Les Concepts fondamentaux de Statistiques

En 3 leçons, nous allons essayer de passer en revue
l'estimation ponctuelle et par intervalle
et les généralités sur les tests d'hypothèses.

IN2P3
Institut National de Physique Nucléaire et de Physique des Particules

Irfu
cea
saclay

La statistique classique date de la fin du 18^{eme} siècle
GAUSS, BERNOUILLI, BAYES, ...

Il y a eu deux "explosions".

1. au début du 20^{eme} siècle, liée au développement des probabilités modernes:
   KOLMOGOROV, FISHER, PEARSON, DARMOIS
2. depuis 50 ans, grâce aux ordinateurs.

On peut distinguer 4 grands domaines :

1. La statistique descriptive
   recensements, mesures météo, résultats bruts des sondages
2. La statistique explicative : estimation, corrélations entre variables
3. La statistique décisionnelle : tests d'hypothèses
4. La statistique prévisionnelle (le temps est un param. supp.)
   météo, santé

- la Bible : Kendall et Stuart (Ch. Griffin)
  The advanced theory of statistics, 3 tomes.
- ma faveur : Frederick James (World Scientific)
  Statistical Methods in Experimental Physics.
- plutôt math. Pugatchev (Mir)
  Théorie des probas et statistique mathématique.
- pour les débrouillés Les conférences PHYSTAT
  `http://www.physics.ox.ac.uk/phystat05/reading.htm`

Il y a deux "écoles" de statisticiens, qui ne sont pas d'accord sur certains principes de base !

- Les Bayésiens ou modernes
- Les anti-Bayésiens ou fréquentistes

EN PROBAS, le théorème de Bayes date de 1763, et tout le monde s'en sert.

$$\mathcal{P}(A|B) = \mathcal{P}(B|A)\mathcal{P}(A)/\mathcal{P}(B)$$

$$p(x|y) = q(y|x)f(x)/g(y)$$

EN STAT, les bayésiens considèrent les paramètres comme une v.a.

$$f(\theta|x) = f(x|\theta)f(\theta)/f(x)$$

Le modèle de l'URNE avec N boules dont B sont blanches

Probabilités	Statistiques
Données : B, N connus p =B/N connu	p, B, N inconnus. On a effectué n tirages et obtenu k boules blanches.
Problèmes : * Trouver la loi de proba. du nombre de boules blanches en n tirages, avec ou sans remise. Moments de ce nombre. * Trouver la loi du nombre de tirages nécéssaires pour obtenir la première boule blanche, esp. math. de ce nombre, ...	* Affecter à p une valeur *raisonnable* estimation ponctuelle * Trouver un intervalle où p a de bonnes chances de se trouver. estimation par intervalle * Décider si la vraie valeur de p est inférieure à telle valeur donnée (ou comprise entre 2 bornes) tests d'hypothèses
Conclusions : *certaines* et en principe avec une *précision arbitraire.*	

Le modèle de l'URNE avec N boules dont B sont blanches

Probabilités	Statistiques
Données : B, N connus p =B/N connu	p, B, N inconnus. On a effectué n tirages et obtenu k boules blanches.
Problèmes : * Trouver la loi de proba. du nombre de boules blanches en n tirages, avec ou sans remise. Moments de ce nombre. * Trouver la loi du nombre de tirages nécéssaires pour obtenir la première boule blanche, esp. math. de ce nombre, ...	* Affecter à p une valeur *raisonnable* estimation ponctuelle * Trouver un intervalle où p a de bonnes chances de se trouver. estimation par intervalle * Décider si la vraie valeur de p est inférieure à telle valeur donnée (ou comprise entre 2 bornes) tests d'hypothèses
Conclusions : *certaines* et en principe avec une *précision arbitraire.*	Impossible de donner une réponse *certaine.* (notion de risque). Plus la réponse est *précise*, plus grande la proba qu'elle soit *fausse.*

Définitions

Soit une grandeur $\mathcal{G}$ de valeur réelle (inconnue) G. Pour déterminer G, on fait un certain nombre de mesures portant sur G, puis un calcul (estimation) aboutit au résultat g.

Définition : $\epsilon = G - g$ est l'erreur sur G.

On peut toujours se ramener au cas où c'est la grandeur elle-même que l'on mesure. Posons donc : $X = G$, $x_0 = g_0 \ldots x = g$. x et $\epsilon = X - x$ sont des v.a.

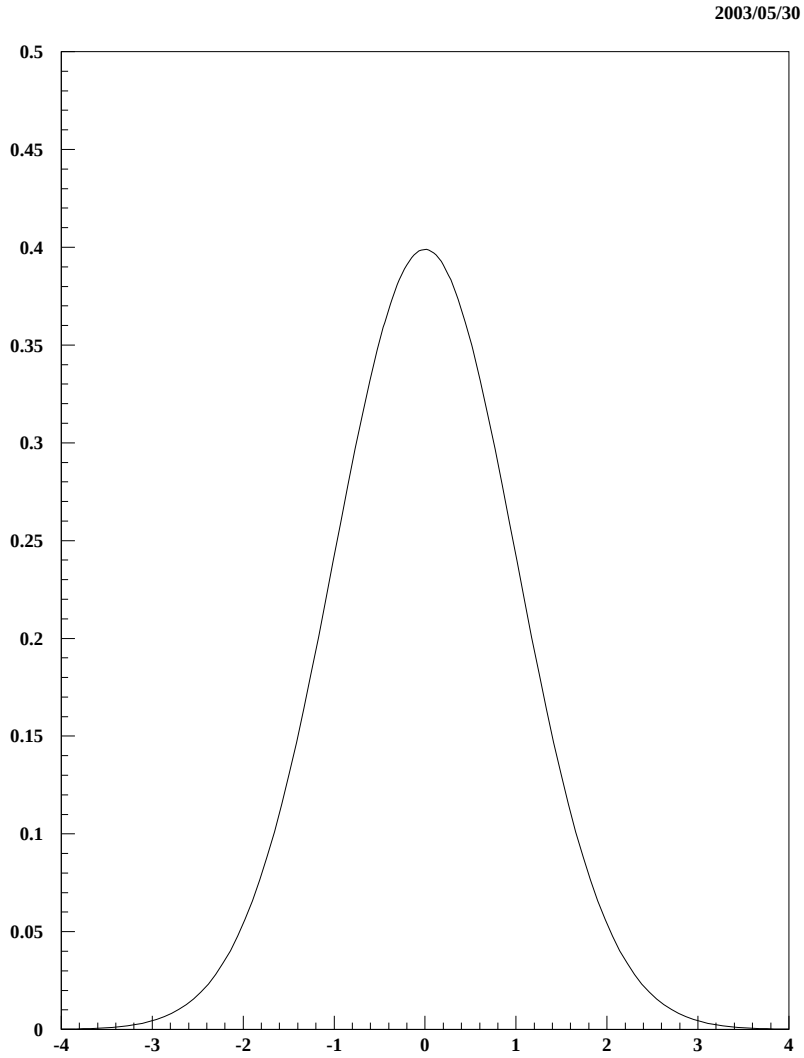
Définition : $\epsilon_s = E[\epsilon]$ est l'erreur systématique

Définition : $\epsilon_a = \epsilon - \epsilon_s$ est l'erreur accidentelle, ou statistique.

Etude de l'erreur accidentelle

La loi des Grands Nombres nous dit que $\overline{x}_n = \frac{1}{n}\sum x_i$ converge vers $E[x]$, donc $\epsilon_a = E[x] - x = \lim_{n\to\infty} \overline{x}_n - x$. La dispersion de ϵ_a est donc la dispersion des x_i, mesures faites dans les mêmes conditions par le même instrument.

Définition : c'est ce qu'on appelle la dispersion de l'instrument



On associe à ϵ_a une densité, gaussienne en général, qu'on appelle courbe de dispersion de l'instrument. On appelle dispersion totale de l'instrument la quantité 4σ.
Plus la dispersion est faible, plus l'appareil est dit fidèle.2σ s'appelle la précision de l'appareil.
Dispersion (ou précision) s'estiment facilement, par l'estimateur usuel de σ

$$S_a^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$$

Etude de l'erreur systématique

Par sa définition, $\epsilon_s = E[\epsilon]$, ce n'est PAS une v.a.
Elle est liée à l'appareil : plus elle est petite, plus l'appareil est dit juste.
Pour l'estimer, on utilise (par la pensée en général) plusieurs instruments de mesure du même type, pour mesurer la même quantité X, et on admet que les appareils sont fabriqués de façon à ce que l'erreur systématique soit nulle en moyenne. Ceci revient à rendre ϵ_s aléatoire !

Soit donc k appareils, et un (grand) nombre de mesures sur chaque appareil : on a donc k moyennes $\overline{x}_j$, où j varie de 1 à k. Pour chaque appareil, $\overline{x}_j \xrightarrow{\mathcal{P}} X - \epsilon_j^s$. Par suite, (puisque $E[\epsilon_s] = 0$), $\frac{1}{k}\sum \epsilon_j^s \xrightarrow{\mathcal{P}} 0$ et donc $\hat{x} = \frac{1}{k}\sum \overline{x}_j \xrightarrow{\mathcal{P}} X$. L'estimateur de ϵ_j^s est donc $\hat{x} - \overline{x}_j$ et la variance de ϵ_s est donc estimée par $\hat{\sigma_s^2} = \frac{1}{k-1}\sum(\overline{x}_j - \hat{x})^2$

Résumé : Le résultat d'une mesure sur X s'écrit par convention

$$\hat{x} \pm 2\hat{\sigma_s} \pm 2\hat{\sigma_a}$$

et ϵ_s et ϵ_a sont indépendantes en probabilité.

N.B. : sémantique ; estimation et mesure au sens large.

Définitions (rappels)

- Echantillon (et la (ou les) ddp sous-jacentes)
- Estimation ponctuelle / par intervalle
- Estimation paramétrique / non paramétrique
- Vraisemblance d'un échantillon, fonction de vraisemblance.
- Information (selon Fisher)
- Statistique (changement de variable)
- Statistique exhaustive (minimale)
- Estimateur et ses propriétés : convergence, biais, efficacité et robustesse.

Echantillon

- au premier niveau (stat. descriptive), c'est la collection des mesures brutes. On parle de taille d'un échantillon, n.
- mais c'est surtout la réalisation d'une suite de v.a. indépendantes.

Notations

- $\{x_i\}$ représente l'ensemble des valeurs numériques des n mesures
- $\{X_i\}$ l'ensemble des n v.a. sous-jacentes
- $f_i(x)$ désigne la ddp de X_i : elle peut être
  - totalement précisée
  - analytiquement connue : on la note alors $f_i(x|\theta)$
  - inconnue

Estimation

- L'estimation ponctuelle paramétrique consiste à déduire de l'échantillon une valeur plausible pour θ.
- L'estimation par intervalle consiste à trouver un intervalle (plus généralement un domaine) où θ a une probabilité donnée de se trouver.
- L'estimation non paramétrique, c'est lorsque f_i est inconnu.

Vraisemblance

La fonction de vraisemblance d'un échantillon, c'est la ddp de la suite

$$\mathcal{L}(x) = \prod_{i=1}^{n} f_i(x)$$

En estimation paramétrique, elle est notée $\mathcal{L}(x|\theta)$
On appelle vraisemblance d'un échantillon la valeur (numérique) que prend la fonction de vraisemblance pour l'échantillon dont on dispose.
En estimation paramétrique, c'est une fonction de θ.

Information (selon Fisher)

L'idée est de permettre la plus grande réduction possible des données, avec une perte d'information minimale. Cette quantité doit donc dépendre

- de l'échantillon : plus n est grand, plus elle est grande
- de la ddp sous-jacente, donc de la vraisemblance, et bien sûr des paramètres.
- elle doit pouvoir être calculée sur "l'échantillon réduit".

Définitions :

1. : on appelle "statistique" toute fonction de l'échantillon.
   exemple : la moyenne empirique $\bar{x} = 1/n \sum x_i$
2. : l'information est définie comme la matrice

$$I_\theta(x) = E[(\frac{\partial \ln \mathcal{L}(x|\theta)}{\partial \theta})^2]$$

Propriétés

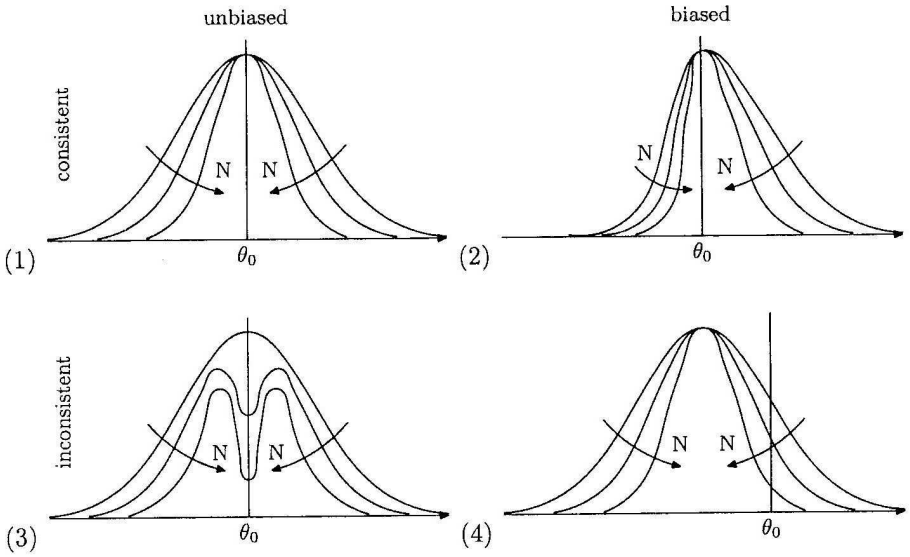
- $I_\theta(x) > 0$
- $I_n = nI_1$
- $I_\theta(x) = -E[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}]$

- Une statistique $T = \phi(X)$ est dite exhaustive pour le paramètre θ si la densité de probabilité conditionnelle de X à T fixé, soit $h(X|T)$, est indépendante de θ.
- N.B.1 T, X, θ peuvent être multidimensionnels.
- N.B.2 Si T est une statistique exhaustive pour θ, toute fonction strictement monotone de T est aussi une statistique exhaustive pour θ.
- N.B.3 La statistique $T = X$ est évidemment exhaustive pour tout paramètre, mais ce n'est pas très utile.
- N.B.4 Si l'on veut traduire la définition en langage utile, il faut la formuler de la façon suivante : il y a autant d'information sur θ dans T que dans les données X. On voit alors clairement que le problème de la réduction des données est équivalent à celui d'une recherche de statistique exhaustive.

Soit un n-échantillon $\{x_i\}$ d'une v.a. X de ddp $f(x|\theta)$
Nous noterons θ_0 la vraie valeur (inconnue) de θ.
On appelle estimateur de θ toute fonctionnelle (et surtout son résultat) de l'échantillon qui est "proche" (au sens des proba.) de θ_0. On le note $\hat{\theta}$. Ces estimateurs (qui sont donc des statistiques) possèdent 4 propriétés importantes

- La convergence (en anglais "consistency")
- Le biais
- L'efficacité
- La robustesse
- Convergence : on dit que $\hat{\theta}$ est convergent si $\lim_{n\to\infty} \hat{\theta} = \theta_0$.
  (La convergence "en probabilité" suffit). Il s'agit donc d'une propriété asymptotique.

- Biais : par définition, c'est $b(\hat{\theta}) = E[\hat{\theta} - \theta_0] = \int (\hat{\theta} - \theta_0) g(\hat{\theta}) d\hat{\theta}$



- Un estimateur convergent est asymptotiquement non-biaisé
- Théorème : un estimateur asymptotiquement non biaisé et dont la variance tend vers 0 (avec n) est convergent.

Efficacié d'un estimateur

Intuitivement, la précision d'un estimateur doit être reliée à la largeur de sa distribution, donc à $V[\hat{\theta}]$. Cette variance est bornée inférieurement.

- Inégalité de Cramer - Rao :

$$V[\hat{\theta}] \geq \frac{(1 + db/d\theta)^2}{I_\theta[\hat{\theta}]}$$

  La variance d'un estimateur non biaisé est borné inférieurement par l'inverse de l'information qu'il porte.

- Estimateur de variance minimum : c'est un estimateur non biaisé qui atteint sa borne, c'est à dire que $V[\hat{\theta}]I_\theta[\hat{\theta}] = 1$. Un théorème existe, qui dit que ceci est équivalent à

$$\frac{\partial \ln g(\hat{\theta}|\theta)}{\partial \theta} = A(\theta)(\hat{\theta} - \theta)$$

- Estimateur efficace : c'est un estimateur de variance minimum qui de plus est une statistique exhaustive pour θ.

Un estimateur est dit parfaitement robuste si sa ddp $g(\hat{\theta})$ ne dépend pas de $f(x|\theta)$. Autant dire qu'il n'existe pas de tels estimateurs. La robustesse est une propriété assez qualitative : on dit qu'un estimateur est robuste si $g(\hat{\theta})$ est peu sensible à de petites variations de f. Ceci est utile lorsqu'on n'est pas très sûr de la forme analytique choisie pour f, en particulier dans les "queues" de distribution.

Les estimateurs "Moyenne tronquée" et "Moyenne de Winsor" sont des estimateurs plus robustes que la moyenne empirique pour estimer le centre d'une distribution sysmétrique.

Estimation ponctuelle paramétrique

De la Théorie à la pratique

- Premiers exemples : biais vs précision.
- Le théorème général
- Maximum de vraisemblance
- Est. à W-distance minimale : (moindres carrés)
- Ajustement d'histogrammes
- Estimation robuste.
- Estimation ponctuelle en pratique.

Soit un n-échantillon de ddp sous-jacente $f(x)$ quelconque, mais telle que $\mathbf{E}[x] = \mu$ et $\mathbf{V}[x] = \sigma^2$ existent et sont finies.

- $\overline{x}$ est un estimateur convergent, sans biais, de μ.
  - sans biais : $\mathrm{E}[\overline{x}] = \frac{1}{n}\sum_1^n \mathrm{E}[x] = \mu$
  - convergent : $\mathrm{V}[\overline{x}] = \frac{1}{n^2}\sum_1^n \mathrm{V}[x] = \sigma^2/n$
- s^2 est un estimateur convergent, biaisé, de σ^2.

$$s^2 = \frac{1}{n}\sum_1^n (x_i - \overline{x})^2 \quad \text{variance empirique}$$

  - biais : remplaçons $x_i - \overline{x}$ par $(x_i - \mu) + (\mu - \overline{x})$ dans s^2. Il vient :

$$s^2 = \frac{1}{n}\sum [(x_i - \mu)^2 + 2(x_i - \mu)(\mu - \overline{x}) + (\mu - \overline{x})^2]$$

$$s^2 = \frac{1}{n}\sum (x_i - \mu)^2 - (\mu - \overline{x})^2$$

    Donc : $\mathrm{E}[s^2] = \sigma^2$ - $\mathrm{V}[\overline{x}] = \frac{n-1}{n}\sigma^2$
    s^2 est donc biaisé, mais asymptotiquement non biaisé.
  - exo : calculer $\mathrm{V}[s^2]$ et en déduire la convergence.

Premier exemple (2)

- corollaire : $s'^2 = \frac{n}{n-1} s^2$ est un estimateur convergent sans biais de σ^2.
  - évident
  - exo : comparer $\mathrm{V}[s^2]$ et $\mathrm{V}[s'^2]$.
- si μ est connu, il est préférable d'estimer σ^2 par $\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \mu)^2$.
- cas où $f(x) = \mathcal{N}(\mu, \sigma^2)$

$$\overline{x} \text{ suit la loi } \mathcal{N}(\mu, \sigma^2/n)$$
$$ns^2/\sigma^2 \text{ suit la loi de } \chi^2 \text{ à (n-1) ddl.}$$

  - Théorème de Fisher-Geary
    $f(x) = \mathcal{N}(\mu, \sigma^2) \Longleftrightarrow \overline{x}$ et s^2 sont des v. a. indépendantes.
  - exo : montrer que $\mathrm{V}[s'^2] > \mathrm{V}[s^2] > \mathrm{V}[\hat{\sigma}^2]$

Ceci se généralise, c'est la méthode des moments : on obtient des estimateurs convergents mais biaisés. Facile et rapide.

Le théorème général.

Rappels : Toute construction d'estimateurs repose sur l'utilisation de la loi des grands nombres et du théorème central limite.

- soient n v.a. indépendantes X_i, avec $\mathrm{E}[X_i] = \mu_i$ et $\mathrm{V}[X_i] = \sigma_i^2$
- et la condition $\lim_{n\to\infty} \left(\frac{1}{n^2} \sum \sigma_i^2\right) = 0$, alors
- Grands Nombres : $\sum (x_i - \mu_i) \overset{m.q.}{\to} 0$
- TCL : $\sum (x_i - \mu_i) / \sqrt{\sum \sigma_i^2} \overset{\mathcal{L}}{\to} \mathcal{N}(0,1)$

Théorème : S'il existe une fonction ξ (des observations), telle que :

- $\xi(\{x_i\}, \theta_0) \to 0$
- ξ suit asymptotiquement une loi normale de variance σ^2
- $\frac{\partial \xi}{\partial \theta} \neq 0$ pour $\theta = \theta_0$

alors, une racine (notée $\hat{\theta}$) de l'équation $\xi = 0$ est un estimateur convergent de θ, distribué asymptotiquement autour de θ_0 suivant une loi normale de variance $\mathrm{V}_\infty[\hat{\theta}] = \sigma^2 / \left[\frac{\partial \xi}{\partial \theta}\right]^2_{\theta=\theta_0}$

- Définition : (1921 par R. A. Fisher)

  L'estimateur MV est défini par $\hat{\theta}$ tel que

  $$\mathcal{L}([x]|\hat{\theta}) = \max_{\theta} \mathcal{L}([x]|\theta)$$

  C'est la fonction de θ que l'on considére, pour des valeurs fixées de l'échantillon (les x_i). C'est choisir la valeur de θ qui maximise la probabilité d'observer ... ce que j'ai effectivement observé.

- Conditions d'utilisation : $\mathcal{L}$ deux fois dérivable par rapport à θ.
  Attention : si θ a un domaine de définition, si $\mathcal{L}$ est maximum sur le bord, sans que $\frac{\partial \mathcal{L}}{\partial \theta}$ soit nul, l'estimateur MV est biaisé.

$$\begin{array}{ccc} \frac{\partial \mathcal{L}}{\partial \theta} = 0 & \text{extremum} & \frac{\partial \ln \mathcal{L}}{\partial \theta} = 0 \\ & \text{(local)} & \\ \frac{\partial^2 \mathcal{L}}{\partial \theta^2} < 0 & \text{maximum} & \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} < 0 \end{array}$$

Propriétés du Maximum de Vraisemblance.

- S'il existe un estimateur (non biaisé) de variance minimum, alors il est donné par le M. V.
  - var. min. $\Leftrightarrow \frac{\partial \mathcal{L}}{\partial \theta} = A(\theta)(t-\theta)$
  - Max Vrai. $\Leftrightarrow \frac{\partial \mathcal{L}}{\partial \theta} = 0$ soit $\hat{\theta} = t$
- Comportement asymptotique le théorème général marche avec

$$\xi(x_i, \theta) = \frac{\partial \mathcal{L}}{\partial \theta}$$

Par suite, $\hat{\theta}$ est convergent, et sa ddp asymptotique est $\mathcal{N}(\theta_0, 1/\sqrt{I_x})$. Il est donc asymptotiquement efficace.

Exemple pratique (Echantillon fini et θ_0 inconnu)

Soit X une variable aléatoire bidimensionnelle (x, y) telle que : x suit $\mathcal{N}(0,1)$ et y suit $\mathcal{N}(0,1)$, mais telle que $\text{cov}(x,y) = \rho$ est le paramétre à estimer à partir d'un échantillon de taille n.

$$f(X|\rho) = (1/2\pi\sqrt{1-\rho^2})\exp[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}]$$

$$\ln \mathcal{L} = \sum_{i=1}^{n} \ln f(X_i|\rho)$$

$$\frac{d\ln\mathcal{L}}{d\rho} = 0 \Leftrightarrow \rho^3 - S_{12}\rho^2 - (1 - S_{11} - S_{22})\rho - S_{12} = 0$$

avec $nS_{11} = \sum x_i^2$; $nS_{12} = \sum x_i y_i$; $nS_{22} = \sum y_i^2$
Le changement de variable $\rho = \lambda + S_{12}/3$ ramène l'equation sous la forme $\lambda^3 + p\lambda + q = 0$, qui a une solution si $4p^3 + 27q^2 > 0$, a fortiori si $p > 0$.
Or $p = S_{11} + S_{22} - 1 - S_{12}^2/3 \rightarrow^{\mathcal{P}} 1 + 1 - 1 - \rho^2/3 > 0$ car $|\rho| \leq 1$

Maximum de Vraisemblance (4)

Quelques "trucs".

- Si le calcul analytique est difficile ou impossible, la résolution de l'équation $\omega(\theta) = \frac{\partial \ln \mathcal{L}}{\partial \theta} = 0$ nécessite un calcul par itérations. Soit θ_1 une valeur approchée (fixée) de $\hat{\theta}$, $\omega(\theta_1)$ calculable (voisin de 0), donc
  $\mathrm{E}[\omega]_{\theta=\theta_1} = \omega(\theta_1)$
  Cherchons une meilleure approximation, θ_2 (inconnue) :

$$\begin{cases} \omega(\theta_2) = \omega(\theta_1) + (\theta_2 - \theta_1)\omega'(\theta_1) \\ \mathrm{E}[\omega]_{\theta=\theta_2} = 0 \end{cases}$$

  En prenant l'espérance de la première équation, il vient :

$$0 = \omega(\theta_1) + (\theta_2 - \theta_1)\mathrm{E}[\omega']$$

  soit $\theta_2 = \theta_1 + \omega(\theta_1)/\mathrm{I}_X(\theta_1)$

- Variance de l'estimateur MV : on le connait asymptotiquement, mais on l'estime en pratique par l'inverse de l'information pour $\hat{\theta}$ ou même par l'inverse de l'information "observée" $\widehat{\mathrm{V}[\hat{\theta}]} \simeq -1/\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}$ pour $\theta = \hat{\theta}$

Ou Moindres Carrés ! (Gauss vers 1820)

- Soit un échantillon de n mesures $\{y_i\}$, correspondant à des valeurs différentes (fixes) d'une "entrée" x_i, et supposons qu'il existe une loi (au sens de la physique) ou un modêle liant mesures et entrés. On écrit donc (vectoriellement) $y = h(x, \theta) + \epsilon$. h traduit le "déterminisme", ϵ l'aléatoire. On parle alors d'estimation des paramètres θ ou d'ajustement du modèle.
- Soyons bayésiens, avec $\pi(\theta) = \text{Cte}$, et supposons que les mesures sont gaussiennes, de moyennes h et de variance σ_i^2 et indépendantes. Alors $\pi(\theta|x_i, y_i) \propto \prod \mathcal{N} \propto \exp\left[-\sum \frac{(y_i - h(x_i,\theta))^2}{2\sigma_i^2}\right]$ Le maximum de π (maximum de vraisemblance) est obtenu pour le minimum de la somme : c'est les moindres carrés. De plus π est une densité normale.

Méthode : on minimise la norme de ϵ, soit S^2.
La norme est définie par un tenseur métrique W

$W = I$	$S^2 = \sum [y_i - h(x_i, \theta)]^2$
W diagonale	$S^2 = \sum [y_i - h(x_i, \theta)]^2 / \sigma_i^2$
W quelconque	$S^2 = (y - H) W (h - H)$

- Si H est quelconque, ce n'est pas plus simple qu'un max $\mathcal{L}$
- Si H est linéaire (en θ), soit $H = A\theta$, alors

$$\hat{\theta} = (\tilde{A} W A)^{-1} \tilde{A} W y$$

- Si $W = V^{-1}$, alors $V[\hat{\theta}] = (\tilde{A} V^{-1} A)^{-1}$

Estimation à W distance minimale (3)

Liaison avec le théorème général.

- Cet estimateur est obtenu avec un ξ tel que

$$\xi(y_i, \theta) = \frac{1}{2}\frac{\partial S^2}{\partial \theta} = -\widetilde{\frac{\partial H}{\partial \theta}} W(y - H)$$

- Si H est linéaire en θ, $(H = A\theta)$, l'estimateur $\hat{\theta}$ est donc donné par $\xi(\hat{\theta}) = 0$, soit

$$-\tilde{A}Wy + \tilde{A}WA\hat{\theta} = 0$$

et la variance asymptotique de $\hat{\theta}$ par

$$\mathrm{V}_\infty[\hat{\theta}] = (\tilde{A}WA)^{-1}\tilde{A}WV\tilde{W}A(\tilde{A}\widetilde{WA})^{-1}$$

où V est la matrice de var.-covar. des y_i

- Cette variance se simplifie fortement si on choisit $W = V^{-1}$ et devient :

$$\mathrm{V}_\infty[\hat{\theta}] = (\tilde{A}V^{-1}A)^{-1}$$

Remarques.

- Si y est normal, S^2 est un χ^2 à n ddl (taille de l'échantillon) lorsque θ est libre, mais à $n-m$ ddl (m dimension de θ) lorsqu'il atteint son minimum, pour $\theta = \hat{\theta}$. $\hat{\theta}$ et S^2 sont des v.a. indépendantes, et $\hat{\theta}$ est un estimateur de variance minimale.
- Si y n'est pas normal, RIEN de tout cela.
- Moindres carrés non linéaires : lourd et long, car il faut linéariser et itérer.
- Moindres carrés en présence de contraintes : pas simple
  Si les paramètres doivent vérifier une équation $g(\theta) = 0$, on passe par les multiplicateurs de Lagrange. Maximum de Vraisemblance ou Moindres Carrés restent équivalents.
  Si les paramètres sont bornés ($g(\theta) \geq 0$), il faut mieux faire comme s'ils n'étaient pas bornés, de façon à éviter les biais.

Relations avec la notion d'innovation (résidus).

- $r = y - \hat{y} = y - A\hat{\theta}$ Remplaçons $\hat{\theta}$ par sa formule :
  $r = \left[1 - A(\tilde{A}WA)^{-1}\tilde{A}W\right] y = By$
- B est un opérateur de projection, orthogonal à A.

$$B^2 = 1 - 2A(\tilde{A}WA)^{-1}\tilde{A}W + A(\tilde{A}WA)^{-1}\tilde{A}WA(\tilde{A}WA)^{-1}\tilde{A}W$$

$$B^2 = 1 - A(\tilde{A}WA)^{-1}\tilde{A}W = B$$

$$BA = A - A(\tilde{A}WA)^{-1}\tilde{A}WA = 0$$

- $trB = tr1_n - tr[A(\tilde{A}WA)^{-1}\tilde{A}W] = n - m$
- $\epsilon = y - A\theta = r + A(\hat{\theta} - \theta)$ et par suite

$$S^2 = \tilde{\epsilon}W\epsilon = \tilde{r}Wr + (\hat{\theta} - \theta)(\tilde{A}WA)(\hat{\theta} - \theta)$$

L'histogramme

- C'est un moyen pratique de visualiser un échantillon de grande taille.
- Il consiste à remplacer $n\{x_i\}$ par $k\{n_j\}$
- Soit A le domaine de définition des x_i, et fixons k.
- On effectue une partition de A en k classes (bins,boîtes) (de a_0 à a_k)
- Les n_j sont alors les fréquences d'observation des x_i dans l'intervalle $[a_{j-1},a_j]$
- Au niveau du dessin, les rectangles ont des aires prop. aux n_j
- Il y a une infinité d'histogrammes visualisant le même échantillon. On peut faire varier k et les bornes des boîtes.
- La simplicité pousse à choisir des boîtes de largeur constante, indépendante de j. La hauteur des boîtes est alors prop. à n_j.
- On peut montrer que les histogrammes qui minimisent la perte d'information sont ceux pour lesquels $n_j \approx n/k$, indépendant de j.

1. n fixé : la v.a. (n_j), vecteur de k v.a. entières, suit une loi **multinomiale**.

$$P(n_1, n_2, \ldots, n_k) = n! \prod_{j=1}^{k} \frac{p_j^{n_j}}{n_j!}$$

Les p_j sont des paramètres, probabilité d'avoir une observation dans la boîte j.

$$\begin{array}{l|l} E[n_j] = np_j & \mathrm{cov}(n_i, n_j) = -np_ip_j \\ V[n_j] = np_j(1-p_j) & \mathrm{corr}(n_i, n_j) = -\sqrt{p_ip_j/(1-p_i)(1-p_j)} \end{array}$$

Si n est grand (et k aussi), $V[n_j] \approx np_j \approx n_j$

2. n libre $\equiv n$ v.a.
   Nous avons donc $k+1$ v.a. En négligeant les corrélations, chacune suit une loi de **Poisson**, et la ddp totale s'écrit :

$$P(n_1, n_2, \ldots, n_k) = \prod_{j=1}^{k} e^{-\mu_j} \frac{\mu_j^{n_j}}{n_j!}$$

   avec une relation avec le cas multinomial : $\mu_j = \mu p_j$ où μ est le paramètre de la loi de Poisson de n.

$$\begin{array}{l|l} E[n_j] = \mu_j & \mathrm{cov}(n_i, n_j) = 0 \\ V[n_j] = \mu_j & E[n] = V[n] = \mu \end{array}$$

3. Si $n \to \infty$, le TCL nous dit que la distribution est **multinormale**.
   Chacun des n_j suit une loi $\mathcal{N}(np_j, np_j)$.

1. Normalisation non ajustée (le modèle ne prédit pas n)
   n n'est pas une v.a. et $\sum n_j = n$ est une contrainte.
   Nous sommes dans le cas d'une distribution multinomiale avec des p_j qui dépendent du modèle, donc des paramètres à estimer. On peut :
   - Max. Vrais. : on maximise $\ln \mathcal{L} = \sum n_j \ln p_j(\theta)$
     C'est un estimateur convergent, asymptotiquement efficace et normal.
   - Moindres carrés : on minimise $S^2(\theta) = \sum \frac{(n_j - np_j(\theta))^2}{np_j(\theta)}$
     C'est un estimateur convergent, asymptotiquement normal. On remarque que le tenseur métrique W n'est plus indépendant de θ !
     $S^2(\hat{\theta})$ se comporte asymptotiquement comme un χ^2 à $(k - s - 1)$ ddl, où s est la dimension du vecteur θ.
   - Moindre carrés **modifiés** : pour ne pas avoir une dépendance en θ au dénominateur, on remplace $np_j(\theta)$ par n_j. On minimise donc $S^2(\theta) = \sum \frac{(n_j - np_j(\theta))^2}{n_j}$. W est maintenant aléatoire ! Pourtant, on a les mêmes propriétés que précédemment.
   - La convergence (TCL) est plus rapide pour le MV, la plus lente pour les Moindres carrés modifiés.

2. Normalisation ajustée : n est une variable aléatoire supplémentaire, et la vraisemblance suit une loi de Poisson. Les 3 estimateurs vus ci-dessus restent valables, mais $S^2(\theta)$ converge maintenant vers un χ^2 à $(k-s)$ ddl.