V - Distributions d'échantillons

Les distributions d'echantillons :

- **normale ou gaussienne ,**
- **chi-carré, χ^2,**
- **Student ,**
- **Fisher**

s'appliquent excatement à des variables aléatoires construites à partir de statistiques

$$(\bar{x}, s^2, S^2)$$

d'échantillons de variables aléatoires indépendantes

$$\underline{x} = \left(x_1, x_2, \ldots, x_n \right)$$ **distribuées**

suivant des distributions gaussiennes

$$N(\mu_i, \sigma_i^2).$$

Elles s'appliquent asymptotiquement à la limite des grands échantillons aux autres distributions de variables aléatoires au travers du Théorème Central Limite

Distribution normale - Théorème Central Limite

Soit $\underline{x} = (x_1, x_2, \ldots, x_n)$ des variables aléatoires indépendantes issues de fdp

de moyennes $(\mu_1, \mu_2, \ldots, \mu_n)$

et de variances $(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$

Soit $X = \sum_{i=1}^{n} x_i \qquad \Rightarrow \qquad \mu_X = \sum_{i=1}^{n} \mu_i$ **et** $\sigma_X^2 = \sum_{i=1}^{n} \sigma_i^2$

Théorème Central Limite:

Si $n \to \infty$, $X = \sum_{i=1}^{n} x_i$ **est distribué suivant une normale** $N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$

La distribution d'une mesure complexe résultant d'un grand nombre de mesures élémentaires est distribuée normalement autour de la vraie valeur avec une variance égale à la somme des variances

Si $\underline{x}$ issus de la même fdp de moyenne μ et de variance σ^2

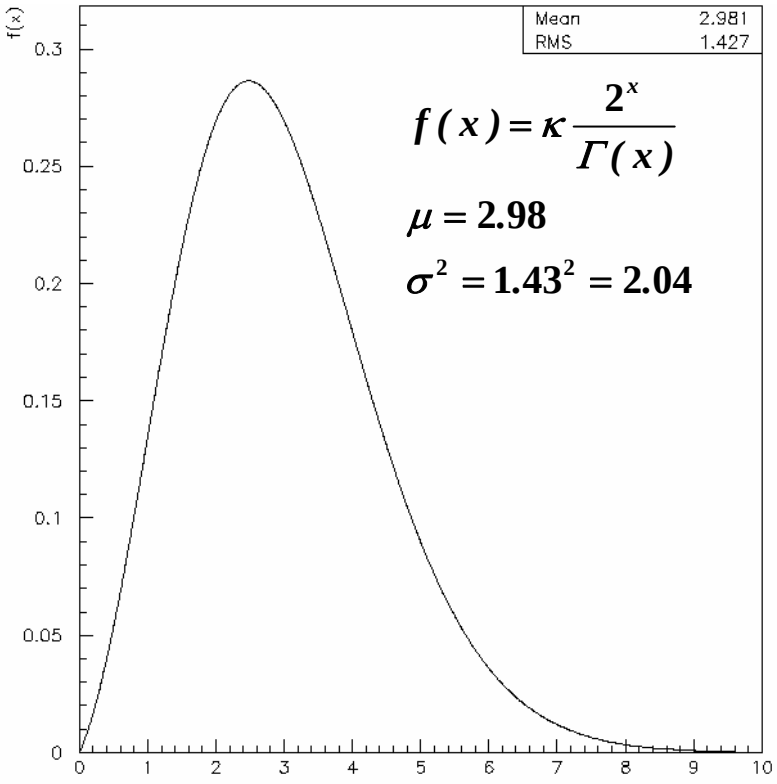
$$\text{Si } n \to \infty, \quad \bar{x} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ est distribué suivant une normale } N\left(\mu, \frac{\sigma^2}{n}\right)$$

La moyenne d'un grand échantillon de taille *n* est distribuée normalement autour de la vraie valeur avec une variance *n* fois plus petite que la variance de la population.

Ou encore

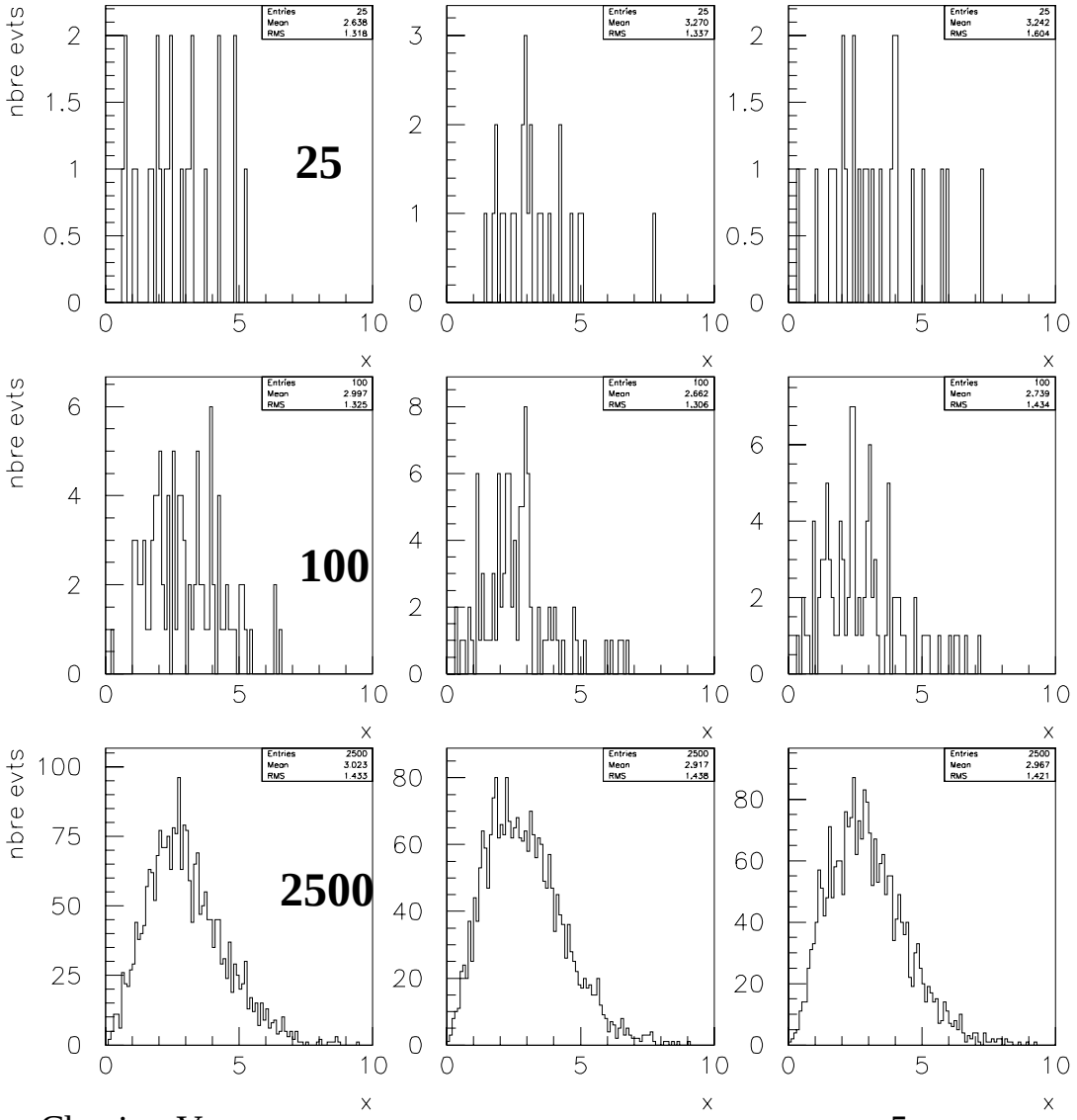
$$\text{Si } n \to \infty, \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \text{ est distribué suivant une } N(0,1)$$

Théorème Central Limite: exemple



Extraction (par simulation) de
10000 échantillons de 25 événements
10000 échantillons de 100 événements
10000 échantillons de 2500 événements

3 exemples d'échantillons des 3 tailles



Distributions des moyennes des 10000 échantillons des 3 tailles

$n = 25$

$$\frac{\sigma^2}{n} = \frac{2.04}{25} = 0.08$$

$$\sigma_x^2 = 0.285^2 = 0.08$$

$n = 100$

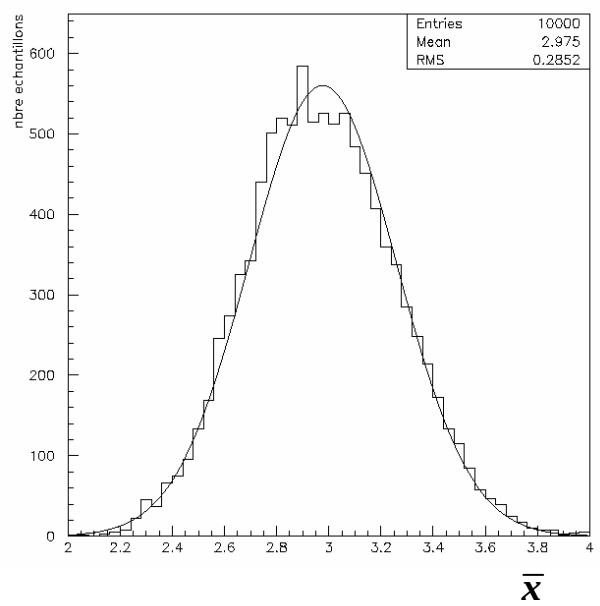
$$\frac{\sigma^2}{n} = \frac{2.04}{100} = 0.020$$

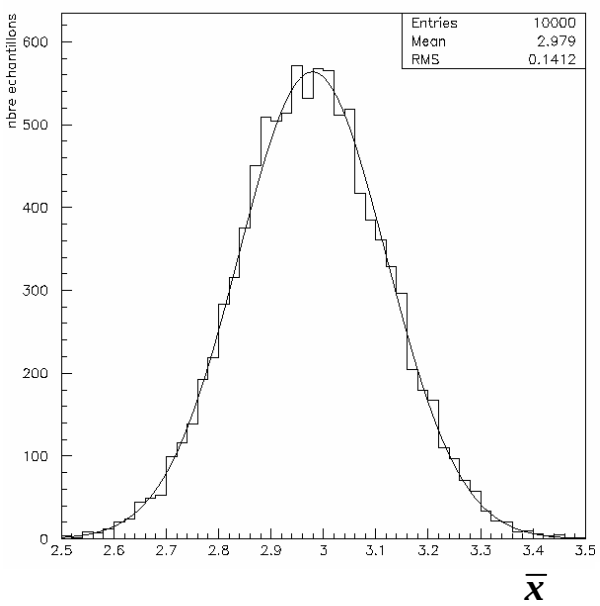
$$\sigma_x^2 = 0.141^2 = 0.020$$

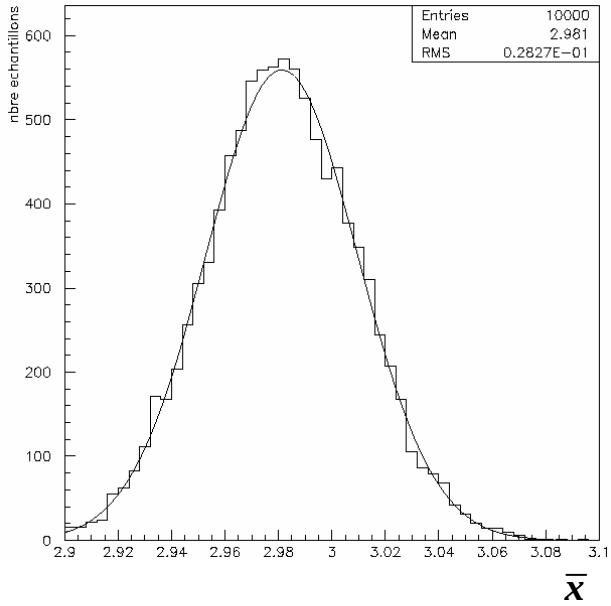
$n = 2500$

$$\frac{\sigma^2}{n} = \frac{2.04}{2500} = 0.0008$$

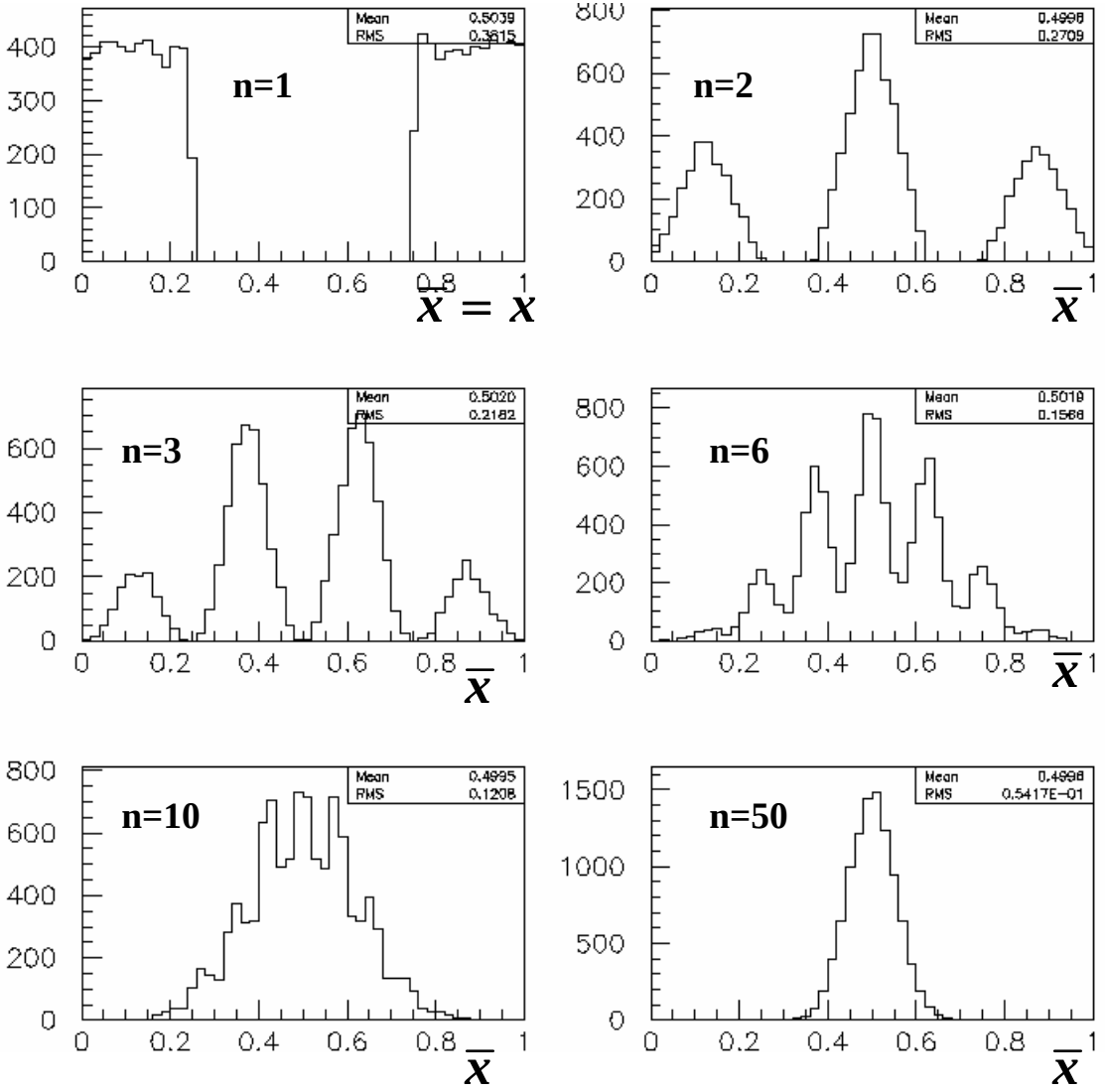
$$\sigma_x^2 = 0.028^2 = 0.0008$$





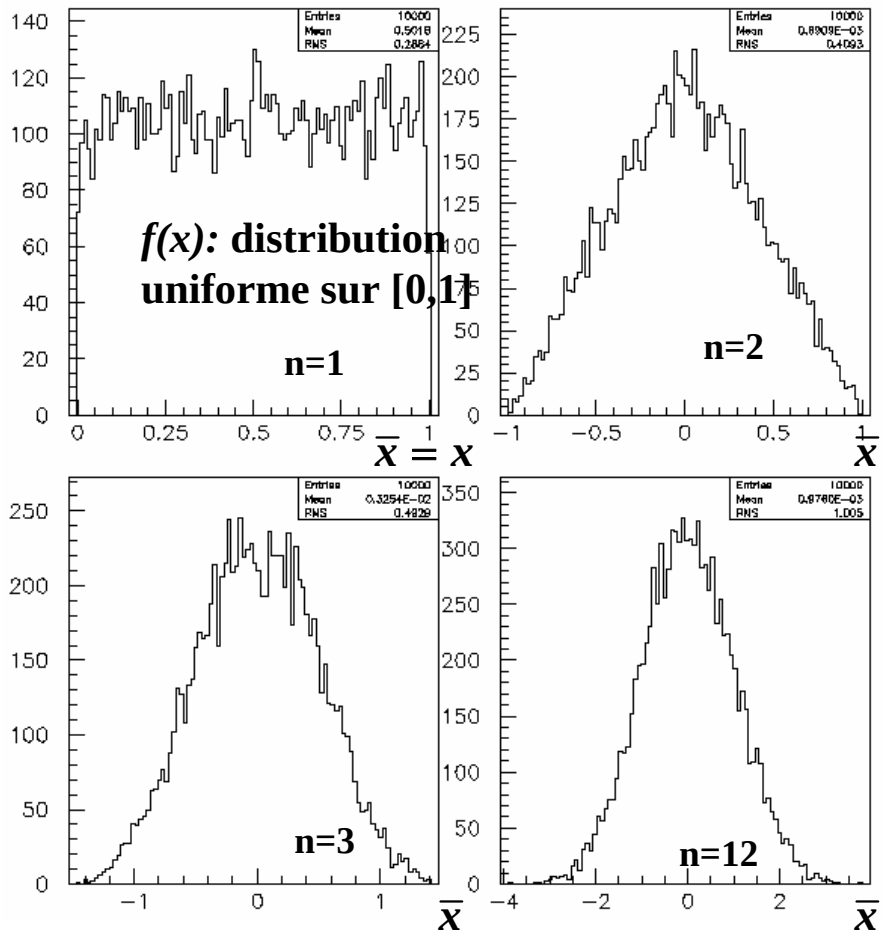


Théorème Central Limite: moyennes des distributions de 10000 échantillons de 1, 2, 3, 6, 10, 50 événements



f(x): **distribution uniforme sur [0,0.25] et [0.75,1]**

Théorème Central Limite: moyennes des distributions de 10000 échantillons de 1, 2, 3, 12 événements



Construction d'un générateur de nombres aléatoires distribués suivant une *N(0,1)* à partir d'un générateur de nombres distribués suivant une fdp uniforme sur [0,1]

$N(0,1)$ **Soient *n* nombres aléatoires** $\xi_1, \xi_2, ...\xi_n$

tirés suivant une fdp uniforme sur [0,1]

$$f(x) = 1$$

$$\mu = 1/2$$

$$\sigma^2 = \int_0^1 \left(x - 1/2\right)^2 dx = 1/12$$

soit $x = \sum_{i=1}^{n} \xi_i$

lim$_{n\to\infty}$ **:** x **est distribué suivant une** $N\left(\frac{n}{2}, \frac{n}{12}\right)$

$$z = \frac{x - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$$ **est distribué suivant** $N(0,1)$

si $n = 12 : z = \sum_{i=1}^{12} \xi_i - 6$ **est distibué** $\approx$ **suivant** $N(0,1)$

sur l'intervalle $\left[-6 \le z \le 6\right]$

Erreur Standard et Niveau de Confiance

La somme de variables aléatoires indépendantes est approximativement distribuée suivant un distribution normale:
Erreur statistique sur des mesures complexes résultant de mesures élémentaires indépendantes est approximativement normale.

Exemple simple:

L'erreur sur la mesure de la longueur d'une feuille de papier mesurée à l'aide d'une latte de 30 *cm* graduée au mm est typiquement de 0.5 *mm*. Un grand nombre de mesures répétées par des observateurs différents prendront deux, voire au plus trois valeurs, sauf faute de lecture.

Les mesures répétées de la longueur d'une pièce de 6 m par reports successifs de la latte auront une distribution approximativement gaussienne centrée sur la vraie valeur, avec un écart type de l'ordre $\sigma_p \approx \sqrt{20}\ \sigma$ où σ est de l'ordre de 0.5 à 1 mm suivant le soin apporté à la mesure : $\sigma_p \approx 2.2 - 4.5mm$.

Notation de l'erreur de mesure

Si l'erreur de mesure est gaussienne d'écart type σ, la probabilié est de $\alpha = 68.3\%$ pour que la valeur mesurée x soit dans l'intervalle $[x_0 - \sigma, x_0 + \sigma]$ si $x_0$est la vraie valeur, de 99% pour que x soit mesuré dans $[x_0 - 2.576\,\sigma, x_0 + 2.576\,\sigma]$,...

La notation $x \pm \Delta x$ du résultat d'une mesure sous entend une probabilité de 68.3% même si l'erreur de mesure n'est pas gaussienne. Si l'erreur est gaussienne, $\sigma = \Delta x$.

Si on désire donner un intervalle de valeur correspondant à une probabilité α différente de 68.3%, dans le cas d'une erreur non gaussienne, on écrira

$x \pm \Delta x$ ou $x^{+\Delta x_+}_{-\Delta x_-}$ au niveau de confiance de α% (α% C.L.)

n	$P(\mu - n\,\sigma < x < \mu + n\,\sigma)$
1	0.683
1.645	0.900
1.960	0.950
2	0.955
2.576	0.990
3	0.997
3.29	0.999

Propagation des erreurs de mesure

Soit une variable y accessible à la mesure au travers des variables directement mesurables $\underline{x} = (x_1, x_2, \ldots, x_n)$ par la relation $y = y(\underline{x})$,

Soit y_0 et $\underline{x}_0 = (x_{1,0}, x_{2,0}, \ldots, x_{n,0})$ les vraies valeurs,

Soient $\hat{\underline{x}}$ les valeurs des mesures,

Soit σ_i l'erreur de mesure gaussienne sur x_i

Développement en série de Taylor au premier ordre

$$y(\underline{x}) = y(\underline{x}_0) + \sum_{i=1}^{n} (x_i - x_{i,0}) \left.\frac{\partial y}{\partial x_i}\right|_{\underline{x}=\underline{x}_0} + \ldots = \sum_{i=1}^{n} x_i \left.\frac{\partial y}{\partial x_i}\right|_{\underline{x}=\underline{x}_0} + \text{Cste} \quad \Rightarrow \quad \boxed{\sigma_y^2 = \sum_{i=1}^{n} \left(\left.\frac{\partial y}{\partial x_i}\right|_{\underline{x}=\underline{x}_0} \right)^2 \sigma_i^2}$$

$\underline{x}_0$ inconnu est approximé par sa mesure $\hat{\underline{x}}$

$$\boxed{\sigma_y^2 = \sum_{i=1}^{n} \left(\left.\frac{\partial y}{\partial x_i}\right|_{\underline{x}=\hat{\underline{x}}} \right)^2 \sigma_i^2}$$

Approximation valable si $\dfrac{\partial^k y}{\partial x_i^k}$, $k > 1$ négligeables

Exemple: $tg\,\theta = y / x$

$$\sigma_{tg\,\theta} = \frac{1}{y} \sqrt{\sigma_y^2 + tg^2\theta\, \sigma_x^2}$$

$$\sigma_\theta = \frac{1}{1 + tg^2\theta}\, \sigma_{tg\,\theta}$$

$$\sigma_\theta = \frac{1}{1 + tg^2\theta} \frac{1}{y} \sqrt{\sigma_y^2 + tg^2\theta\, \sigma_x^2}$$

Cas général de propagation des erreurs de mesures entres variables non indépendantes vers une seule variable

$$\sigma_y^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \left.\frac{\partial y}{\partial x_i}\right|_{\underline{x}=\underline{\hat{x}}} \left.\frac{\partial y}{\partial x_j}\right|_{\underline{x}=\underline{\hat{x}}} \sigma_{x_{ij}}$$

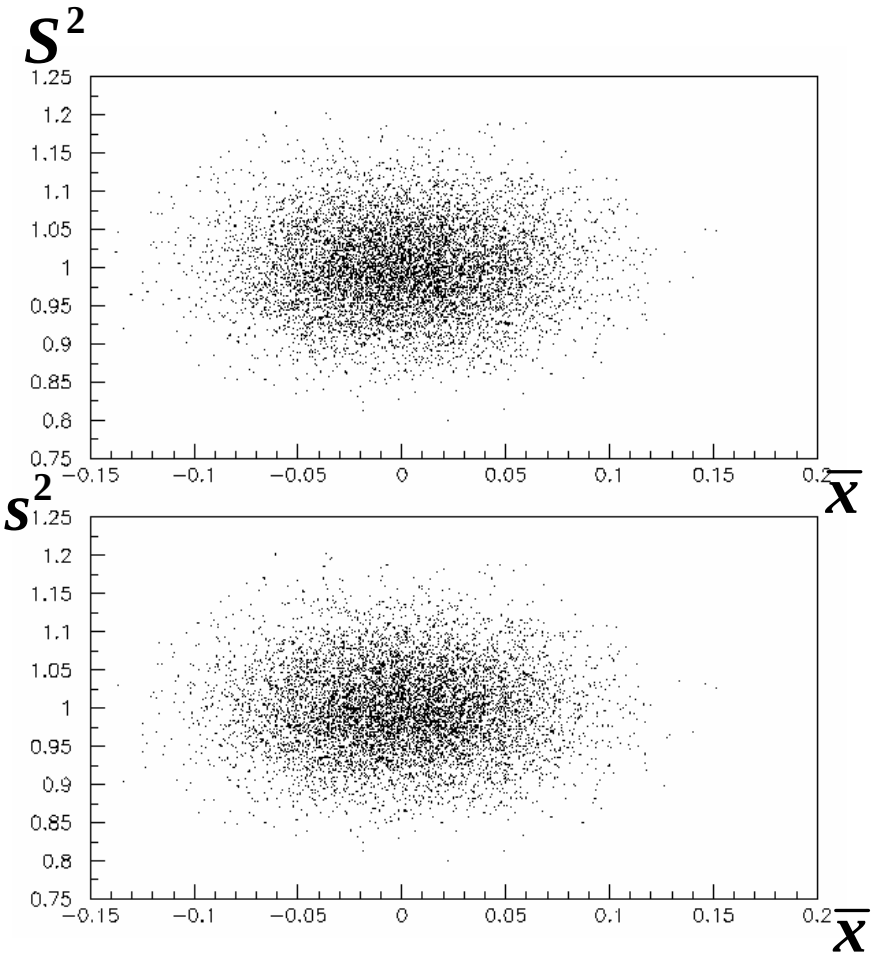
rappel : $\sigma_{ii} = \sigma_i^2$

Cas général de propagation des erreurs de mesures de variables non indépendantes vers plusieurs variables

Soit m **variables** $\underline{y} = (y_1, y_2, \ldots, y_m)$ **accessibles à la mesure au travers des** n **variables directement mesurables** $\underline{x} = (x_1, x_2, \ldots, x_n)$ **par les** m **relations** $\underline{y} = \underline{y}(\underline{x})$**,**

$$\sigma_{y_{kl}} = \sum_{i=1}^{n}\sum_{j=1}^{n} \left.\frac{\partial y_k}{\partial x_i}\right|_{\underline{x}=\underline{\hat{x}}} \left.\frac{\partial y_l}{\partial x_j}\right|_{\underline{x}=\underline{\hat{x}}} \sigma_{x_{ij}} \qquad k, l = 1, m$$

Les statistiques $\bar{x}$ et S^2, et $\bar{x}$ et s^2 d'échantillons extraits de distributions normales sont des variables aléatoires indépendantes.



Principe de la démonstration:

On montre que $\phi\left(\bar{x}, S^2\right) = \phi\left(\bar{x}\right)\phi\left(S^2\right)$

$$\phi\left(\bar{x}, s^2\right) = \phi\left(\bar{x}\right)\phi\left(s^2\right)$$

Démonstration par l'exemple:

10 000 échantillons de 625 valeurs tirées suivant une $N\left(0.,1.\right)$

Pas de corrélation entre les 10 000 paires de valeurs $\left(\bar{x}, S^2\right)$ **ou** $\left(\bar{x}, s^2\right)$

Distribution χ^2

Soient *n* variables indépendantes $\underline{x} = (x_1, ..., x_n)$ distribuées suivant des $N(\mu_i, \sigma_i^2)$

et $y_i = \dfrac{x_i - \mu_i}{\sigma_i}$ les variables standardisées correspondantes distribuées suivant des $N(0,1)$

$$\chi_n^2 = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = \sum_{i=1}^{n} y_i^2$$ est distribuée suivant une f.d.p. χ_n^2 à n degrés de libertés

χ_n^2 mesure le carré de la distance entre $\underline{x}$ et son espérance mathématique $\underline{\mu}$

dans l'espace à *n* dimensions en utilisant $\underline{\sigma}$ comme unité de mesure.

Fonction de densité de probabilité:

$$f(\chi_n^2) = \frac{1}{2^{n/2}\,\Gamma(n/2)} (\chi_n^2)^{n/2-1} e^{-\chi_n^2/2}$$

$\mu = n$ $\qquad \chi_n^2 \geq 0$

$\sigma^2 = 2n$

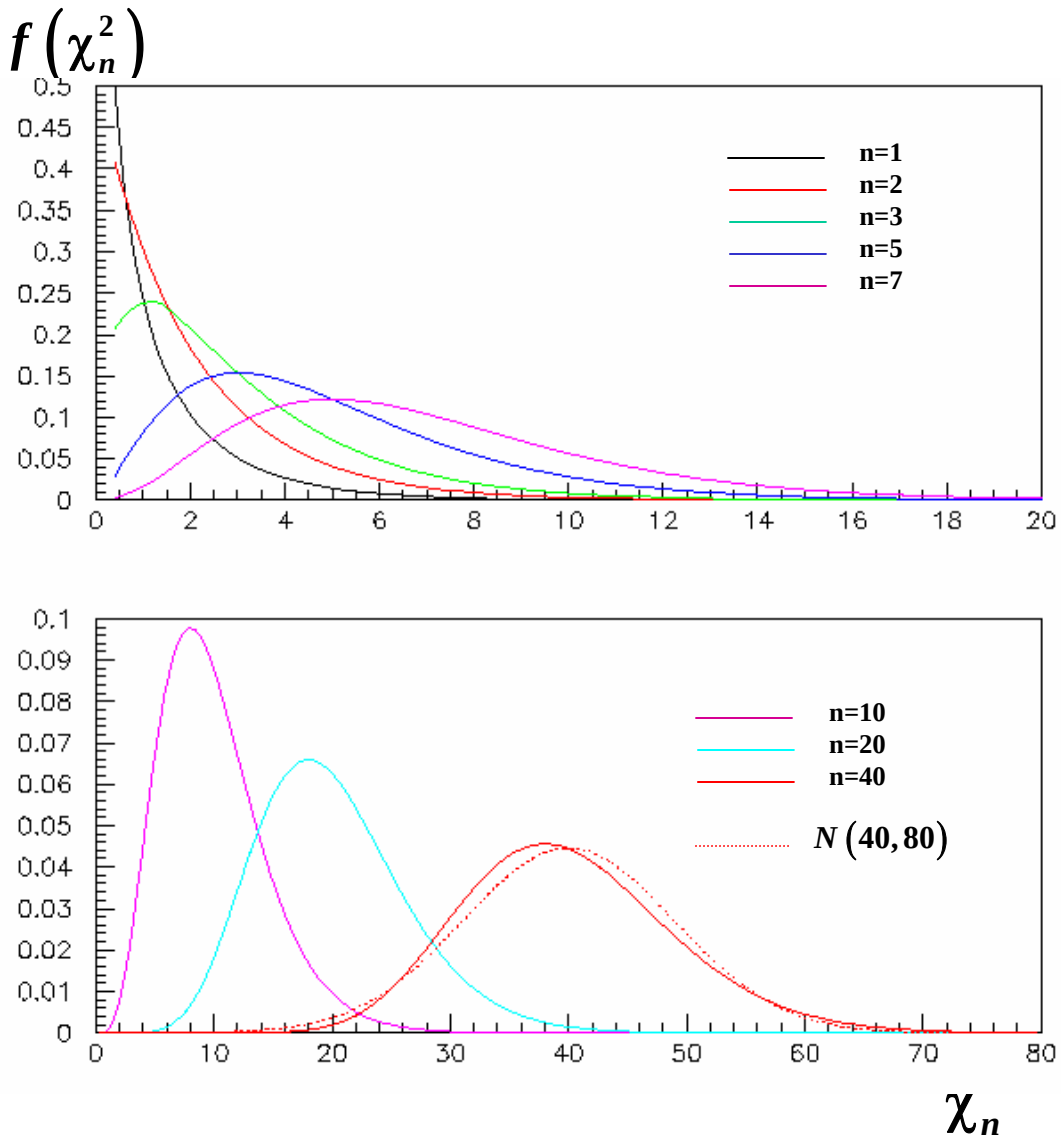
$\chi_n^2 \equiv$ fdp $gamma(\alpha = n/2, \beta = 2)$

Démonstration pour n=2

changements de variables: $y_i = \dfrac{x_i - \mu_i}{\sigma_i}$ $\quad \begin{cases} y_1 = \chi \cos\phi \\ y_2 = \chi \sin\phi \end{cases}$

$$f(\chi,\phi) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_1^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_2^2} \begin{vmatrix} \cos\phi & -\chi\sin\phi \\ \chi\cos\phi & \sin\phi \end{vmatrix}$$

$$f(\chi) = \int_0^{2\pi} f(\chi,\phi)\,d\phi \quad \Rightarrow \quad f(\chi^2) = \frac{1}{2\chi} f(\chi)$$



Cas limite: $n \to \infty$

$lim_{n\to\infty}\ f\left(\chi_n^2\right) = N\left(n, 2n\right)$

Exemples de distributions en χ^2

Conditions : les mesures sont indépendantes et non biaisées,
les erreurs sont correctement estimées,
le modèle est correct.

- E expériences mesurent P paramètres d'un modèle théorique.

Soit $x_{pe} \pm \sigma_{pe}$ la mesure obtenue pour le paramètre p par l'expérience e

Soit μ_p la prédiction du modèle.

Les E variables $X_e^2 = \sum_{p=1}^{P} \frac{\left(x_{pe} - \mu_p\right)^2}{\sigma_{pe}^2}$, $e = 1, E$ caractérisant les fluctuations statistiques des mesures des P paramètres autour de leurs prédictions, obtenues par chacune des expériences, seront distribuées suivant une χ_P^2

Les P variables $X_p^2 = \sum_{j=1}^{E} \frac{\left(x_{pe} - \mu_p\right)^2}{\sigma_{pe}^2}$, $p = 1, P$ caractérisant les fluctuations statistiques de l'ensemble des E mesures d'un paramètre autour de sa prédiction, seront distribuées suivant une χ_E^2

Si aucun modèle ne prédit μ_p :

$\bar{x}_p = \frac{1}{E}\sum_{e=1}^{E} x_{pe}$ **sont les moyennes des** E **mesures des paramètres** $p = 1, P$

Les P **variables** $X_p^2 = \sum_{j=1}^{E} \frac{\left(x_{pe} - \bar{x}_p\right)^2}{\sigma_{pe}^2}$, $p = 1, P$ **caractérisant les fluctuations statistiques de l'ensemble des** E **mesures d'un paramètre autour de leur moyenne, seront distribuées suivant une** χ^2_{E-1}.

Le nombre de degrés de liberté est $E - 1$**. Un degré est perdu parcequ'une partie de l'information est utilisée dans l'estimation de** μ_p **par** $\bar{x}_p$**. La démonstration exacte est similaire à celle de la définition de l'estimateur non biaisé** s^2 **de** σ^2**.**

Statistiques distribuées suivant une χ^2

- **Soient** $(x_1,\ldots,x_n)$ **indépendantes et extraites d'une** $N(\mu,\sigma^2)$

$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$ **est distribuée comme une** χ_n^2 **et** $S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$

$\Rightarrow$ $n\frac{S^2}{\sigma^2}$ **est distribuée comme une** χ_n^2

$\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$ **est distribuée comme une** χ_{n-1}^2 **et** $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

$\Rightarrow$ $(n-1)\frac{s^2}{\sigma^2}$ **est distribuée comme une** χ_{n-1}^2

Distribution t de Student

- Soient - x distribuée suivant un $N(0,1)$

  - u distribuée suivant une χ^2_n

  - x et u indépendantes

$$t_n = \frac{x}{\sqrt{u/n}}$$ est distribuée suivant une Student à n degrés de liberté

Fonction de densité de probabilité:

$$f(t_n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1+t_n^2\right)^{\frac{n+1}{2}}}$$

$$\mu = 0$$

$$\sigma^2 = \frac{n}{n-2} \text{ si } n > 2$$

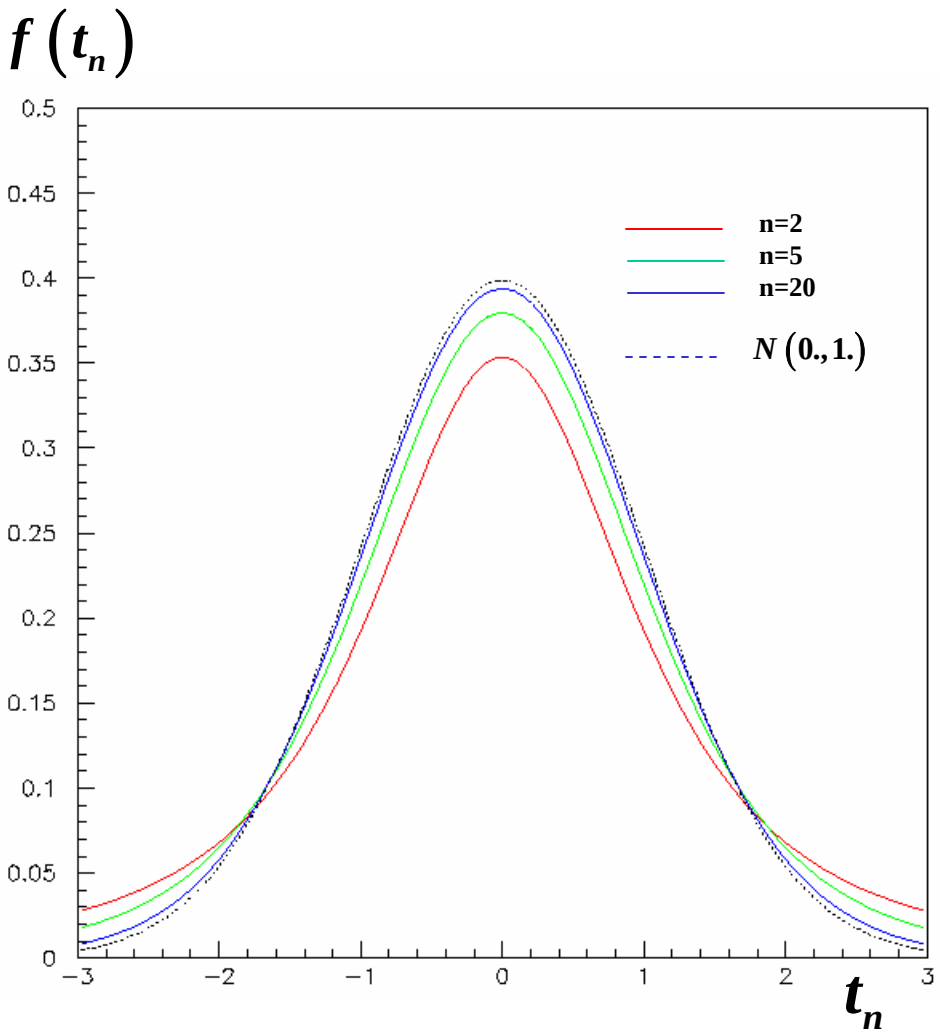
Démonstration:

changement de variables: $\begin{cases} t = \dfrac{x}{\sqrt{u/n}} \\ z = u \end{cases}$

$$f(t,z) = f(x,u)\,|J| = f(x)\,f(u)\,|J| =$$

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2^{n/2}\Gamma(n/2)} (u)^{n/2-1} e^{-u/2} |J|$$

$$f(t) = \int_0^\infty f(t,z)\,dz$$



Cas limite: $n \to \infty$

$$\lim_{n\to\infty} f(t_n) = N(0,1)$$

Statistiques distribuées suivant une Student

- **Soient $(x_1,\ldots,x_n)$ indépendantes et extraites d'une $N(\mu,\sigma^2)$**

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2,\quad s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ **distribué suivant une** $N(0,1)$

$n\dfrac{S^2}{\sigma^2}$ **est distribuée comme une** χ^2_n

$(n-1)\dfrac{s^2}{\sigma^2}$ **est distribuée comme une** χ^2_{n-1}

$$\frac{\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{n\dfrac{S^2}{\sigma^2}}{n}}}=\frac{\bar{x}-\mu}{S/\sqrt{n}}$$

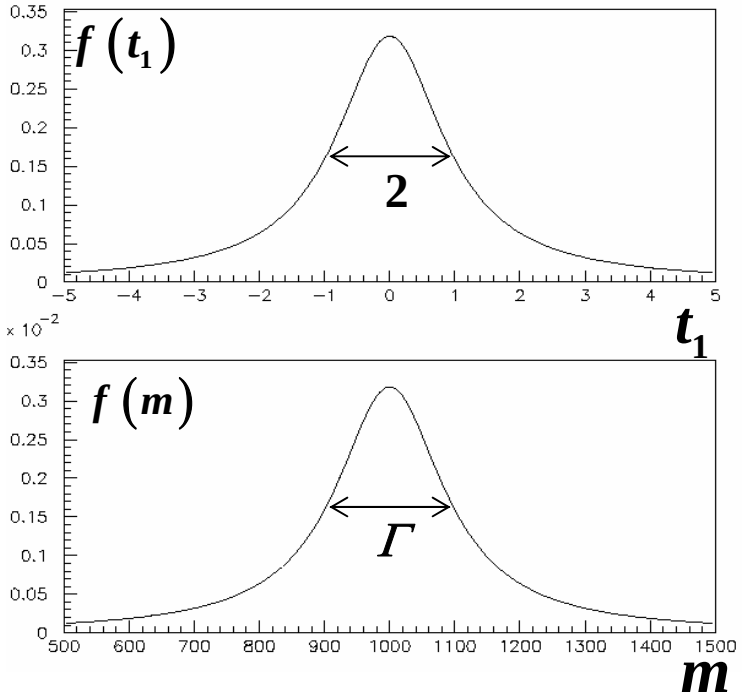
est distribué suivant une t_n de Student à n degrés de liberté

Même si la variance σ^2 de la distribution n'est pas connue et la taille n de l'échantillon est petite, la distribution de la moyenne $\bar{x}$ d'échantillons autour de la moyenne $\mu=E[\bar{x}]$ de la population est connue exactement.

$$\frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\frac{s^2}{\sigma^2}}{(n-1)}}} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$ **est distribué suivant une t_{n-1} Student à $(n-1)$ degrés de liberté**

$$lim_{n\to\infty}\ s^2, S^2 = \sigma^2 \quad \Leftrightarrow \quad lim_{n\to\infty}\ f(t_n) = N(0,1)$$

Distribution de Cauchy ou t_1 de Student - Distribution de Breit-Wigner



Distribution de Cauchy ou t_1 de Student

$$f(t_1) = \frac{1}{\pi\left(1+t_1^2\right)}$$

$$\int_{-\infty}^{\infty} t_1^2 f(t_1)\,dt_1 = \infty \Rightarrow \begin{cases} \sigma^2 \text{ non défini !} \\ \text{le théorème central limite ne s'applique pas !} \\ \text{largeur à mi-hauteur (FWHM) : 2} \end{cases}$$

Distribution de Breit-Wigner

$\Gamma = \frac{\hbar}{\tau}$ =largeur intrinsèque de la masse d'une particule,

d'une particule instable de spin 0 et de vie moyenne τ

(processus poissonien $\rightarrow$ loi de désintégration exponentielle)

Interaction faible $\left(\tau \approx 10^{-6} - 10^{-12}\,s\right)$:

Γ << résolution expérimentale

Interaction forte $\left(\tau \approx 10^{-23}\,s\right)$: $\Gamma \approx \frac{6.6 \cdot 10^{-22}\,MeVs}{10^{-23}\,s} \approx 100\,MeV$

changement de variable $m = m_0 + t_1\,\Gamma/2$

$$f(m) = \frac{\Gamma/2}{\pi}\frac{1}{\left(m-m_0\right)^2 + \left(\Gamma/2\right)^2} \qquad \Gamma = \text{largeur à mi hauteur}$$

Distribution F de Fisher-Snedecor

- Soient - u_1, u_2 distribuées suivant des $\chi^2_{n_1}, \chi^2_{n_2}$

  - u_1 et u_2 indépendantes

$$F_{n_1 n_2} = \frac{u_1 / n_1}{u_2 / n_2}$$ est distribuée suivant une Fisher à n_1, n_2 degrés de liberté

Fonction de densité de probabilité:

$$f\left(F_{n_1 n_2}\right) = \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \times \frac{F_{n_1 n_2}^{\frac{n_1}{2}-1}}{\left(1 + \frac{1}{n_2} F_{n_1 n_2}\right)^{\frac{n_1+n_2}{2}}}$$

$$\mu = \frac{n_2}{n_2 - 2} \text{ si } n_2 > 2$$

$$\sigma^2 = \frac{2 n_2^{\,2}\left(n_1 + n_2 - 2\right)}{n_1\left(n_2 - 2\right)^2\left(n_2 - 4\right)} \text{ si } n_2 > 4$$

Démonstration:

changement de variables: $\begin{cases} F = \dfrac{u_1 / n_1}{u_2 / n_2} \\ z = v \end{cases}$

$$f\left(F, z\right) = f\left(u_1, u_2\right)\left|J\right| = f\left(u\right) f\left(v\right)\left|J\right| =$$

$$f\left(F\right) = \int_0^\infty f\left(F, z\right) dz$$

Statistiques distribuées suivant une F

- **Soient $(x_1,\ldots,x_n)$ indépendantes et extraites d'une $N(\mu_x,\sigma_x^2)$**

  et $(y_1,\ldots,y_m)$ indépendantes et extraites d'une $N(\mu_y,\sigma_y^2)$

$\dfrac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$ **est distribué suivant une** F_{nm}

$\dfrac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$ **est distribué suivant une** $F_{(n-1)(m-1)}$

Le rôle central de la distribution normale