyenne

Cas gaussien

On a n variables X_1, \dots, X_n indépendantes, de loi $\mathcal{N}(\mu, \sigma^2)$. La valeur de σ^2 est **inconnue**.

On veut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.

On considère la statistique de test

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{1}{n} S^2}}$$

où S^2 est l'estimateur de la variance : $S^2 \sim \frac{\sigma^2}{n-1} \chi^2(n-1)$.

👉 Sous H_0 , $T \sim t(n-1)$.

👉 On obtient un test de risque α en rejetant H_0 si $|T| > t_{1-\alpha/2}^{n-1}$.

Test sur une moyenne

Grands échantillons

On a n variables X_1, \dots, X_n indépendantes, de même loi d'espérance μ .
On veut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.

On considère la statistique de test

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{1}{n} S^2}}$$

où S^2 est l'estimateur de la variance.

- 👉 Sous H_0 , T est approximativement $\mathcal{N}(0, 1)$.
- 👉 On obtient un test de risque approx. α en rejetant H_0 si $|T| > z_{1-\alpha/2}$.

Comparaison de deux moyennes

Cas gaussien

On a un échantillon de taille n_1 issu d'une loi $\mathcal{N}(\mu_1, \sigma^2)$ et un échantillon de taille n_2 issu d'une loi $\mathcal{N}(\mu_2, \sigma^2)$ (même variance). On note $n = n_1 + n_2$, le nombre total d'observations.

On teste $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$ (test bilatéral).

Les moyennes empiriques des deux échantillons sont \bar{X}_1 et \bar{X}_2 . On a

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma^2\right).$$

On a dans chaque échantillon une estimation de σ^2 :

$$S_1^2 \sim \frac{\sigma^2}{n_1-1} \chi^2(n_1 - 1) \text{ et } S_2^2 \sim \frac{\sigma^2}{n_2-1} \chi^2(n_2 - 1).$$

Comparaison de deux moyennes

Cas gaussien

On estime σ^2 par

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n - 2} \sim \frac{\sigma^2}{n - 2} \chi^2(n - 2).$$

Sous H_0 , on a

$$T = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^2}} \sim t(n - 2).$$

Un test de risque α est obtenu en rejetant H_0 quand $|T| > t_{1-\alpha/2}^{n-2}$.

Variance d'une loi de Gauss

On a X_1, \dots, X_n indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$.

On teste $H_0 : \sigma^2 = \sigma_0^2$ contre $H_1 : \sigma^2 \neq \sigma_0^2$.

Sous H_0 ,

$$Y = \frac{n-1}{\sigma_0^2} S^2 \sim \chi^2(n-1).$$

👉 Test de risque α en rejetant H_0 quand $Y < x_{\alpha/2}^{n-1}$ ou $Y > x_{1-\alpha/2}^{n-1}$.

Comparaison de deux variances (lois de Gauss)

On a un échantillon de taille n_1 issu d'une loi $\mathcal{N}(\mu_1, \sigma_1^2)$ et un échantillon de taille n_2 issu d'une loi $\mathcal{N}(\mu_2, \sigma_2^2)$.

On teste $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$.

La variance empirique du premier échantillon est $S_1^2 \sim \frac{\sigma_1^2}{n_1-1} \chi^2(n_1 - 1)$
celle du second est $S_2^2 \sim \frac{\sigma_2^2}{n_2-1} \chi^2(n_2 - 1)$.

👉 sous H_0

$$\frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1).$$

👉 On obtient un test de risque α en rejetant H_0 quand

$$\frac{S_1^2}{S_2^2} < F_{\alpha/2}^{n_1-1, n_2-1} \text{ ou } \frac{S_1^2}{S_2^2} > F_{1-\alpha/2}^{n_1-1, n_2-1}.$$

Test sur une proportion

Soient X_1, \dots, X_n indépendantes de loi $B(p)$.

On teste $H_0 : p = p_0$ contre $H_1 : p \neq p_0$ (test bilatéral).

On estime p par $\hat{p} = \frac{1}{n}(X_1 + \dots + X_n)$. Sous H_0 et pour n « assez grand » la loi de \hat{p} est approximativement normale :

$$\hat{p} \sim \mathcal{N}\left(\mu = p_0, \sigma^2 = \frac{p_0(1-p_0)}{n}\right).$$

On obtient un test de risque α en rejetant H_0 quand

$$\left| \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right| > z_{1-\alpha/2}.$$

Comparaison de deux proportions

On a un échantillon de taille n_1 issu d'une loi $B(p_1)$ et un de taille n_2 issu d'une loi $B(p_2)$.

On teste $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$.

Sous H_0 , en notant $p = p_1 = p_2$, si les échantillons sont « assez grands » on a

$$(\hat{p}_1 - \hat{p}_2) \sim \mathcal{N} \left(\mu = 0, \sigma^2 = p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right).$$

👉 Problème : estimer la variance, qui dépend de p .

Comparaison de deux proportions

Première solution : en supposant H_0 vraie, on estime p par

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}.$$

On obtient un test de risque α en rejetant H_0 quand

$$\left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \right| > z_{1-\alpha/2}.$$

Comparaison de deux proportions

Deuxième solution : on utilise la variance de $\hat{p}_1 - \hat{p}_2$ dans le cas général, qui est

$$\frac{1}{n_1}p_1(1 - p_1) + \frac{1}{n_2}p_2(1 - p_2)$$

qu'on estime en remplaçant p_1 et p_2 par \hat{p}_1 et \hat{p}_2 .

On obtient un test de risque α en rejetant H_0 quand

$$\left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{1}{n_1}\hat{p}_1(1 - \hat{p}_1) + \frac{1}{n_2}\hat{p}_2(1 - \hat{p}_2)}} \right| > z_{1-\alpha/2}.$$

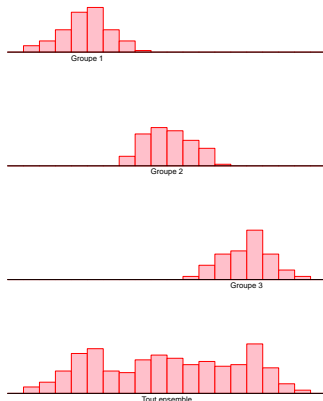
Remarque : Lequel de ces tests est le plus puissant ? Une analyse superficielle montre que si p_1 et p_2 sont suffisamment différents, la variance utilisée par la solution (1) peut être plus grande que celle utilisée par la solution (2). Mais la qualité de l'approximation par la loi normale doit être prise en compte également...

ANOVA 1

Analyse de la variance à un facteur

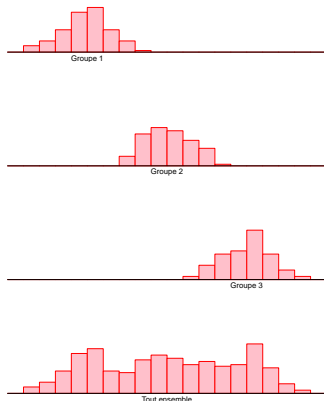
But : comparer la moyenne de p groupes en ne faisant qu'un seul test.

Moyen : estimer la variance de deux façons différentes, et comparer ces estimations.



Analyse de la variance à un facteur

Dans chacun des groupes, les données sont assez peu étalées : la variance est faible ; si on considère toutes les données ensemble, la variance est élevée.



Analyse de la variance à un facteur

Modèle

- n observations réparties en p groupes d'effectifs n_1, \dots, n_p .
- X_{ij} = j^{e} observation du groupe i , $i = 1, \dots, p$ et $j = 1, \dots, n_i$
- les X_{ij} sont normaux, l'espérance dépend du groupe i

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$$

On teste $H_0 : \mu_1 = \mu_2 = \dots = \mu_p$, vs H_1 : au moins deux des μ_i sont différents.

Analyse de la variance à un facteur

Reformulation comme modèle linéaire

De façon équivalente

$$X_{ij} = \mu + \alpha_i + E_{ij}$$

où les E_{ij} sont indépendantes de loi $\mathcal{N}(0, \sigma^2)$.

On doit imposer $\sum_i n_i \alpha_i = 0$ pour que la paramétrisation soit unique.

La moyenne du groupe i est $\mu_i = \mu + \alpha_i$. On a

$$\sum_i n_i \mu_i = \sum_i n_i (\mu + \alpha_i) = n\mu.$$

Avec ces notations, on teste $H_0 : \alpha_1 = \dots = \alpha_p = 0$ vs H_1 : au moins un α_i est non nul.

Analyse de la variance à un facteur

Notations

Pour alléger un peu les calculs, on note

$$X_{i+} = \sum_{j=1}^{n_i} X_{ij}$$

$$X_{i\bullet} = \frac{1}{n_i} X_{i+}$$

$$X_{++} = \sum_{ij} X_{ij} = \sum_{i=1}^p X_{i+}$$

$$X_{\bullet\bullet} = \frac{1}{n} X_{++} = \frac{1}{n} \sum_{i=1}^p n_i X_{i\bullet}$$

Ainsi, $X_{i\bullet}$ est la moyenne empirique de l'échantillon issu du groupe i ,
et $X_{\bullet\bullet}$ est la moyenne empirique de l'ensemble.

On note également les sommes de carrés des observations par

$$X_{i+}^2 = \sum_{j=1}^{n_i} X_{ij}^2$$

$$X_{++}^2 = \sum_{ij} X_{ij}^2$$

Analyse de la variance à un facteur

Estimation des paramètres

Paramètres μ_1, \dots, μ_p

La moyenne dans le groupe i est estimée par $\hat{\mu}_i = X_{i\bullet}$.

Paramètres $\mu, \alpha_1, \dots, \alpha_p$

On estime μ par la moyenne de l'ensemble

$$\hat{\mu} = X_{\bullet\bullet}$$

et on a

$$\hat{\alpha}_i = X_{i\bullet} - \hat{\mu} = X_{i\bullet} - X_{\bullet\bullet}$$

Analyse de la variance à un facteur

Sommes de carrés

Pour estimer la variance des données **prises dans leur ensemble** on définit la somme des carrés totaux :

$$SCT = \sum_{ij} (X_{ij} - X_{\bullet\bullet})^2 = X_{++}^2 - \frac{1}{n} (X_{++})^2 \sim \sigma^2 \chi^2(n-1) \text{ sous } H_0$$

et le carré moyen total

$$CMT = \frac{1}{n-1} SCT$$

qui est sous H_0 un estimateur sans biais de σ^2 (avec $n-1$ ddl).

Analyse de la variance à un facteur

Sommes de carrés

Pour estimer la variance des données **groupe par groupe**, on définit la somme des carrés du groupe i :

$$SC_i = \sum_{j=1}^{n_i} (X_{ij} - X_{i\bullet})^2 = X_{i+}^2 - \frac{1}{n_i} (X_{i+})^2 \sim \sigma^2 \chi^2(n_i - 1)$$

Pour chaque groupe on a une estimation de σ^2 par $\frac{1}{n_i-1} SC_i$.
On somme les SC_i pour obtenir la somme des carrés résiduels

$$SCR = \sum_{i=1}^p SC_i = \sum_{i=1}^p \sum_{j=1}^{n_i} (X_{ij} - X_{i\bullet})^2 \sim \sigma^2 \chi^2(n - p)$$

Le carré moyen résiduel : $CMR = \frac{1}{n-p} SCR$ est donc un estimateur sans biais de σ^2 (avec $n - p$ ddl).

Analyse de la variance à un facteur

Comparer CMR et CMT ?

Des valeurs de CMT « beaucoup plus grandes » que CMR plaident pour H_1 . Mais CMT et CMR ne sont pas indépendantes : comparaison difficile. On est amené à considérer la somme des carrés factoriels

$$SCF = \sum_{i=1}^p n_i (\hat{\alpha}_i)^2 = \sum_{i=1}^p n_i (X_{i\bullet} - X_{\bullet\bullet})^2 = \sum_{i=1}^p \frac{1}{n_i} (X_{i+})^2 - \frac{1}{n} (X_{++})^2$$

On a également

$$SCF = \frac{1}{n} \sum_{i < k} n_i n_k (\hat{\mu}_i - \hat{\mu}_k)^2 = \frac{1}{n} \sum_{i < k} n_i n_k (X_{i\bullet} - X_{k\bullet})^2.$$

Analyse de la variance à un facteur

Le théorème

On a

$$SCT = SCF + SCR.$$

D'autre part sous H_0 on a

$$SCR \sim \sigma^2 \chi^2(n - p)$$

$$SCT \sim \sigma^2 \chi^2(n - 1)$$

et **SCF et SCR sont indépendantes**, d'où

$$SCF \sim \sigma^2 \chi^2(p - 1).$$

Analyse de la variance à un facteur

Le test

De tout ce qui précède, on déduit que le quotient

$$F = \frac{CMF}{CMR} = \frac{SCF/(p-1)}{SCR/(n-p)}$$

suit sous H_0 une loi F de degrés de liberté $p-1$ et $n-p$. On obtient un test de risque α en rejetant H_0 quand $F > F_{1-\alpha}^{p-1, n-p}$ (test unilatéral).

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur	SCF	$p - 1$	$CMF = SCF/(p-1)$	$F = CMF/CMR$
résidus	SCR	$n - p$	$CMR = SCR/(n-p)$	
Total	SCT	$n - 1$		

Analyse de la variance à un facteur

Exemple

Temps de clairance parasitaire (TCP), pour un traitement anti-paludéen, patients de six régions différentes : trois régions africaines (groupes 1, 2 et 3) et trois régions asiatiques (groupes 3, 4 et 5).

Tester l'hypothèse selon laquelle le traitement a la même efficacité dans ces six régions.

	n_i	x_{i+}	x_{i+}^2
Groupe 1	10	568	37 186
Groupe 2	10	621	42 965
Groupe 3	10	612	44 479
Groupe 4	10	879	85 781
Groupe 5	10	732	55 390
Groupe 6	10	788	65 767
Total	60	4 200	331 568

Analyse de la variance à un facteur

Exemple

On peut compléter le tableau par les sommes de carrés de chacun des groupes, et par la SCT.

Par ex. $SC_1 = x_{1+}^2 - \frac{1}{n_1}(x_{1+})^2 = 37186 - 568^2/10 = 4923,6$.

$SCT = x_{++}^2 - \frac{1}{n}(x_{++})^2 = 331568 - 4200^2/60 = 37568$.

	n_i	x_{i+}	x_{i+}^2	SC_i
Groupe 1	10	568	37 186	4 923,6
Groupe 2	10	621	42 965	4 400,9
Groupe 3	10	612	44 479	7 024,6
Groupe 4	10	879	85 781	8 516,9
Groupe 5	10	732	55 390	1 807,6
Groupe 6	10	788	65 767	3 672,6
Total	60	4 200	331 568	37 568

Analyse de la variance à un facteur

Exemple

On obtient SCR en sommant les SC_i :

$$SCR = 4923,6 + 4400,9 + 7024,6 + \dots = 30346,2.$$

On calcule ensuite

$$SCF = SCT - SCR = 37568 - 30346,2 = 7221,8.$$

	n_i	x_{i+}	x_{i+}^2	SC_i
Groupe 1	10	568	37 186	4 923,6
Groupe 2	10	621	42 965	4 400,9
Groupe 3	10	612	44 479	7 024,6
Groupe 4	10	879	85 781	8 516,9
Groupe 5	10	732	55 390	1 807,6
Groupe 6	10	788	65 767	3 672,6
Total	60	4 200	331 568	37 568

Analyse de la variance à un facteur

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur				
résidus	30 346,2	54		
Total	37 568	59		

Analyse de la variance à un facteur

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur	7 221,8	5		
résidus	30 346,2	54		
Total	37 568	59		

Analyse de la variance à un facteur

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur	7 221,8	5	1 444,36	$F = 2,57$
résidus	30 346,2	54	561,97	
Total	37 568	59		

Analyse de la variance à un facteur

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur	7 221,8	5	1 444,36	$F = 2,57$
résidus	30 346,2	54	561,97	
Total	37 568	59		

La valeur de F est à comparer avec le quantile 0,95 de $F(5, 54) = 2,4$:
on rejette l'hypothèse d'égalité.

Analyse de la variance à un facteur

Contrastes

On peut vouloir comparer seulement deux groupes parmi les p groupes. On définit le contraste

$$C_{ik} = \mu_i - \mu_k = \alpha_i - \alpha_k.$$

On estime sa valeur par

$$\hat{C}_{ik} = X_{i\bullet} - X_{k\bullet} \sim \mathcal{N}\left(C_{ik}, \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_k}\right)\right)$$

On estime σ^2 par *CMR* avec $(n - p)$ ddl

- 👉 test t à $n - p$ ddl pour tester l'égalité
- 👉 quantiles de la loi $t(n - p)$ pour un intervalle de confiance

Analyse de la variance à un facteur

Exemple

Si on revient à notre exemple : on peut comparer les groupes 1 et 2, de moyenne respective 56,8 et 62,1, en utilisant le $CMR = 561,97$ pour estimer la variance commune.

On a alors

$$t = \frac{62,1 - 56,8}{\sqrt{\left(\frac{1}{10} + \frac{1}{10}\right) 561,97}} = 0,50$$

à comparer à un quantile de la loi $t(54)$ (non significatif).

Analyse de la variance à un facteur

Anova partielle

On peut vouloir faire l'anova sur seulement une partie des p groupes, par exemple sur les groupes 1, 2 et 3 (tester $H_0 : \mu_1 = \mu_2 = \mu_3$). Comme dans le cas des contrastes, il est souhaitable d'**utiliser la totalité des p groupes** pour l'estimation de la variance par le CMR.

Reste à calculer la SCF pour les seuls groupes dont on veut comparer la moyenne. Si I est l'ensemble des indices des groupes à comparer,

$$SCF_I = \frac{1}{\sum_{i \in I} n_i} \sum_{\substack{i < k \\ i, k \in I}} n_i n_k (X_{i\bullet} - X_{k\bullet})^2 = \frac{1}{\sum_{i \in I} n_i} \sum_{\substack{i < k \\ i, k \in I}} n_i n_k \hat{C}_{i,k}^2.$$

Analyse de la variance à un facteur

Anova partielle

Dans le cas où $I = \{1, 2, 3\}$ cela donne

$$SCF_{1,2,3} = \frac{1}{n_1 + n_2 + n_3} (n_1 n_2 (X_{1\bullet} - X_{2\bullet})^2 + n_1 n_3 (X_{1\bullet} - X_{3\bullet})^2 + n_2 n_3 (X_{2\bullet} - X_{3\bullet})^2)$$

On réalise ensuite le test en calculant

$$\frac{SCF_I / (p_I - 1)}{SCR / (n - p)}$$

où p_I est le nombre de groupes dans I (3 dans notre exemple).

Ce quotient suit une loi $F(p_I - 1, n - p)$.

Analyse de la variance à un facteur

Exemple

Revenons à notre exemple : les groupes 1, 2 et 3 sont les trois groupes africains.

On calcule

$$\begin{aligned}SCF_{\text{Afr}} &= \frac{1}{30} (100 \times (56,8 - 62,1)^2 + 100 \times (56,8 - 61,2)^2 + 100 \times (62,1 - \\ &= 160,87\end{aligned}$$

d'où un $CMF = 160,87/2 = 80,43$ et

$$F = \frac{80,43}{561,97} = 0,14,$$

à comparer à un quantile de la loi $F(2, 54)$ (quantile de niveau 0,95 : 3,2).

Analyse de la variance à un facteur

Comparaison de modèles emboîtés

En testant $H_0 : \mu_1 = \dots = \mu_6$ on a implicitement comparé les modèles suivants :

- on contraint $\mu_1 = \dots = \mu_p$ (un paramètre)
- μ_1, \dots, μ_p varient librement (p paramètres)

Le modèle à un paramètre est un cas particulier du modèle à p paramètres : si μ_1, \dots, μ_p varient librement, alors il est possible d'avoir $\mu_1 = \dots = \mu_p$. On parle de modèles emboîtés.

Le test de l'anova peut être interprété comme répondant à la question :
« est-ce que le deuxième modèle explique significativement mieux les observations que le premier » ?

Analyse de la variance à un facteur

Comparaison de modèles emboîtés

On peut de façon générale comparer deux modèles emboîtés. Pour cela on définit la somme de carrés (résiduels) associée à un modèle : elle se calcule en fusionnant les groupes qui sont confondus par le modèle.

- Dans le modèle à un seul paramètre, il n'y a qu'un groupe : la somme des carrés est SCT
- Dans le modèle à p paramètres, il y a p groupes, et la somme des carrés SCR

Le nombre de degrés de liberté de la somme de carrés associée à un modèle à k paramètres est $n - k$.

Analyse de la variance à un facteur

Comparaison de modèles emboîtés

Pour comparer deux modèles emboîtés, on part de leur somme de carrés. On fera une table de Fisher, avec une ligne par modèle : la ligne associée à un modèle à k paramètres est

Source	Somme des carrés	degrés de liberté	Carrés moyens
<i>modèle</i>	SCR_{mod}	$n - k$	$CMR_{mod} = SCR_{mod} / (n - k)$

Analyse de la variance à un facteur

Comparaison de modèles emboîtés

Si un modèle (1) à k_1 paramètres (ou k_1 groupes) est emboîté dans un modèle (2) à $k_2 > k_1$ paramètres (ou groupes), la somme de carrés SCR_1 est plus grande SCR_2 (les groupes dans le modèle (1) sont plus grands). Pour tester si le modèle (2) « explique significativement mieux les données » que le modèle (1), on fait un tableau d'analyse de la variance comme ceci :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(diff.)	$SCF = SCR_1 - SCR_2$	$n = n_1 - n_2$	$CMF = SCF/n$	$F = CMF/CMR_2$
mod ₂	SCR_2	$n_2 = n - k_2$	$CMR_2 = SCR_2/n_2$	
mod ₁	SCR_1	$n_1 = n - k_1$		

Analyse de la variance à un facteur

Exemple

Revenons à notre exemple. On va comparer les modèles

- (a) Le modèle à un paramètre où le traitement a la même efficacité dans les six groupes ;
- (b) le modèle à deux paramètres où le traitement a la même efficacité dans les groupes africains, et la même efficacité dans les groupes asiatiques ;
- (c) le modèle à six paramètres, un pour chaque groupe.

La SCR pour le modèle (a) (un groupe) est $SCR_a = SCT = 37568$ avec 59 ddl ; pour le modèle (c) (6 groupes) c'est $SCR_c = 30346,2$ avec 54 ddl ; la comparaison entre ces deux modèles a été faite par une anova classique. Reste à comparer (a) et (b), puis (b) et (c).

Analyse de la variance à un facteur

Exemple

Pour calculer la somme des carrés pour le modèle (b), il faut fusionner les groupes 1, 2 et 3 (groupes africains) d'une part, 4, 5 et 6 d'autre part (groupes asiatiques).

	n_i	x_{i+}	x_{i+}^2
Groupe 1+2+3 (afr)	30	1 801	124 630
Groupe 4+5+6 (asi)	30	2 399	206 938

Analyse de la variance à un facteur

Exemple

Pour calculer la somme des carrés pour le modèle (b), il faut fusionner les groupes 1, 2 et 3 (groupes africains) d'une part, 4, 5 et 6 d'autre part (groupes asiatiques).

	n_i	x_{i+}	x_{i+}^2	SC_i
Groupe 1+2+3 (afr)	30	1 801	124 630	16 509,97
Groupe 4+5+6 (asi)	30	2 399	206 938	15 097,97

Analyse de la variance à un facteur

Exemple

Pour calculer la somme des carrés pour le modèle (b), il faut fusionner les groupes 1, 2 et 3 (groupes africains) d'une part, 4, 5 et 6 d'autre part (groupes asiatiques).

	n_i	x_{i+}	x_{i+}^2	SC_i
Groupe 1+2+3 (afr)	30	1 801	124 630	16 509,97
Groupe 4+5+6 (asi)	30	2 399	206 938	15 097,97

La somme des carrés pour le modèle (b) est donc

$$SCR_b = 16509,97 + 15097,97 = 31607,94$$

avec $60 - 2 = 58$ degrés de liberté.

Analyse de la variance à un facteur

Exemple

Comparons les modèles (a) et (b) (c'est une anova à deux groupes) :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(a)-(b)				
(b) (2 par.)	31 607,94	58		
(a) (1 par.)	37 568	59		

Analyse de la variance à un facteur

Exemple

Comparons les modèles (a) et (b) (c'est une anova à deux groupes) :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(a)-(b)	5 960,06	1		
(b) (2 par.)	31 607,94	58		
(a) (1 par.)	37 568	59		

Analyse de la variance à un facteur

Exemple

Comparons les modèles (a) et (b) (c'est une anova à deux groupes) :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(a)-(b)	5 960,06	1	5 960,06	$F = 10,94$
(b) (2 par.)	31 607,94	58	544,96	
(a) (1 par.)	37 568	59		

Analyse de la variance à un facteur

Exemple

Comparons les modèles (a) et (b) (c'est une anova à deux groupes) :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(a)-(b)	5 960,06	1	5 960,06	$F = 10,94$
(b) (2 par.)	31 607,94	58	544,96	
(a) (1 par.)	37 568	59		

La statistique $F = 10,94$ est à comparer au quantile d'ordre 0,95 de $F(1, 58)$, qui vaut environ 4 : on rejette l'hypothèse nulle (le modèle (a) suffit à expliquer les observations) au profit du modèle (b).

Analyse de la variance à un facteur

Exemple

Comparons les modèles (b) et (c) :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(c) (6 par.)	30 346,2	54		
(b) (2 par.)	31 607,94	58		

Analyse de la variance à un facteur

Exemple

Comparons les modèles (b) et (c) :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(b)-(c)	1 261,74	4		
(c) (6 par.)	30 346,2	54		
(b) (2 par.)	31 607,94	58		

Analyse de la variance à un facteur

Exemple

Comparons les modèles (b) et (c) :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(b)-(c)	1 261,74	4	315,44	$F = 0,56$
(c) (6 par.)	30 346,2	54	561,97	
(b) (2 par.)	31 607,94	58		

Analyse de la variance à un facteur

Exemple

Comparons les modèles (b) et (c) :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(b)-(c)	1 261,74	4	315,44	$F = 0,56$
(c) (6 par.)	30 346,2	54	561,97	
(b) (2 par.)	31 607,94	58		

La statistique $F = 0,56$ est à comparer au quantile d'ordre 0,95 de $F(4, 54)$, qui vaut environ 2,5 : le modèle (c) n'explique pas significativement mieux les observations que le modèle (b).

...autrement dit, la variabilité entre les groupes est pour l'essentiel due à une variabilité inter-continentale.

ANOVA 2

Analyse de la variance à deux facteurs

Dans l'anova 1, on a une source possible de variations, par exemple des niveaux de traitement. On considère ici le cas où on a deux sources de variations possibles qui se croisent :

- deux traitements simultanés (par ex. deux molécules différentes administrées simultanément à des doses variées)
- un traitement (à plusieurs niveaux), et une covariable : par ex. le sexe des patients (deux niveaux) ; on peut même inclure un « effet patient » dans le modèle (mesures répétées).

Analyse de la variance à deux facteurs

Notations

X_{ijk} = observation numéro k dans le groupe où le premier facteur (facteur A) est au niveau i et le second facteur (facteur B) est au niveau j ;

$i = 1, \dots, p, j = 1, \dots, q, k = 1, \dots, n_{ij}$.

Comme pour l'anova 1, on a des notations abrégées pour les sommes et moyennes prises sur un indice :

$$\begin{aligned} n_{i+} &= \sum_{j=1}^q n_{ij}, & n_{+j} &= \sum_{i=1}^p n_{ij}, \\ X_{\bar{j}+} &= \sum_{k=1}^{n_{\bar{j}}} X_{ijk}, & X_{\bar{j}\bullet} &= \frac{1}{n_{\bar{j}}} X_{\bar{j}+}, \\ X_{i++} &= \sum_{j=1}^q X_{\bar{j}+}, & X_{i\bullet\bullet} &= \frac{1}{n_{i+}} X_{i++}, \\ X_{+j+} &= \sum_{i=1}^p X_{\bar{j}+}, & X_{\bullet j\bullet} &= \frac{1}{n_{+j}} X_{+j+}, \\ X_{+++} &= \sum_{ijk} X_{ijk}, & X_{\bullet\bullet\bullet} &= \frac{1}{n} X_{+++}. \end{aligned}$$

Analyse de la variance à deux facteurs

Plans d'expérience équilibrés

On ne considèrera que des plans d'expérience équilibrés, c'est-à-dire

$$n_{ij} = \frac{1}{n} n_{i+} n_{+j}$$

Si on présente les données dans un tableau, dans toutes les lignes les nombres d'observations sont dans des proportions identiques ; et dans toutes les colonnes les nombres d'observations sont dans des proportions identiques.

	B_1	B_2	B_3
Niveau A_1	$n_{11} = 2$	$n_{12} = 4$	$n_{13} = 10$
Niveau A_2	$n_{21} = 1$	$n_{22} = 2$	$n_{23} = 5$
Niveau A_3	$n_{31} = 3$	$n_{32} = 6$	$n_{33} = 15$
Niveau A_4	$n_{41} = 4$	$n_{42} = 8$	$n_{43} = 20$

Analyse de la variance à deux facteurs

Modèle

Le modèle est

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk},$$

où les E_{ijk} sont indépendantes de loi $\mathcal{N}(0, \sigma^2)$, avec les contraintes suivantes sur les paramètres :

$$\begin{aligned} \sum_{i=1}^p n_{i+} \alpha_i &= 0, & \sum_{j=1}^q n_{+j} \beta_j &= 0, \\ \forall i, \sum_{j=1}^q n_{+j} \gamma_{ij} &= 0, & \forall j, \sum_{i=1}^p n_{i+} \gamma_{ij} &= 0. \end{aligned}$$

Les différentes hypothèses à tester sont

- $H_{0A} : \forall i, \alpha_i = 0$ (pas d'effet du facteur A) ;
- $H_{0B} : \forall j, \beta_j = 0$ (pas d'effet du facteur B) ;
- $H_{0AB} : \forall i, j, \gamma_{ij} = 0$ (pas d'interaction entre A et B).

Analyse de la variance à deux facteurs

Estimation des paramètres $\alpha_i, \beta_j, \gamma_{ij}$

Avec les contraintes mentionnées, on a les estimations suivantes :

$$\hat{\mu} = X_{\dots}$$

$$\begin{aligned}\hat{\alpha}_i &= X_{i\bullet\bullet} - \hat{\mu} \\ &= X_{i\bullet\bullet} - X_{\dots}\end{aligned}$$

$$\begin{aligned}\hat{\beta}_j &= X_{\bullet j\bullet} - \hat{\mu} \\ &= X_{\bullet j\bullet} - X_{\dots}\end{aligned}$$

$$\begin{aligned}\hat{\gamma}_{ij} &= X_{ij\bullet} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu} \\ &= X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\dots}\end{aligned}$$

Analyse de la variance à deux facteurs

Les sommes de carrés

$$SCT = \sum_{ijk} (X_{ijk} - X_{\bullet\bullet\bullet})^2 = X_{+++}^2 - \frac{1}{n}(X_{+++})^2 \quad n - 1 \text{ ddl}$$

$$SCF_A = \sum_{i=1}^p n_{i+} (\hat{\alpha}_i)^2 = \sum_{i=1}^p \frac{1}{n_{i+}} (X_{i++})^2 - \frac{1}{n} (X_{+++})^2 \quad p - 1 \text{ ddl}$$

$$SCF_B = \sum_{j=1}^q n_{+j} (\hat{\beta}_j)^2 = \sum_{j=1}^q \frac{1}{n_{+j}} (X_{+j+})^2 - \frac{1}{n} (X_{+++})^2 \quad q - 1 \text{ ddl}$$

$$SCF_{AB} = \sum_{i=1}^p \sum_{j=1}^q n_{ij} (\hat{\gamma}_{ij})^2 \quad (p - 1)(q - 1) \text{ ddl}$$

$$SCR = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^{n_{ij}} (X_{ijk} - X_{ij\bullet})^2 \quad n - pq \text{ ddl}$$

On a

$$SCT = SCF_A + SCF_B + SCF_{AB} + SCR$$

Analyse de la variance à deux facteurs

Les tests

On a les statistiques de test suivantes.

$$\text{sous } H_{0A} : F_A = \frac{CMF_A}{CMR} \sim F(p-1, n-pq)$$

$$\text{sous } H_{0B} : F_B = \frac{CMF_B}{CMR} \sim F(q-1, n-pq)$$

$$\text{sous } H_{0AB} : F_{AB} = \frac{CMF_{AB}}{CMR} \sim F((p-1)(q-1), n-pq)$$

Analyse de la variance à deux facteurs

Exemple

	Traitement A	Traitement B	Traitement C	
Hommes	$n_{11} = 4$	$n_{12} = 4$	$n_{13} = 6$	$n_{1+} = 14$
	$x_{11+} = 38,8$	$x_{12+} = 53,5$	$x_{13+} = 53,9$	$x_{1++} = 146,2$
	$x_{11+}^2 = 382,90$	$x_{12+}^2 = 732,81$	$x_{13+}^2 = 505,73$	$x_{1++}^2 = 1621,44$
Femmes	$n_{21} = 6$	$n_{22} = 6$	$n_{23} = 9$	$n_{2+} = 21$
	$x_{21+} = 75,1$	$x_{22+} = 72,7$	$x_{23+} = 100,0$	$x_{2++} = 247,8$
	$x_{21+}^2 = 966,97$	$x_{22+}^2 = 894,45$	$x_{23+}^2 = 1139,72$	$x_{2++}^2 = 3001,14$
	$n_{+1} = 10$	$n_{+2} = 10$	$n_{+3} = 15$	$n_{++} = 35$
	$x_{+1+} = 113,9$	$x_{+2+} = 126,2$	$x_{+3+} = 153,9$	$x_{++++} = 394,0$
	$x_{+1+}^2 = 1349,87$	$x_{+2+}^2 = 1627,26$	$x_{+3+}^2 = 1645,45$	$x_{++++}^2 = 4622,58$

Analyse de la variance à deux facteurs

Exemple

	Traitement A	Traitement B	Traitement C	
Hommes	$n_{11} = 4$	$n_{12} = 4$	$n_{13} = 6$	$n_{1+} = 14$
	$x_{11+} = 38,8$	$x_{12+} = 53,5$	$x_{13+} = 53,9$	$x_{1++} = 146,2$
	$x_{11+}^2 = 382,90$	$x_{12+}^2 = 732,81$	$x_{13+}^2 = 505,73$	$x_{1++}^2 = 1621,44$
Femmes	$n_{21} = 6$	$n_{22} = 6$	$n_{23} = 9$	$n_{2+} = 21$
	$x_{21+} = 75,1$	$x_{22+} = 72,7$	$x_{23+} = 100,0$	$x_{2++} = 247,8$
	$x_{21+}^2 = 966,97$	$x_{22+}^2 = 894,45$	$x_{23+}^2 = 1139,72$	$x_{2++}^2 = 3001,14$
	$n_{+1} = 10$	$n_{+2} = 10$	$n_{+3} = 15$	$n_{++} = 35$
	$x_{+1+} = 113,9$	$x_{+2+} = 126,2$	$x_{+3+} = 153,9$	$x_{++++} = 394,0$
	$x_{+1+}^2 = 1349,87$	$x_{+2+}^2 = 1627,26$	$x_{+3+}^2 = 1645,45$	$x_{++++}^2 = 4622,58$

Première chose à vérifier : il s'agit bien d'un plan d'expérience équilibré...

Analyse de la variance à deux facteurs

Exemple

	Traitement A	Traitement B	Traitement C	
Hommes	$n_{11} = 4$	$n_{12} = 4$	$n_{13} = 6$	$n_{1+} = 14$
	$x_{11+} = 38,8$	$x_{12+} = 53,5$	$x_{13+} = 53,9$	$x_{1++} = 146,2$
	$x_{11+}^2 = 382,90$	$x_{12+}^2 = 732,81$	$x_{13+}^2 = 505,73$	$x_{1++}^2 = 1621,44$
Femmes	$n_{21} = 6$	$n_{22} = 6$	$n_{23} = 9$	$n_{2+} = 21$
	$x_{21+} = 75,1$	$x_{22+} = 72,7$	$x_{23+} = 100,0$	$x_{2++} = 247,8$
	$x_{21+}^2 = 966,97$	$x_{22+}^2 = 894,45$	$x_{23+}^2 = 1139,72$	$x_{2++}^2 = 3001,14$
	$n_{+1} = 10$	$n_{+2} = 10$	$n_{+3} = 15$	$n_{++} = 35$
	$x_{+1+} = 113,9$	$x_{+2+} = 126,2$	$x_{+3+} = 153,9$	$x_{+++} = 394,0$
	$x_{+1+}^2 = 1349,87$	$x_{+2+}^2 = 1627,26$	$x_{+3+}^2 = 1645,45$	$x_{+++}^2 = 4622,58$

On calcule la somme des carrés dans chacun des 6 groupes, par exemple

$$SC_{11} = x_{11+}^2 - \frac{1}{n_{11}}(x_{11+})^2 = 382,90 - \frac{1}{4}38,8^2 = 6,54, \text{ etc.}$$

Analyse de la variance à deux facteurs

Exemple

	Traitement A	Traitement B	Traitement C	
Hommes	$n_{11} = 4$ $x_{11+} = 38,8$ $x_{11+}^2 = 382,90$	$n_{12} = 4$ $x_{12+} = 53,5$ $x_{12+}^2 = 732,81$	$n_{13} = 6$ $x_{13+} = 53,9$ $x_{13+}^2 = 505,73$	$n_{1+} = 14$ $x_{1++} = 146,2$ $x_{1++}^2 = 1621,44$
	$n_{21} = 6$ $x_{21+} = 75,1$ $x_{21+}^2 = 966,97$	$n_{22} = 6$ $x_{22+} = 72,7$ $x_{22+}^2 = 894,45$	$n_{23} = 9$ $x_{23+} = 100,0$ $x_{23+}^2 = 1139,72$	$n_{2+} = 21$ $x_{2++} = 247,8$ $x_{2++}^2 = 3001,14$
	$n_{+1} = 10$ $x_{+1+} = 113,9$ $x_{+1+}^2 = 1349,87$	$n_{+2} = 10$ $x_{+2+} = 126,2$ $x_{+2+}^2 = 1627,26$	$n_{+3} = 15$ $x_{+3+} = 153,9$ $x_{+3+}^2 = 1645,45$	$n_{++} = 35$ $x_{+++} = 394,0$ $x_{+++}^2 = 4622,58$

On calcule $SCT = x_{+++}^2 - \frac{1}{n}(x_{+++})^2 = 4622,58 - \frac{1}{35}394^2 = 187,27$.

Analyse de la variance à deux facteurs

Exemple

	Traitement A	Traitement B	Traitement C	
Hommes	$n_{11} = 4$	$n_{12} = 4$	$n_{13} = 6$	$n_{1+} = 14$
	$x_{11+} = 38,8$	$x_{12+} = 53,5$	$x_{13+} = 53,9$	$x_{1++} = 146,2$
	$x_{11+}^2 = 382,90$	$x_{12+}^2 = 732,81$	$x_{13+}^2 = 505,73$	$x_{1++}^2 = 1621,44$
Femmes	$n_{21} = 6$	$n_{22} = 6$	$n_{23} = 9$	$n_{2+} = 21$
	$x_{21+} = 75,1$	$x_{22+} = 72,7$	$x_{23+} = 100,0$	$x_{2++} = 247,8$
	$x_{21+}^2 = 966,97$	$x_{22+}^2 = 894,45$	$x_{23+}^2 = 1139,72$	$x_{2++}^2 = 3001,14$
	$n_{+1} = 10$	$n_{+2} = 10$	$n_{+3} = 15$	$n_{++} = 35$
	$x_{+1+} = 113,9$	$x_{+2+} = 126,2$	$x_{+3+} = 153,9$	$x_{++++} = 394,0$
	$x_{+1+}^2 = 1349,87$	$x_{+2+}^2 = 1627,26$	$x_{+3+}^2 = 1645,45$	$x_{++++}^2 = 4622,58$

$$\begin{aligned}
 SCF_{\text{sexe}} &= \frac{1}{n_{1+}}(x_{1++})^2 + \frac{1}{n_{2+}}(x_{2++})^2 - \frac{1}{n}(x_{++++})^2 \\
 &= \frac{1}{14}146,2^2 + \frac{1}{21}247,8^2 - \frac{1}{35}394^2 \\
 &= 15,47
 \end{aligned}$$

Analyse de la variance à deux facteurs

Exemple

	Traitement A	Traitement B	Traitement C	
Hommes	$n_{11} = 4$	$n_{12} = 4$	$n_{13} = 6$	$n_{1+} = 14$
	$x_{11+} = 38,8$	$x_{12+} = 53,5$	$x_{13+} = 53,9$	$x_{1++} = 146,2$
	$x_{11+}^2 = 382,90$	$x_{12+}^2 = 732,81$	$x_{13+}^2 = 505,73$	$x_{1++}^2 = 1621,44$
Femmes	$n_{21} = 6$	$n_{22} = 6$	$n_{23} = 9$	$n_{2+} = 21$
	$x_{21+} = 75,1$	$x_{22+} = 72,7$	$x_{23+} = 100,0$	$x_{2++} = 247,8$
	$x_{21+}^2 = 966,97$	$x_{22+}^2 = 894,45$	$x_{23+}^2 = 1139,72$	$x_{2++}^2 = 3001,14$
	$n_{+1} = 10$	$n_{+2} = 10$	$n_{+3} = 15$	$n_{++} = 35$
	$x_{+1+} = 113,9$	$x_{+2+} = 126,2$	$x_{+3+} = 153,9$	$x_{++++} = 394,0$
	$x_{+1+}^2 = 1349,87$	$x_{+2+}^2 = 1627,26$	$x_{+3+}^2 = 1645,45$	$x_{++++}^2 = 4622,58$

$$\begin{aligned}
 SCF_{\text{traitement}} &= \frac{1}{n_{+1}}(x_{+1+})^2 + \frac{1}{n_{+2}}(x_{+2+})^2 + \frac{1}{n_{+3}}(x_{+3+})^2 - \frac{1}{n}(x_{++++})^2 \\
 &= \frac{1}{10}113,9^2 + \frac{1}{10}126,2^2 + \frac{1}{15}153,9^2 - \frac{1}{35}394^2 \\
 &= 33,66
 \end{aligned}$$

Analyse de la variance à deux facteurs

Exemple

On a donc $SCT = 187,27$, $SCF_{\text{sexe}} = 15,47$, $SCF_{\text{traitement}} = 33,66$, et les sommes des carrés de chacun des groupes sont :

	Traitement A	Traitement B	Traitement C
Hommes	$SC_{11} = 6,54$	$SC_{12} = 17,25$	$SC_{13} = 21,53$
Femmes	$SC_{21} = 26,97$	$SC_{22} = 13,57$	$SC_{23} = 28,61$

En sommant on a $SCR = 114,46$.

On est prêts à remplir la table d'anova.

Analyse de la variance à deux facteurs

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Sexe	15,47	1		
Traitement	33,66	2		
Interaction				
Résidus	114,46	29		
Total	187,27	34		

Analyse de la variance à deux facteurs

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Sexe	15,47	1		
Traitement	33,66	2		
Interaction	23,67	2		
Résidus	114,46	29		
Total	187,27	34		

Analyse de la variance à deux facteurs

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Sexe	15,47	1	15,47	$F = 3,92$
Traitement	33,66	2	16,83	$F = 4,26$
Interaction	23,67	2	11,83	$F = 2,99$
Résidus	114,46	29	3,95	
Total	187,27	34		

Analyse de la variance à deux facteurs

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Sexe	15,47	1	15,47	$F = 3,92$
Traitement	33,66	2	16,83	$F = 4,26$
Interaction	23,67	2	11,83	$F = 2,99$
Résidus	114,46	29	3,95	
Total	187,27	34		

Après comparaison aux valeurs critiques : $F_{0,95}^{1,29} = 4,2$ et $F_{0,95}^{2,29} = 3,3$, on constate que seul l'effet du traitement est significatif.

Analyse de la variance à deux facteurs

Cas particulier : une observation par cellule

Si tous les n_{ij} sont égaux à 1, on a $n = pq$ et dans les calculs qui précèdent on trouve $SCR = 0$ avec 0 ddl. On ne peut pas réaliser les tests ainsi !
La solution est de postuler qu'il n'y a pas d'interaction, que H_{0AB} est vrai.

$$X_{ij} = \mu + \alpha_i + \beta_j + E_{ij}.$$

On peut alors prendre SCF_{AB} comme somme de carrés résiduels (à $(p-1)(q-1)$ ddl) (c'est bien la somme des carrés résiduels du modèle où les γ_{ij} sont nuls), et la noter $SCR = SCF_{AB}$.

On a alors $SCT = SCF_A + SCF_B + SCR$ et on peut tester H_{0A} au moyen de $F_A = \frac{CMF_A}{CMR} \sim F(p-1, n-p-q+1)$ et H_{0B} au moyen de $F_B = \frac{CMF_B}{CMR} \sim F(q-1, n-p-q+1)$.

Analyse de la variance à deux facteurs

Exemple

Patient	P_1	P_2	P_3	P_4	P_5	P_6
Mesure 1	5,2	5,5	7,1	5,9	5,0	7,1
Mesure 2	6,0	5,3	8,0	5,8	6,5	7,8
Mesure 3	6,5	6,5	8,1	6,0	6,5	7,6
x_{+j+}	17,7	17,3	23,2	17,7	18,0	22,5
x_{+j+}^2	105,29	100,59	180,02	104,45	109,50	169,01

P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	x_{i++}	x_{i++}^2
6,4	7,4	6,5	7,0	6,5	6,0	75,6	483,14
8,1	7,3	6,7	6,5	6,6	6,6	81,2	557,98
7,5	7,5	6,6	7,1	6,6	6,7	83,2	581,04
22,0	22,2	19,8	20,6	19,7	19,3	240,0	
162,82	164,30	130,70	141,66	129,37	124,45		1622,12

Analyse de la variance à deux facteurs

Exemple

Patient	P_1	P_2	P_3	P_4	P_5	P_6
Mesure 1	5,2	5,5	7,1	5,9	5,0	7,1
Mesure 2	6,0	5,3	8,0	5,8	6,5	7,8
Mesure 3	6,5	6,5	8,1	6,0	6,5	7,6
x_{+j+}	17,7	17,3	23,2	17,7	18,0	22,5
x_{+j+}^2	105,29	100,59	180,02	104,45	109,50	169,01

P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	x_{i++}	x_{i++}^2
6,4	7,4	6,5	7,0	6,5	6,0	75,6	483,14
8,1	7,3	6,7	6,5	6,6	6,6	81,2	557,98
7,5	7,5	6,6	7,1	6,6	6,7	83,2	581,04
22,0	22,2	19,8	20,6	19,7	19,3	240,0	
162,82	164,30	130,70	141,66	129,37	124,45		1622,12

La somme des carrés factoriels associée à la mesure est

$$\begin{aligned}
 SCF_{\text{mesure}} &= \frac{1}{12} ((x_{1++})^2 + (x_{2++})^2 + (x_{3++})^2) - \frac{1}{n}(x_{+++})^2 \\
 &= \frac{1}{12}(75,6^2 + 81,2^2 + 83,2^2) - \frac{1}{36}240^2 \\
 &= 2,587.
 \end{aligned}$$

Analyse de la variance à deux facteurs

Exemple

Patient	P_1	P_2	P_3	P_4	P_5	P_6
Mesure 1	5,2	5,5	7,1	5,9	5,0	7,1
Mesure 2	6,0	5,3	8,0	5,8	6,5	7,8
Mesure 3	6,5	6,5	8,1	6,0	6,5	7,6
x_{+j+}	17,7	17,3	23,2	17,7	18,0	22,5
x_{+j+}^2	105,29	100,59	180,02	104,45	109,50	169,01

P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	x_{i++}	x_{i++}^2
6,4	7,4	6,5	7,0	6,5	6,0	75,6	483,14
8,1	7,3	6,7	6,5	6,6	6,6	81,2	557,98
7,5	7,5	6,6	7,1	6,6	6,7	83,2	581,04
22,0	22,2	19,8	20,6	19,7	19,3	240,0	
162,82	164,30	130,70	141,66	129,37	124,45		1622,12

La somme des carrés factoriels associée aux patients est

$$\begin{aligned}
 SCF_{\text{patients}} &= \frac{1}{3} ((x_{+1+})^2 + \dots + (x_{+12+})^2) - \frac{1}{n} (x_{++++})^2 \\
 &= \frac{1}{3} (17,7^2 + 17,3^2 + 23,2^2 + \dots) - \frac{1}{36} 240^2 \\
 &= 16,06.
 \end{aligned}$$

Analyse de la variance à deux facteurs

Exemple

Patient	P_1	P_2	P_3	P_4	P_5	P_6
Mesure 1	5,2	5,5	7,1	5,9	5,0	7,1
Mesure 2	6,0	5,3	8,0	5,8	6,5	7,8
Mesure 3	6,5	6,5	8,1	6,0	6,5	7,6
x_{+j+}	17,7	17,3	23,2	17,7	18,0	22,5
x_{+j+}^2	105,29	100,59	180,02	104,45	109,50	169,01

P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	x_{i++}	x_{i++}^2
6,4	7,4	6,5	7,0	6,5	6,0	75,6	483,14
8,1	7,3	6,7	6,5	6,6	6,6	81,2	557,98
7,5	7,5	6,6	7,1	6,6	6,7	83,2	581,04
22,0	22,2	19,8	20,6	19,7	19,3	240,0	
162,82	164,30	130,70	141,66	129,37	124,45		1622,12

La somme des carrés totaux est

$$SCT = x_{+++}^2 - \frac{1}{n}(x_{+++})^2 = 1622,16 - \frac{1}{36}240^2 = 22,16.$$

Analyse de la variance à deux facteurs

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Mesure	2,59	2		
Patients	16,06	11		
Résidus				
Total	22,16	35		

Analyse de la variance à deux facteurs

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Mesure	2,59	2		
Patients	16,06	11		
Résidus	3,51	22		
Total	22,16	35		

Analyse de la variance à deux facteurs

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Mesure	2,59	2	1,29	$F = 8,13$
Patients	16,06	11	1,46	$F = 9,18$
Résidus	3,51	22	0,16	
Total	22,16	35		

Analyse de la variance à deux facteurs

Exemple

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Mesure	2,59	2	1,29	$F = 8,13$
Patients	16,06	11	1,46	$F = 9,18$
Résidus	3,51	22	0,16	
Total	22,16	35		

Au seuil 5% la valeur critique pour l'effet mesure est (avec 2 et 22 ddl) vaut environ 3,44, et pour l'effet patient (avec 11 et 22 ddl) 2,25. On conclut que les deux facteurs ont un effet.