T.D. (+formulaire) du cours "Bruits et Signaux" 2ème année de Master Spécialité Astro-Astrophys

Frédéric Arenou
UMR 8111 du CNRS et Gepi (Observatoire de Paris)

Revision: 1.3 (préliminaire), Date: 2004/10/22 06:53:11
Mise à jour sur `http://wwwhip.obspm.fr/~arenou`

- « *La statistique est la première des sciences inexactes.* » E. et J. Goncourt
- « *Je ne crois qu'aux statistiques que j'ai falsifiées moi-même.* » W. Churchill
- « *Il ne faut pas utiliser les statistiques commes les ivrognes utilisent les réverbères : pour s'appuyer et non pour s'éclairer.* » Lord Thorneycroft
- « *Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales.* » G. Elgozy
- « *La statistique est un bikini. Ce qu'elle révèle est suggestif, ce qu'elle cache est vital.* » A. Kœstler

Des débuts difficiles...

- Newton (Principia Naturalis, 1687): la Terre est un ellipsoïde aplati aux pôles.
- Triangulation des Cassini: semble donner le résultat inverse
- 2 expéditions: Laponie et équateur
- 1740, Bouguer et La Condamine mesurent l'angle au zénith de ϵ Ori au Pérou
- Au bout d'un an: leur deux séries de mesures ne coïncident pas
- La Condamine suggère à Bouguer d'établir une *moyenne arithmétique*, pour annuler simultanément toutes les causes d'erreur en les faisant se compenser. Bouguer est horrifié: il s'agirait de tricher! A quoi servirait toute l'expédition si elle se terminait par une évaluation fantaisiste, fruit de la seule imagination? Jamais un savant digne de ce nom n'accepterait de donner un résultat artificiel, qui ne proviendrait pas de mesures directes de la Vérité...

Probabilités

- **Conventions**
  Pour une v.a. X et sa réalisation x, on notera souvent $f(x)$ au lieu de $f_X(x)$ sa densité de probabilité et $F(x)$ sa fonction de répartition.

  On s'intéressera essentiellement à des fonctions continues.

  On utilisera les lettres grecques pour les paramètres inconnus,

- **Fonction de répartition (distribution)**

$$F_X(x) = P(X \leq x), \quad x \in [-\infty, +\infty]$$

  ☞ Propriétés

$$F(x) \in [0,1], \quad F(-\infty) = 0, \quad F(+\infty) = 1$$
$$F(x) \leq F(x'), \quad \forall x \leq x'$$

- **Densité de probabilité (p.d.f.)**

$$f(x) = \frac{dF}{dx}$$

☞ densité marginale en X d'une loi $f(x,y)$

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

$$P(X \leq a) = \int_{-\infty}^{a} f_X(x)dx$$

☞ indépendance X et Y sont indépendantes $\Longleftrightarrow$

$$f(x,y) = f_X(x)f_Y(y), \quad \forall (x,y)$$

☞ densité conditionnelle

$$f(x \mid y) = \frac{f(x,y)}{f_Y(y)}$$

$$P(a \leq X \leq b \mid Y = y) = \int_a^b f(x \mid y)dx$$

Moments

- **Espérance mathématique**

$$E[X] = \int_{-\infty}^{+\infty} x f(x) dx = \mu$$

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

- **Variance**

$$\begin{array}{rcccl} \sigma^2(X) & = & E\left[(X - E[X])^2\right] & = & E[X^2] - E^2[X] \\ & = & \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx & = & \int_{-\infty}^{+\infty} x^2 f(x) dx - \mu^2 \end{array}$$

- **Écart-type** $\sigma(X)$

- **Covariance**

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E\left[(X - E[X])(Y - E[Y])\right] \\ &= \int_{-\infty}^{+\infty} (x - \mu_X)(y - \mu_Y) f(x,y) dx dy \\ &= \int_{-\infty}^{+\infty} xy f(x,y) dx dy - \mu_X \mu_Y \end{aligned}$$

Dans le cas multidimensionnel d'un vecteur $\mathbf{X} = (X_i)$, on introduit la matrice de variance-covariance

$$\mathbf{V} = E\left[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T\right] = (\mathrm{Cov}(X_i, X_j))$$

dont la diagonale est formée des variances, et qui est définie non-négative.

- **Corrélation**

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

Si X, Y sont des v.a. et a, b des réels, on a

$$\text{Var}(aX + bY) = a^2\sigma^2(X) + 2ab\rho(X,Y)\sigma(X)\sigma(Y) + b^2\sigma^2(Y)$$

- **Moment d'ordre r**

$$E[X^r] = \int_{-\infty}^{+\infty} x^r f(x)dx$$

et le moment centré est :

$$E[(X-\mu)^r] = \int_{-\infty}^{+\infty} (x-\mu)^r f(x)dx$$

- **Quantiles**
  Q_α est un quantile $(1-\alpha)$ si $P(X \leq Q_\alpha) = 1-\alpha$. En particulier, la médiane est $Q_{0.5}$

- **Mode**
  C'est un maximum de la pdf (il peut y en avoir plusieurs).

L'erreur sur la moyenne

- **Retour à Bouguer et La Condamine...**
  Après avoir démontré les propriétés de l'espérance et de la variance, on va voir l'intérêt d'utiliser la moyenne.
  - Montrer que $E[aX] = aE[X]$ et $E[a] = a$ où a constante
  - Calculer la distribution $F_Y(y)$ puis la p.d.f $f_Y(y)$ où $Y = X_1 + X_2$ avec X_1 et X_2 suivant des lois pouvant être différentes.
  - En déduire que $E[X_1 + X_2] = E[X_1] + E[X_2]$
  - Dans ce qui suit, les X_i sont indépendantes et identiquement distribuées (i.i.d.) d'espérance μ et d'écart-type σ et la moyenne arithmétique est $m = \frac{1}{N}\sum x_i$. Montrer que m est non biaisé.
  - Démontrer que $\text{Var}(aX) = a^2\text{Var}(X)$ puis que $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Covar}(X_1, X_2)$
  - Montrer que l'erreur sur la moyenne est $\sigma_m = \frac{\sigma}{\sqrt{N}}$

❑ **Indice**

La linéarité est claire si les X_i suivent la même loi. Sinon, on en profite pour voir comment calculer la loi de la somme de deux variables aléatoires quelconques :

$$\begin{aligned} F(y) &= P(X_1 + X_2 \leq y) = \int\int_{x_1+x_2\leq y} f(x_1, x_2)dx_1dx_2 && (1)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{y-x_1} f(x_1, x_2)dx_2dx_1 = \int_{-\infty}^{\infty}\int_{-\infty}^{y} f(x_1, u - x_1)dudx_1 && (2)\\ & && (3)\end{aligned}$$

D'où, en différenciant, $f_{X_1+X_2}(y) = \int_{-\infty}^{\infty} f(x_1, y - x_1)dx_1$. L'espérance de la somme est alors

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x_1, y - x_1)dx_1\right]dy && (4)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1 + x_2)f(x_1, x_2)dx_1dx_2 && (5)\\ &= \int_{-\infty}^{\infty} x_1\int_{-\infty}^{\infty} f(x_1, x_2)dx_2dx_1 + \int_{-\infty}^{\infty} x_2\int_{-\infty}^{\infty} f(x_1, x_2)dx_1dx_2 && (6)\\ &= \int_{-\infty}^{\infty} x_1 f(x_1)dx_2 + \int_{-\infty}^{\infty} x_2 f(x_2)dx_2 = E[X_1] + E[X_2] && (7)\end{aligned}$$

Par linéarité, on a $E[m] = \frac{1}{N}\sum E[x_i] = \frac{1}{N}\sum \mu = \mu$.

$$\mathrm{Var}(aX) = E[(aX)^2] - E^2[aX] = E[a^2X^2] - (aE[X])^2 = a^2\mathrm{Var}(X)$$

$$\begin{aligned} \mathrm{Var}(X_1 + X_2) &= E[(X_1 + X_2)^2] - E^2[X_1 + X_2] && (8)\\ &= E[X_1^2] + E[X_2^2] + 2E[X_1X_2] - E^2[X_1] - E^2[X_2] - 2E[X_1]E[X_2] && (9)\\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2(E[X_1X_2] - E[X_1]E[X_2]) && (10)\end{aligned}$$

Le dernier terme, $\mathrm{Covar}(X_1, X_2)$ est nul quand les variables ne sont pas corrélées. Quand elles sont indépendantes, c'est *a fortiori* le cas : par un calcul analogue à ci-dessus, la loi suivie par le produit X_1X_2 est $f_{X_1X_2}(y) = \int_{-\infty}^{\infty}\frac{1}{|x_1|}f_{X_1,X_2}(x_1, \frac{y}{x_1})dx_1$ dont l'espérance est

$$\begin{aligned} E[X_1X_2] &= \int_{-\infty}^{\infty} y\int_{-\infty}^{\infty}\frac{1}{|x_1|}f_{X_1,X_2}(x_1, \frac{y}{x_1})dx_1dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f_{X_1,X_2}(x_1, x_2)dx_1dx_2 && (11)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f_{X_1}(x_1)f_{X_2}(x_2)dx_1dx_2 = E[X_1]E[X_2] && (12)\end{aligned}$$

Les variables X_i étant indépendantes, leur covariance est nulle donc

$$\mathrm{Var}(m) = \frac{1}{N^2}\sum \mathrm{Var}(x_i) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 = \frac{\sigma^2}{N}$$

La moyenne permet donc de gagner un facteur $\sqrt{N}$ en précision.

En pratique

On suppose que l'on a un n-échantillon $x_1, .., x_n$, réalisation des v.a. $X_1, .., X_n$.

- **Statistique**
  C'est une fonction $g(x_1, .., x_n)$, qui est une réalisation de la v.a. $g(X_1, .., X_n)$, comme par exemple la moyenne de l'échantillon

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  estimation de l'espérance de la population parente.

- **Distribution empirique**

$$F_n(x) = \frac{\text{(nombre de } x_i \leq x)}{n}$$

- **Moments empiriques**
  d'ordre 1 : $m = \frac{1}{n} \sum_{i=1}^{n} x_i$

Si μ n'est pas connu, il est estimé par m, et alors un estimateur non biaisé de la variance est

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2$$

et une approximation non biaisée de l'écart-type est

$$s_n = \frac{\Gamma(\frac{n-1}{2})\sqrt{n-1}}{\Gamma(\frac{n}{2})\sqrt{2}} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2} \approx \frac{n - 0.75}{n-1} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2}$$

- **Statistique d'ordre**
  On note $x_{(1)} \leq .. \leq x_{(n)}$ l'échantillon trié par ordre croissant :

- **Quantile q_α**
  q_α est un quantile $(1 - \alpha)$ de l'échantillon si

$$\frac{\text{(nombre de } x_i < q_\alpha)}{n} \leq 1 - \alpha \leq \frac{\text{(nombre de } x_i \leq q_\alpha)}{n}$$

Moyenne pondérée

- **Soient x_i indépendantes, d'espérance μ, et d'erreurs gaussiennes d'écart-type σ_i, tous différents. Soit $m = \frac{\sum_{i=1}^{n} p_i x_i}{\sum_{i=1}^{n} p_i}$ la moyenne pondérée où $p_i = \frac{1}{\sigma_i^2}$.**
  - Montrer que cette moyenne pondérée est non biaisée
  - Calculer la précision sur cette moyenne
- **Indice**

$$E[m] = \frac{\sum_{i=1}^{n} p_i E[x_i]}{\sum_{i=1}^{n} p_i} \quad (13)$$

$$= \mu \quad (14)$$

$$\mathrm{Var}(m) = \frac{\sum_{i=1}^{n} p_i^2 \mathrm{Var}(x_i)}{(\sum_{i=1}^{n} p_i)^2} \quad (15)$$

$$= \frac{\sum_{i=1}^{n} p_i}{(\sum_{i=1}^{n} p_i)^2} \quad (16)$$

$$\sigma_m = 1/\sqrt{\sum \frac{1}{\sigma_i^2}}$$

Loi binomiale

- **Définition**
  Probabilité de x succès sur n essais ayant deux résultats possibles, de probabilités respectives p et $1-p$
- **Propriétés**
  La somme de variables binomiales indépendantes de probabilité p est binomiale : si $X_i \rightsquigarrow b(x\,;n_i,p)$, alors

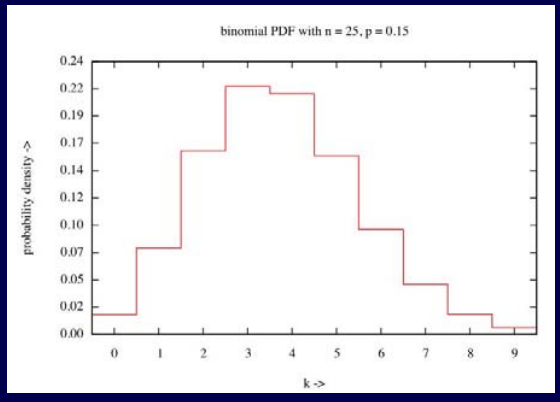
$$Y = \sum_{i=1}^{k} X_i \rightsquigarrow b(y\,;\sum_{i=1}^{k} n_i, p)$$

- **Densité**

$$b(x\,;n,p) = C_n^x p^x (1-p)^{n-x}$$

- **Moments**

$$E[X] = np \qquad \sigma(X) = \sqrt{np(1-p)}$$



Loi de Poisson

- **Définition**
  Probabilité d'apparition d'un évènement rare (en moyenne $\lambda \neq 0$) sur un grand nombre d'observations

- **Densité**

$$p(x\,;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$$

- **Moments**

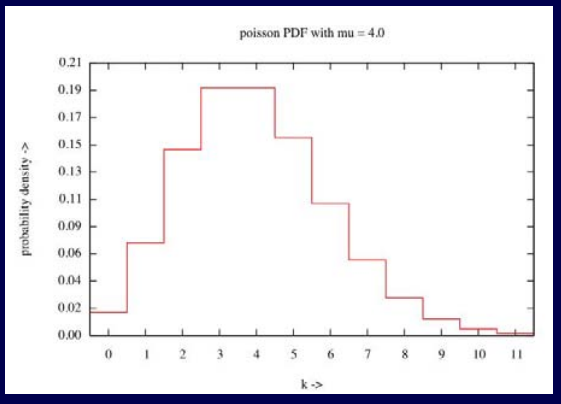
$$E[X] = \lambda \qquad \sigma(X) = \sqrt{\lambda}$$

- **Propriétés**
  La somme de variables de Poisson indépendantes est de Poisson : si $X_i \rightsquigarrow p(x\,;\lambda_i)$, alors

$$Y = \sum_{i=1}^{k} X_i \rightsquigarrow p(y\,;\sum_{i=1}^{k}\lambda_i)$$

- **Convergence**
  Vers loi normale : si $\lambda \to +\infty$, alors

$$p(x\,;\lambda) \to N(x\,;\lambda,\sqrt{\lambda}) \qquad (\lambda \gtrsim 20)$$



Loi Uniforme

- **Définition**
  Equiprobabilité de se trouver dans un intervalle $[a, b]$

- **Densité**

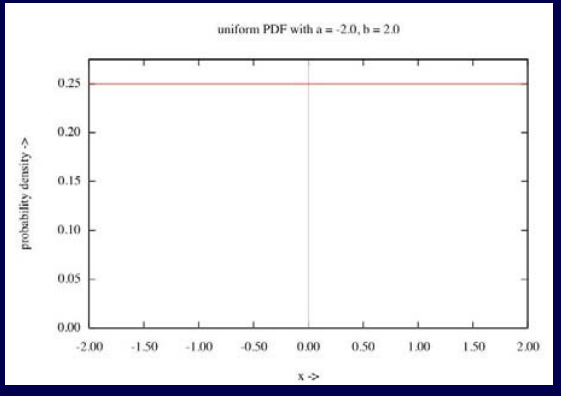
$$u(x\,;a,b) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a,b] \\ 0 & \text{sinon.} \end{cases}$$

- **Moments**

$$E[X] = \frac{a+b}{2} \qquad \sigma(X) = \frac{b-a}{2\sqrt{3}}$$

- **Propriétés**
  la loi la plus simple en l'absence d'autres informations. . .

- **Applications**
  erreur d'arrondi dans les calculs



Loi Exponentielle

- **Définition**
  Probabilité d'attendre un temps $> x$ quand $\frac{1}{\alpha}$ est le temps moyen

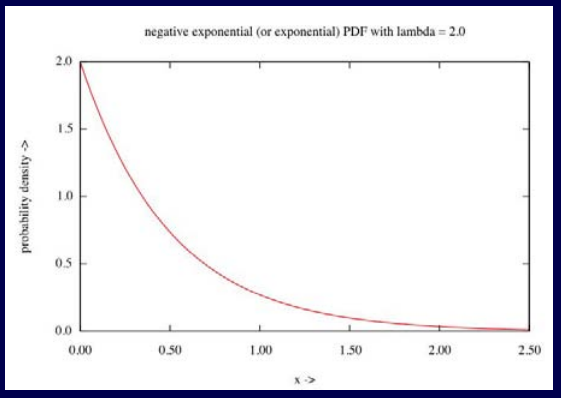
- **Densité**

$$e(x\,;\alpha) = \alpha e^{-\alpha x} \text{ pour } x \geq 0 \text{ et } \alpha > 0$$

- **Moments**

$$E[X] = \frac{1}{\alpha} \qquad \sigma(X) = \frac{1}{\alpha}$$

- **Propriétés**
  - Loi sans mémoire : $P(X > x + x' \mid X > x') = P(X > x)$
  - Si le nombre d'apparitions d'un phénomène pendant le temps t suit une loi $p(x\,;\alpha t)$ alors la distribution du temps entre deux apparitions suit une loi $e(t\,;\alpha)$
  - Si $X \rightsquigarrow u(x\,;a,b)$ alors
    $Y = \frac{-\log[(b-X)/(b-a)]}{b-a} \rightsquigarrow e(y\,;b-a)$
  - Y=Min$(e(x\,;\alpha_1),\ldots,e(x\,;\alpha_k)) \rightsquigarrow e(y\,;\sum_{i=1}^{k}\alpha_i)$



Loi Normale (Gaussienne)

- **Définition**
  Influence d'un grand nombre de facteurs aléatoires, indépendants, petits et additifs

- **Densité**

$$N(x\,;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ pour } \sigma \geq 0$$

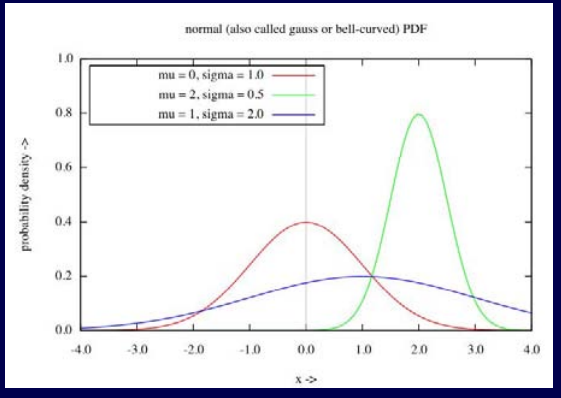
$$N(\mathbf{X}\,;\mu,\mathbf{V}) = \frac{1}{(2\pi)^{p/2}|\mathbf{V}|^{1/2}} e^{-\frac{(\mathbf{X}-\mu)^T\mathbf{V}^{-1}(\mathbf{X}-\mu)}{2}}$$

- **Moments**

$$E[X] = \mu \qquad \sigma(X) = \sigma$$

- **Convergence**
  Théorème Central Limite (TCL) : soient $X_1 \ldots X_n$ des variables indépendantes et identiquement distribuées (suivant n'importe quelle loi), de moyenne μ et de variance σ^2. Quand $n \to +\infty$, alors

$$Y = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(y\,;0,1) \qquad \begin{cases} n \gtrsim 30 & \text{loi sy} \\ n \gtrsim 60 & \text{sinon} \end{cases}$$



Loi de Cauchy

- **Définition**
  Rapport de deux variables iid $N(x\,;0,1)$

- **Densité**

$$C(x) = \frac{1}{\pi(x^2+1)}$$

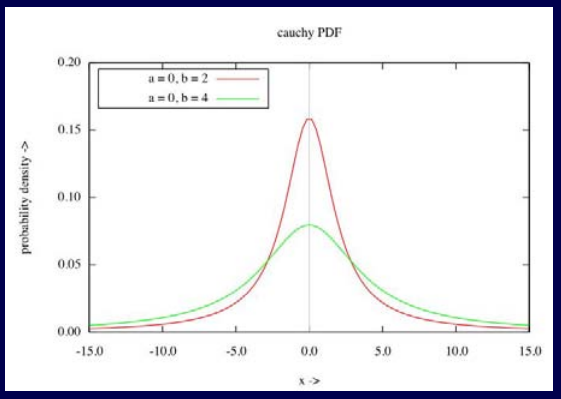
ou pour généraliser

$$C(x\,;a,b) = \frac{b}{\pi((x-a)^2+b^2)} \text{ où } b>0$$

- **Moments**
  Aucun ! Même si a et b ressemblent respectivement à des facteurs de position et d'échelle, la moyenne et l'écart-type de cette loi ne sont pas définis.

- **Propriétés**
  La somme de variables de Cauchy indépendantes est de Cauchy : si $X_i \rightsquigarrow C(x\,;a_i,b_i)$, alors

$$Y = \sum_{i=1}^{k} X_i \rightsquigarrow C(y\,;\sum_{i=1}^{k} a_i, \sum_{i=1}^{k} b_i)$$



Loi du Khi-deux (χ^2)

- **Définition**
  χ^2 à ν degrés de liberté: somme des carrés de ν variables iid $N(x\,;0,1)$

- **Densité**

$$\chi^2(x\,;\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu-2)/2} e^{-x/2} \text{ pour } x \geq 0$$

où $\Gamma(k) = \int_0^{+\infty} y^{k-1}e^{-y}dy$ (si k entier, $\Gamma(k) = (k-1)\,!$)

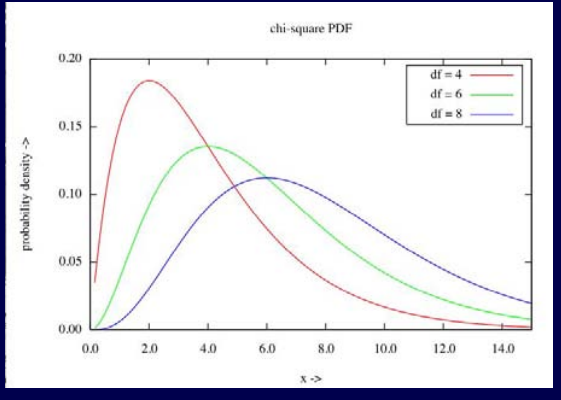
- **Moments**

$$E[X] = \nu \qquad \sigma(X) = \sqrt{2\nu}$$

- **Propriétés**
  La somme de variables χ^2 indépendantes est χ^2 : si $X_i \rightsquigarrow \chi^2(x\,;\nu_i)$, alors

$$Y = \sum_{i=1}^{k} X_i \rightsquigarrow \chi^2(y\,;\sum_{i=1}^{k} \nu_i)$$

- **Convergence**
  Si $X \rightsquigarrow \chi^2(x\,;\nu)$, quand $\nu \to +\infty$:
  - $X \to N(x\,;\nu,\sqrt{2\nu})$
  - $Y = \sqrt{2X} \to N(y\,;\sqrt{2\nu-1},1)$ $(n \gtrsim 30)$



Simulations

- **Générateur d'une loi uniforme**

- **Autres distributions :**
  Une fois que l'on sait générer une valeur uniforme, on peut générer une valeur suivant une autre loi, par l'une des méthodes ci-dessous :

  ☞ cas où l'on peut utiliser des propriétés de la loi

  - Poisson : loi $p(x\,;\alpha t)$ si $\Delta t \rightsquigarrow e(t\,;\alpha)$
  - Cauchy : rapport de deux $N(0,1)$

  ☞ cas où $F^{-1}(y)$ est facile à calculer
  On utilise la propriété que $F(x)$ suit une loi uniforme. On tire $Y \rightsquigarrow u(y\,;0,1)$, puis on calcule $x = F^{-1}(y)$ pour obtenir une variable X qui suivra la loi désirée (ex : Cauchy)

  ☞ sinon, méthode du rejet
  tirer x uniforme dans l'intervalle $[x_{\min}, x_{\max}]$, puis tirer y uniforme dans $[0, y_{\max}]$, où $y_{\max} > \max f(x)$, et garder la réalisation x si $y \leq f(x)$. Cette méthode est évidemment pénalisante en temps-calcul.

Simulations

- **Trouver une procédure (approximative) pour simuler une variable gaussienne $N(0,1)$ à partir d'un générateur uniforme**

  - Calculer l'écart-type d'une loi uniforme [0,1]
  - Utiliser le T.C.L. et faire appel la fonction `rand()`

- **Indice**
  La densité de probabilité est $f(u) = 1$. La moyenne est $< u >= \int_0^1 u du = \frac{1}{2}$. La variance est $\mathrm{Var}(u) = \int_0^1 (u - < u >)^2 du = \left[\frac{1}{3}(u - \frac{1}{2})^3\right]_0^1 = \frac{1}{12}$

```
#!/usr/bin/perl
MAIN:{
    srand(); # initialisation du generateur aleatoire
    $n=12; # nombre de tirages du generateur uniforme
    for ($j=0;$j<1000;$j++) { print gauss01(),"\n"; }
}
# tirage d'une variable approximativement gaussienne (0,1) via le T.C.L.
sub gauss01 {
    local $i; local $u=0;
    for ($i=0;$i<$n;$i++)  $u+=rand()-0.5;
    return sqrt(12/$n)*$u;
}
```

Simulations

- **Simuler une variable suivant une loi exponentielle**

  - ☞ Montrer que $F(X)$ suit toujours une loi uniforme
  - ☞ Puis calculer F^{-1} dans le cas d'une loi exponentielle

- **Indice**
  Soit $Y = F(X)$. F est croissante, donc $P(Y \leq y) = P(X \leq x)$, donc $f(y)dy = f(x)dx$ mais $\frac{dy}{dx} = f(x)$, d'où $f(y) = 1$ et Y donc suit une loi uniforme.
  La p.d.f. est $e(x\,;\alpha) = \alpha e^{-\alpha x}$ donc la fonction de répartition est

$$F(x) = \int_0^x \alpha e^{-\alpha u} du = 1 - e^{-\alpha x}$$

$$y = F(x) \Longrightarrow x = -\frac{\ln(1-y)}{\alpha}$$

  On tirera donc suivant une loi uniforme $Y \rightsquigarrow u(y\,;0,1)$, puis on calculera $x = F^{-1}(y)$ pour obtenir une variable X qui suivra la loi désirée.

Propagation des erreurs

- **Changement de variable**
  Si l'on connaît la densité de X, celle de $Y = h(X)$ est $f_Y(y) = f_X(x)|\frac{dx}{dy}|$. Pour le montrer, on utilise le fait que $P(Y \leq y) = P(X \leq x)$ si h est croissante, d'où $f_Y(y)dy = f_X(x)dx$

- **Quelle est l'erreur sur $Y = h(X)$ sachant celle sur X ?**
  Si $\mathbf{\Sigma_X} = (\sigma_{ij})$ est la matrice de variance-covariance des $\mathbf{X}$, alors la matrice de variance-covariance des $\mathbf{Y}$ est

$$\mathbf{\Sigma_Y} = \mathbf{J}\mathbf{\Sigma_X}\mathbf{J}^T$$

  où $\mathbf{J} = (\frac{\partial h(X_i)}{\partial X_j})$ est le jacobien.

- **précautions :**

  - ce n'est valable qu'au premier ordre
  - μ étant souvent inconnu, on utilise $g'(x)$, au lieu de $g'(\mu)$, rendant ce terme aléatoire, donc dégradant la précision, et pouvant introduire des biais.

Propagation (cas quasi-linaire)

- **Flux et magnitude**
  On a reçu le flux F de photons par unité de temps, on peut donc calculer la magnitude de l'objet $m = m_0 - 2.5 \log F$ et on veut estimer la précision sur cette magnitude.
  - Soit X de moyenne μ et écart-type σ_X, faire un développement de Taylor de $Y = \phi(X)$ au voisinage de la moyenne puis montrer que $\sigma_y = |\frac{d\phi(x)}{dx}|_{x=\mu}\sigma_X$
  - m_0 est supposé sans erreur, le flux mesuré F est supposé poissonnien. Calculer la précision sur la magnitude m en fonction du flux seulement.

- **Indice**
  - $\phi(X) \approx \phi(\mu) + \frac{d\phi(x)}{dx}|_{x=\mu}(X - \mu)$
  - $E[Y] \approx \phi(\mu) + \frac{d\phi(x)}{dx}|_{x=\mu}E[X - \mu] = \phi(\mu)$ donc $E[\phi(X)] \approx \phi(E[X])$ (au premier ordre seulement, l'égalité est fausse en général!)
  - $\mathrm{Var}(Y) = \left(\frac{d\phi(x)}{dx}|_{x=\mu}\right)^2 \mathrm{Var}(X)$
  - F est la seule chose que l'on connaît, et on suppose donc que $\mu = F$. On a alors $\sigma_m = \frac{2.5}{\ln 10}\frac{\sigma_F}{F}$. Le flux est poissonnien donc $\sigma_F = \sqrt{F}$ et $\sigma_m = \frac{2.5}{\ln 10}\frac{1}{\sqrt{F}}$

Estimation bayésienne

- **Théorème de Bayes**
  Pour 2 évènements A et B, la probabilité conjointe est

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

$$\text{donc } f(\theta \mid x) = \frac{f(x \mid \theta)f(\theta)}{\int_{-\infty}^{+\infty} f(x \mid \theta)f(\theta)d\theta}$$

  $f(\theta)$ est la loi *a priori*, $f(\theta \mid x)$ est la loi *a posteriori* et $f(x \mid \theta)$ est nommée la vraisemblance.

- **loi** *a priori*
  si la loi $f(\theta)$ est inconnue, on la choisit en général :
  - uniforme ($f(\theta)$ constant) pour un paramètre de position (comme la moyenne)
  - inverse ($f(\theta) \propto 1/\theta$) pour un paramètre d'échelle (comme l'écart-type)

- **Espérance** *a posteriori*

$$E[\theta \mid X] = \frac{\int_{-\infty}^{+\infty} \theta f(x \mid \theta)f(\theta)d\theta}{\int_{-\infty}^{+\infty} f(x \mid \theta)f(\theta)d\theta}$$

Biais ou non ?

- **Biais de Dyson/Eddington/Trumpler-Weaver (Lutz-Kelker)**
  Soit ϖ_0 la parallaxe mesurée d'une étoile, non biaisée, d'espérance ϖ, d'erreur gaussienne, où la densité $f(\varpi)$ n'est pas uniforme. On veut estimer la parallaxe moyenne de toutes les étoiles ayant la parallaxe observée ϖ_0
  - Calculer $f(\varpi_i|\varpi_0)$ comme loi a posteriori
  - Commencer par calculer la dérivée de la densité observe $f'(\varpi_0)$.
  - Exprimer $E[\varpi_i|\varpi_0]$ en fonction de la dispersion des mesures σ, de la densité observe $f(\varpi_0)$ et de sa dérivée.
  - Noter qu'ici on n'a pas besoin ici de loi *a priori*.
  - En déduire l'erreur commise quand on tronque une distribution observée

- **Indice**
  La distribution *a posteriori* est $f(\varpi|\varpi_0) = \frac{f(\varpi_0|\varpi)f(\varpi)}{f(\varpi_0)}$ avec la loi marginale $f(\varpi_0) = \int_{-\infty}^{+\infty} f(\varpi_0|\varpi)f(\varpi)d\varpi$. Donc $E[\varpi \mid \varpi_0] = \frac{1}{f(\varpi_0)} \int_{-\infty}^{+\infty} \varpi f(\varpi_0|\varpi)f(\varpi)d\varpi$.
  Calculons maintenant la dérivée de la distribution observée:

$$
\begin{aligned}
f'(\varpi_0) &= \int_{-\infty}^{+\infty} -\frac{(\varpi_0 - \varpi)}{\sigma^2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\varpi_0-\varpi)^2}{\sigma^2}} f(\varpi)d\varpi && (17)\\
&= \int_{-\infty}^{+\infty} -\frac{(\varpi_0 - \varpi)}{\sigma^2} f(\varpi_0|\varpi)f(\varpi)d\varpi && (18)\\
&= -\frac{\varpi_0}{\sigma^2} f(\varpi_0) + \frac{1}{\sigma^2} \int \varpi f(\varpi_0|\varpi)f(\varpi)d\varpi && (19)\\
&= -\frac{\varpi_0}{\sigma^2} f(\varpi_0) + \frac{1}{\sigma^2} E[\varpi \mid \varpi_0]f(\varpi_0) && (20)\\
&= \frac{f(\varpi_0)}{\sigma^2}(-\varpi_0 + E[\varpi \mid \varpi_0]) && (21)
\end{aligned}
$$

  D'où l'espérance *a posteriori* $E[\varpi \mid \varpi_0] = \varpi_0 + \sigma^2 \frac{f'(\varpi_0)}{f(\varpi_0)}$

  Une coupure à une valeur ϖ_0 implique que l'échantillon tronqué aura une parallaxe moyenne biaisée (étoiles plus lointaines que l'estimation qu'on en fait). L'utilisation individuelle de la "correction" ci-dessus reste néanmoins fortement discutable.

Biais de Malmquist

❑ **Troncature en magnitude**

☞ La magnitude absolue des étoiles d'un certain type a une certaine dispersion autour de la valeur moyenne pour ce type. La distribution des magnitudes apparentes est croissante, et forcément tronquée observationnellement. Mettre en évidence le biais sur la valeur moyenne de la magnitude absolue d'un échantillon à l'aide d'une simulation.

- Générer des distances d'étoiles en supposant la densité stellaire volumique constante, jusqu'à par ex. 1 kpc.
- D'autre part, pour chaque étoile, simuler une magnitude absolue gaussienne de dispersion $\sigma = 1$ mag autour de la moyenne (par ex. $M = 5$ pour un type solaire)
- Fabriquer un échantillon limité à la magnitude apparente $m = 11$
- Calculer la magnitude absolue moyenne de l'échantillon et le biais $\approx -1.38\sigma^2$

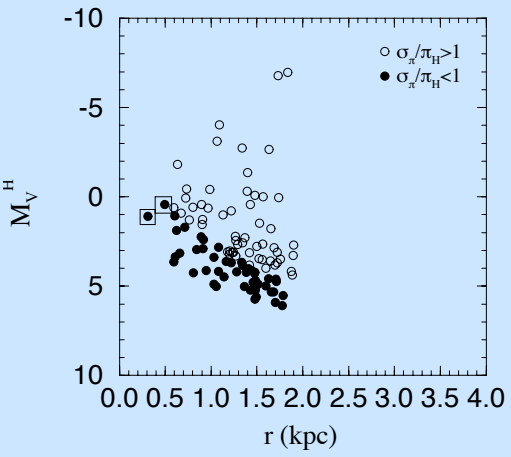
❑ **Indice**

Le nombre d'objets à la distance r est $F(r) = \left(\frac{r}{1000}\right)^3$ avec la normalisation pour avoir $F(1000) = 1$. Comme F suit une loi uniforme, en tirant u uniforme, les distances $r = 1000u^{\frac{1}{3}}$ suivent la loi attendue. La procedure en Perl peut être :

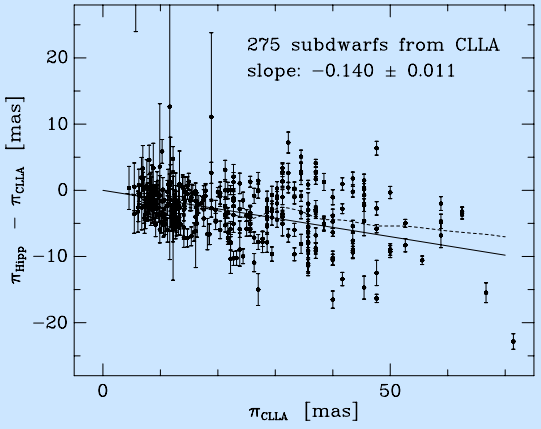
```
$M0 = 5; # magnitude absolue moyenne
  $sigmaM = 1; # dispersion des mag.  abs.
  $dLim = 1000; # distance limite echantillon
  $mLim = 11; # magnitude apparente limite pour troncature
  $moyM = 0; # compteur:  magnitude absolue moyenne de l'echantillon
  $nbStars = 0; # compteur:  nombre d'objets dans l'echantillon
  for ($j=0;$j<10000;$j++) # nombre de tirages
  {
    $d = $dLim * rand()**(1./3.)  ; # distance aleatoire
    $M = $sigmaM*gauss01() + $M0 ; # magnitude absolue
    $m = $M + 5*log($d)/log(10) - 5; # magnitude apparente
    if ($m < $mLim) {
       $moyM += $M; # incrementer la moyenne de l'echantillon
       $nbStars ++; # et le nombre d'objets retenus
    }
  }
  if ($nbStars == 0) { print "Pas d'etoiles dans l'echantillon\n" }
  else {print "biais = ",$moyM/$nbStars-$M0," magnitude, $nbStars etoiles\n"}
```

Graphes

- **Deux paramètres**
  Peut-on tracer la magnitude absolue vs la distance, celles-ci étant calculées en utilisant les parallaxes observées π_H?



- **Deux catalogues**
  On a deux catalogues de parallaxes (notées π_{CCLA} et ϖ_{Hipp}) que l'on veut comparer. Vaut-il mieux tracer π_{CCLA} vs ϖ_{Hipp} ou bien $\pi_{\mathrm{CCLA}} - \pi_{\mathrm{Hipp}}$ vs π_{CCLA} (ou vs ϖ_{Hipp}) ?



Réponse: on ne fait que mettre en évidence les corrélations (aucun effet physique).

Estimation ponctuelle

- **Détermination d'un paramètre (ou vecteur)**
  L'estimation ponctuelle consiste à associer une valeur unique obtenue de l'échantillon à un paramètre de la population.

- **Qualité des estimateurs**
  Quand on veut connaître la valeur centrale d'un échantillon, le premier réflexe est d'en calculer la moyenne arithmétique. En fait, il existe bien d'autres estimateurs.

  ☞ biais contre precision

Qualité des estimateurs

Soit $\widehat{\theta}_n$ un estimateur de θ, calculé à partir d'un n-échantillon.

- **Convergence**
  $\widehat{\theta}_n$ est un estimateur convergent si
  $$\forall \epsilon > 0 \, ; \lim_{n \to +\infty} P(|\widehat{\theta}_n - \theta| \geq \epsilon) = 0$$

- **Absence de biais**
  Le biais d'un estimateur $\widehat{\theta}_n$ est
  $$B_n(\theta) = E[\widehat{\theta}_n] - \theta$$

- **Optimalité**
  $\widehat{\theta}_n$ est un estimateur optimal s'il est à la fois convergent, non-biaisé, et de variance inférieure à celle de tout autre estimateur.

- **Robustesse (ou fiabilité)**
  $\widehat{\theta}_n$ est robuste s'il a une faible sensitivité en cas d'écart aux hypothèses initiales.

Efficacité d'un estimateur

- **Information de Fisher**
  Si les v.a. X_i sont indépendantes, la vraisemblance du n-échantillon est le produit des vraisemblances individuelles
  $$\mathcal{L}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$
  $$I_n(\theta) = E\left[\left(\frac{\partial \log \mathcal{L}}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right]$$
  est nommée l'information de Fisher contenue dans le n-échantillon

- **Inégalité de Fréchet-Darmois-Rao-Cramer**
  Si $\widehat{\theta}_n$ est un estimateur non biaisé de θ, alors sa variance est supérieure à la borne de Fréchet :
  $$\mathrm{Var}_\theta(\widehat{\theta}_n) \geq \frac{1}{I_n(\theta)}$$
  avec $I_n(\theta) = nI_1(\theta)$ si l'échantillon est i.i.d.

- **Estimateur efficace (MVB)**
  C'est un estimateur non biaisé dont la variance atteint la borne de Fréchet (donc le plus précis des estimateurs non-biaisés).

Biais

❑ **Distance d'un amas**
Si l'on veut calculer la distance moyenne d'un amas en utilisant les parallaxes observées des étoiles ϖ_{0i}, deux estimateurs sembleraient à première vue équivalents : la moyenne des distances individuelles $m_1 = < \frac{1}{\varpi_{0i}} >$ ou bien l'inverse de la moyenne des parallaxes $m_2 = \frac{1}{<\varpi_{0i}>}$. Lequel des deux choisir ?

☞ Considérer d'abord m_1 et calculer le biais $B_i = E[\frac{1}{\varpi_{0i}}] - \frac{1}{\varpi_i}$ sur la distance individuelle de l'étoile i

- Utiliser la définition de l'espérance
- Introduire la variable centrée réduite
- Approcher au deuxième ordre en $\frac{\sigma}{\varpi_i}$

☞ Comparer l'effet sur les deux estimateurs

❑ **Indice**

- Biais individuel

$$\begin{aligned} E[\frac{1}{\varpi_0}] - \frac{1}{\varpi} &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{\varpi_0} e^{-\frac{(\varpi_0-\varpi)^2}{2\sigma^2}} d\varpi_0 - \frac{1}{\varpi} \\ &= \frac{-1}{\varpi\sqrt{2\pi}} \int_{-\infty}^{+\infty} (1 - \frac{1}{1+u\frac{\sigma}{\varpi}}) e^{-\frac{u^2}{2}} du \qquad (22) \\ &\neq 0 \text{ en général, dès que } \frac{\sigma}{\varpi} \neq 0 \end{aligned}$$

- La distance calculée avec la parallaxe observée est donc biaisée, avec un biais $\approx \frac{\sigma^2}{\varpi^3} + 3\frac{\sigma^4}{\varpi^5} + \ldots$ aux premiers ordres en $\frac{\sigma}{\varpi}$. Ce biais est aggravé quand on ne conserve que les parallaxes positives.
- Sur la moyenne m_1 de ces distances, le biais est donc $< B_i >$, globalement équivalent à chaque biais individuel. L'estimateur m_2 est également biaisé, et son biais s'obtient en substituant $\frac{\sigma}{\sqrt{n}}$ à σ dans l'expression ci-dessus (car c'est la précision sur la moyenne des parallaxes). Quand la taille de l'échantillon augmente, le biais de m_2 tend ainsi vers 0, rendant cet estimateur nettement préférable à m_1. De plus, on peut montrer que la variance de m_2 est également plus petite que celle de m_1. Donc m_2 est préférable.

Méthodes d'estimation

- **Moments**
  Si un paramètre θ peut s'exprimer en fonction des k premiers moments $h(\mu_1, \ldots, \mu_k)$,
  - on calcule les moments empiriques $\widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$
  - on estime $\widehat{\theta} = h(\widehat{\mu}_1, \ldots, \widehat{\mu}_k)$
  - d'où résolution de k équations à k inconnues.

- **Maximum de vraisemblance (ML)**
  Maximiser $\mathcal{L}(x_1, \ldots x_n; \theta) = f(x_1; \theta) \times \ldots f(x_n; \theta)$ par rapport aux paramètres. On recherche les solutions de

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \text{ avec } \frac{\partial^2 \mathcal{L}}{\partial \theta^2} < 0$$

- **Moindres carrés (LS)**
  Comme son nom l'indique, il s'agit de minimiser l'écart quadratique entre un modèle et les observations censées le représenter :

$$\min_{\Theta} (\mathbf{Y} - h(\mathbf{X}\,; \Theta))^T \mathbf{V}^{-1} (\mathbf{Y} - h(\mathbf{X}\,; \Theta))$$

Ajustement

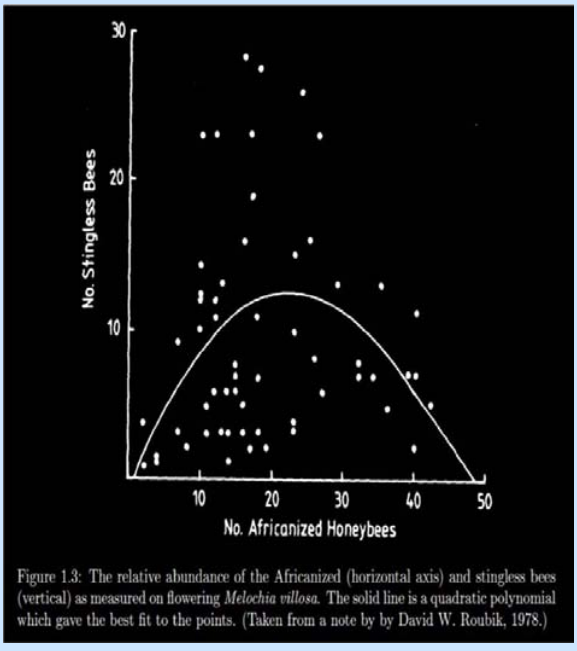


Figure 1.3: The relative abundance of the Africanized (horizontal axis) and stingless bees (vertical) as measured on flowering *Melochia villosa*. The solid line is a quadratic polynomial which gave the best fit to the points. (Taken from a note by by David W. Roubik, 1978.)

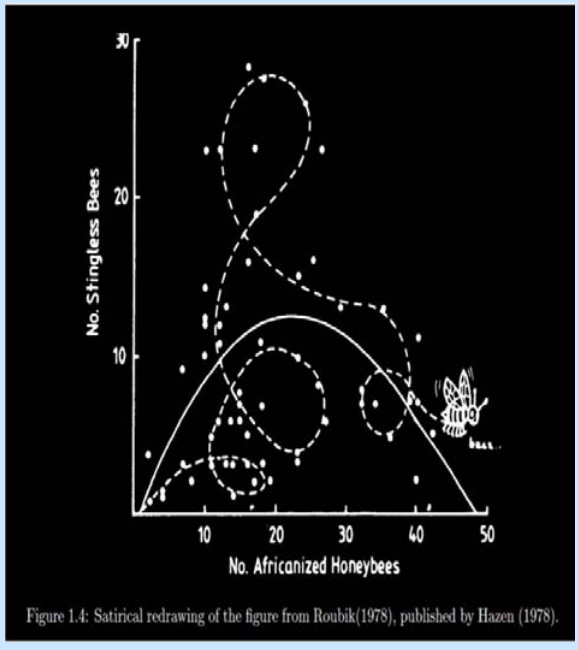


Figure 1.4: Satirical redrawing of the figure from Roubik(1978), published by Hazen (1978).

Ajuster correctement les données!

Moyenne par ML

❑ **Quel est le meilleur estimateur de la parallaxe moyenne ϖ d'un amas ?**
On suppose avoir un échantillon de parallaxes ϖ_{0i} indépendantes, dont les erreurs sont gaussiennes de précision individuelle σ_i (hétéroscédastique), et que l'amas est suffisamment lointain pour que sa profondeur soit négligeable.

- Calculer la densité de probabilité pour une observation
- Puis la log-vraisemblance de l'échantillon
- L'estimateur pour la parallaxe moyenne ϖ

❑ **Indice**

La vraisemblance individuelle est $f(\varpi_{0i} \mid \varpi) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(\varpi_{0i}-\varpi)^2}{2\sigma_i^2}}$ donc la log-vraisemblance est $\log\mathcal{L} = -\sum \frac{(\varpi_{0i}-\varpi)^2}{2\sigma_i^2}$+constante et on calcule $\frac{\partial \log\mathcal{L}}{\partial \varpi} = 0$, d'où l'on trouve, en posant $p_i = \frac{1}{\sigma_i^2}$, l'estimateur de ϖ : $\widehat{\varpi} = \frac{\sum_{i=1}^{n} p_i \varpi_{0i}}{\sum_{i=1}^{n} p_i}$. L'estimateur est donc la moyenne pondérée par l'inverse des variances individuelles. Sa précision s'obtient par l'information de Fisher, ou bien en calculant la variance, d'où $\sigma_{\widehat{\varpi}} = \frac{1}{\sqrt{\sum_{i=1}^{n} p_i}}$.

D'après les propriétés du maximum de vraisemblance, c'est le meilleur estimateur.

Localisation par ML

❑ **Estimer la position et le flux total d'un objet sur des pixels par ML ?**
On a observé un champ d'objets pour lesquel on veut faire l'astrométrie et la photométrie. Il s'agit donc d'utiliser le flux sur tous les pixels prédit par: le signal + le fond de ciel b (supposé connu) + le bruit de Poisson et d'en déduire le centroïde et le flux. On négligera le bruit de lecture R du CCD. On note:

- les indices i, j et les coordonnées x, y en unité de pixel avec (x, y) le centroïde de l'objet, et (i, j) la position (entière) d'un pixel
- $C(x, y)$ la PSF effective, normalisée et centrée en (0,0)
- $C_{ij} = C(i - x, j - y)$ sa valeur sur le pixel (i, j)
- $C'^x_{ij} = \frac{\partial C}{\partial x}(i - x, j - y)$, $C''^{xy}_{ij} = \frac{\partial^2 C}{\partial x \partial y}(i - x, j - y) = C''^{yx}_{ij}$ etc, ses dérivées
- n_{ij} le nombre d'e$^-$ reçus sur le pixel (i, j) et N le nombre total d'e$^-$ de l'objet
- b est le fond de ciel moyen en e$^-$ par pixel
- s_{ij} est l'espérance théorique d'e$^-$ reçus sur un pixel (i, j),
- $\mathbf{p} = (p_m) = (x, y, N)$ est le vecteur de paramètres à déterminer par ML
- $\mathbf{n} = (n_{ij})$ est le vecteur d'observables (le flux sur les pixels)

❑ **Indice**

Sur un pixel (i,j), l'espérance du nombre d'électrons reçus sera $E[n_{ij}] = s_{ij} = NC_{ij} + b$ et suit une loi de Poisson: $p(n_{ij}\,;s_{ij}) = \frac{{s_{ij}}^{n_{ij}}}{n_{ij}!}e^{-s_{ij}}$

La log-vraisemblance du groupe de pixels de l'objet est alors

$$\ln \mathcal{L}(\mathbf{n}|\mathbf{p}) = \sum_{(i,j)} [n_{ij} \ln s_{ij} - s_{ij} - \ln(n_{ij}!)]$$

Ses dérivées (formant le vecteur score $\mathbf{S}$) sont

$\frac{\partial \ln \mathcal{L}}{\partial x} = N \sum_{(i,j)} (1 - \frac{n_{ij}}{s_{ij}}) C'^{x}_{ij} = 0$ et $\frac{\partial \ln \mathcal{L}}{\partial (\ln N)} = -N \sum_{(i,j)} (1 - \frac{n_{ij}}{s_{ij}}) C_{ijK} = 0$

La matrice Hessienne est $\mathbf{H} = \left(\frac{\partial^2 \ln \mathcal{L}}{\partial p_m \partial p_n} \right)$ et mesure la courbure de la vraisemblance. La matrice d'information de Fisher $\mathbf{F} = E[-\mathbf{H}]$, avec n_{ij} la partie aléatoire. Ces matrices sont utiles car $\mathbf{F}^{-1}$ est la matrice de variance-covariance également utilisée pour les itérations.

$-\frac{\partial^2 \ln \mathcal{L}}{\partial x^2} = N^2 \sum_{(i,j)} \frac{n_{ij}}{s_{ij}^2} C'^{x}_{ij}$ et $-\frac{\partial^2 \ln \mathcal{L}}{\partial (\ln N)^2} = N^2 \sum_{(i,j)} \frac{n_{ij}}{s_{ij}^2} C^2_{ij}$

La solution est trouvée itérativement par Newton-Raphson. Notant $\mathbf{p}^{(I)} = \left(p_m^{(I)} \right)$ le vecteur obtenu à l'itération I, la valeur à l'itération suivante est calculée par

$$\mathbf{p}^{(I+1)} = \mathbf{p}^{(I)} + \mathbf{F}^{-1}\mathbf{S}^{(I)}$$

Estimation d'intervalles

❑ **Intervalle de confiance**

L'intervalle de confiance $[m_{\text{inf}}, m_{\text{sup}}]$ contient le paramètre recherché μ avec la probabilité γ si

$$P(m_{\text{inf}} \leq \mu \leq m_{\text{sup}}) = \gamma$$

❑ **Types d'intervalle**

Il y a bien sûr une infinité de solutions à l'équation ci-dessus. Les plus courantes sont les suivantes :

- l'intervalle minimal : tend vers le mode de $f(x)$ quand $\gamma \to 0$
- l'intervalle central symétrique : tend vers la moyenne si $\gamma \to 0$
- l'intervalle bilatéral symétrique : tend vers la médiane quand $\gamma \to 0$

Tests d'hypothèses

- **Test**
  C'est une procédure de décision à partir d'un échantillon, conduisant à choisir entre deux hypothèses, par ex :

$$\begin{array}{|ll} H_0 : \theta = 3 & \text{(hypothèse nulle) contre} \\ H_1 : \theta \neq 3 & \text{(hypothèse alternative)} \end{array}$$

- **Erreurs de :**
  - première espèce : rejet de H_0 alors qu'elle est vraie. C'est le seuil α du test. On prend souvent $\alpha = 0.05$.
  - seconde espèce : acceptation de H_0 alors qu'elle est fausse (de probabilité β ; $1 - \beta$ est alors appelé puissance du test)

Comparaison de proportions

- **Binaires**
  Leinert et al. (A&A 278, 129) trouvent 44 systèmes doubles ou multiples dans un échantillon de $n = 104$ objets de la région de formation d'étoiles Taurus-Auriga. Pour les étoiles du voisinage solaire (statistiquement plus vieilles) de même type et de même gamme de période orbitale (peut-être $\approx \frac{1}{3}$ ou $\frac{1}{2}$ de la distribution totale des périodes), Duquesnoy & Mayor (A&A 248, 455) auraient 22% comme taux de duplicité. Peut-on en conclure que les étoiles naissent préférentiellement dans des systèmes doubles ou multiples, mais que l'évolution ultérieure tend à en transformer une partie en étoiles simples ?
  - Soit Π le vrai taux de duplicité (inconnu) dans Taurus-Auriga. Utiliser la variable auxiliaire X_i valant 0 si l'étoile i est simple et 1 sinon. Quelle loi suit la somme des X_i ? Quelle moyenne, quel écart-type ?
  - Soit $p = \frac{1}{n}\sum X_i$ la proportion observée. Calculer $E[p]$, Var(p), puis utiliser le T.C.L. pour indiquer la loi suivie par la proportion p (Π est supposé proche ni de 0 ni de 1).
  - Construire un intervalle de confiance à 95% pour Π
  - Conclure par un test unilatéral à 3σ de comparaison avec les 22%.

❑ **Indice**

$\sum X_i$ suit une loi binomiale $b(x; n, \Pi)$, de moyenne $E[X] = n\Pi$ et d'écart-type $\sigma(X) = \sqrt{n\Pi(1-\Pi)}$. On a $E[p] = \Pi$ et $\sigma_p = \sqrt{\frac{\Pi(1-\Pi)}{n}}$ et asymptotiquement,

$$\frac{p-\Pi}{\sqrt{\frac{\Pi(1-\Pi)}{n}}} \rightsquigarrow N(y\,;0,1)$$

(Th. de deMoivre-Laplace). Ici on a $p = 42.3\%$ et $\sigma_p = 4.8\%$ si $p \approx \Pi$.

L'intervalle de confiance est $\left[p - Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}, p + Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right]$. Au seuil de signification $\alpha = 0.05$, le quantile $1 - \frac{\alpha}{2}$ de la gaussienne centrée réduite est $Z_{\frac{\alpha}{2}} = 1.96$. On a donc l'intervalle $[32.8\%, 51.7\%]$ pour Π.

Si l'on considère maintenant le test unilatéral H_0 : $\Pi = 22\%$ contre l'hypothèse alternative H_1 : $\Pi > 22\%$, l'hypothèse nulle est rejetée car $p = 42.3 > 22 + 3\sqrt{\frac{22(100-22)}{104}} = 34$ avec 1.3 pour mille de risque d'erreur. Donc il y a effectivement nettement plus de systèmes doubles/multiples dans cette région de formation que dans les étoiles de champ du voisinage solaire. En admettant des processus et des taux de formation universels, en tenant compte du nombre de binaires non détectées (car de séparation/période non accessible) dans cette étude, et en supposant que tous les autres biais de sélection ont effectivement été pris en compte, la conclusion ci-dessus semble logique.

❑ **Sondages**

- Les sondages utilisent un général des échantillons inférieurs à 1000 personnes. Quelle précision a t'on si deux candidats A et B sont à égalité ? Que peut-on dire si le sondage donne 51% pour A et 49% pour B ?
- Quelle taille d'échantillon faudrait-il pour avoir moins de 5% de risque de se tromper en affirmant que A va gagner ?

Type de tests

❑ **Types de tests d'hypothèses**

- tests paramétriques :
  on connaît la loi en présence et on teste un ou plusieurs de ses paramètres. Ex : pour une loi normale, on teste si $\sigma = 1$.
- tests non paramétriques :
  on ne fait pas de supposition sur la loi en présence. Ex : tester si deux échantillons sont indépendants.
- tests d'adéquation :
  on teste le type de la loi en présence. Ex : mon échantillon suit-il une loi Gaussienne ?

Test de détection

❑ **Test sur une variable bidimensionnelle**
Pour la réduction astrométrique des données d'Hipparcos, des étoiles pouvaient avoir un mouvement non-linéaire et un terme d'accélération devait alors être pris en compte. Le problème était de savoir pour quelles étoiles cette accélération $\mathbf{G} = (g_{\alpha*}, g_\delta)$ était significative "à 3σ". On suppose les erreurs sur $\mathbf{G}$ gaussiennes de matrice de variance-covariance

$$\mathbf{V} = \begin{pmatrix} \sigma^2_{g_{\alpha*}} & \rho\sigma_{g_{\alpha*}}\sigma_{g_\delta} \\ \rho\sigma_{g_{\alpha*}}\sigma_{g_\delta} & \sigma^2_{g_\delta} \end{pmatrix}$$

On utilisera le fait que si $\mathbf{X} \rightsquigarrow N(\mu, \mathbf{V})$ de dimension n, $(\mathbf{X}-\mu)^T\mathbf{V}^{-1}(\mathbf{X}-\mu)$ suit une loi du χ^2 à n degrés de liberté.

❑ **Indice**
En l'absence d'accélération réelle (hypothèse nulle H_0 : $\mathbf{G} = 0$), l'accélération observée suit une loi gaussienne de moyenne (0,0) et de variance $\mathbf{V}$. Donc on teste si la statistique $F^2 = \mathbf{G}^T\mathbf{V}^{-1}\mathbf{G}$ suit une loi du χ^2 à 2 degrés de liberté.

Le seuil qui a été choisi pour le test est $\alpha = 0.0027$, qui correspondrait à un test à 3σ pour une gaussienne. Dans une table du $\chi^2(2)$, ceci correspond à la valeur 11.83 ($= 3.44^2$). Pour chaque étoile, on a ainsi calculé l'accélération, puis la statistique F^2, et on a considéré que l'étoile avait une accélération significative quand $F > 3.44$; dans ce cas l'hypothèse alternative était donc adoptée avec un risque d'erreur inférieur à 0.27%.

Bibliographie sommaire

- **Les fondamentaux**
  - *Bruit et signaux*, D. Pelat, Poly du cours de master
  - *Kendall's Advanced Theory of Statistics*, Stuart et al., 3 volumes
- **Aspects numériques**
  - *Numerical Recipes*, Press *et al.*, ed. Cambridge University Press (en C, ISBN 0-521-35465-X)
- **Côtés mathématiques**
  - *Méthodes statistiques*, Tassi, ed. Economica, ISBN 2717816232
  - *Modern Mathematical Statistics*, Dudewicz & Mishra, ed. Wiley & sons, ISBN 0-471-60716-9
  - *L'analyse statistique bayésienne*, C. Robert, ed. Economica, ISBN 2-7178-2199-6 (pour les bayésiens purs et durs)

- **Applications en astronomie**
  - *Errors, Bias and Uncertainties in Astronomy*, Jaschek & Murtagh, Cambridge University Press, ISBN 0-521-39300-0
  - *Statistical Challenges in Modern Astronomy*, Feigelson & Babu, ed. Springer Verlag, Vol I : ISBN 0-387-97911-5, Vol II : ISBN 0-387-98203-0
  - *On-line statistical software for astronomy & related fields,* `http://www.astro.psu.edu`
  - *Statistical Consulting Center for Astronomy,* `http://www.stat.psu.edu/scca/homepage.html`
- **Pour les physiciens**
  - *Statistics in theory and in practice*, Lupton, ed. Princeton University Press,
  - *Statistics for physicists*, B.R. Martin, ed. Academic Press, ISBN 0-12-474750-7
- **Et pour les autres**
  . . . qui veulent des formules rapides :
  - *Guide de Statistique appliquée*, Manoukian, ed. Hermann, ISBN 2705660224