

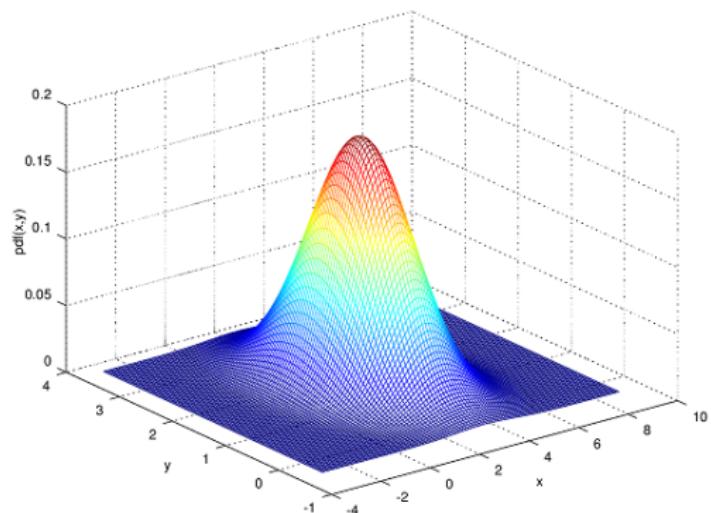


# STT 1000 - STATISTIQUES

ARTHUR CHARPENTIER



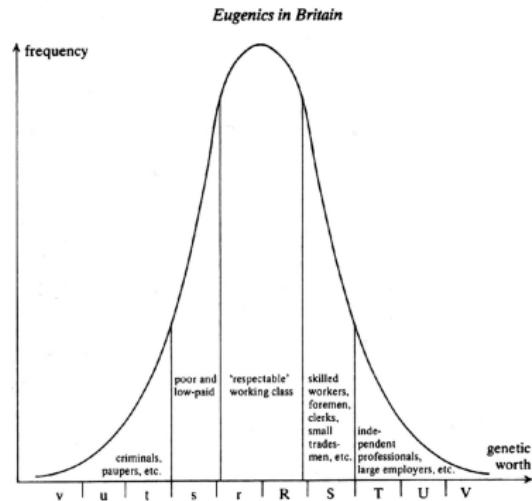
# Karl Friedrich Gauss & the Gaussian distribution



# Gaussian distribution

Legendre and Gauss (or Gauß) introduced the distribution as a *law of errors*...

Quetelet's average man  
Galton's view of British social structure (picture Eugenics in Britain)



Galton needed to revolutionize this branch of mathematics, error theory and the use of the Gauss distribution as a distribution of errors from a mean value. A new statistical paradigm was needed, The Structure of Scientific Revolutions, Kuhn 1970.

# Loi normale centrée & réduite

Loi normale / Gaussienne  $\mathcal{N}(0, 1)$

$X \sim \mathcal{N}(0, 1)$ , with density on  $\mathbb{R}$ ,

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$$

Loi normale / Gaussienne  $\mathcal{N}(0, 1)$

Si  $X \sim \mathcal{N}(0, 1)$ ,  $\mathbb{E}[X] = 0$  et  $\text{Var}[X] = 1$ .

# Gaussian Tables

In many applications we should solve

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp\left[-\frac{x^2}{2}\right] dx = p$$

no simple analytical formula...

Need for a **standard normal table**

Hence  $\Phi(1.64) = 95\%$

and  $\Phi(1.96) = 97.5\%$ ,

$$\Phi^{-1}(0.975) = 1.96$$

$$\Phi^{-1}(0.025) = -1.96$$

```
1 > qnorm(.95)
2 [1] 1.644854
3 > qnorm(.975)
4 [1] 1.959964
```

Table n° 3.

$$\text{VALEURS DE L'INTÉGRALE DÉFINIE } P_3 = \frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2} dt, \text{ POUR DES }$$

VALEURS DE  $t$  EXPRIMÉES EN FONCTION DE  $\varphi$  PRIS POUR UNITÉ.

$\frac{t}{\varphi}$	$\frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2} dt$	Différences	$\frac{t}{\varphi}$	$\frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2} dt$	Différences
0,0	0,000		2,5	0,908	
0,1	0,054	54	2,6	0,921	43
0,2	0,107	53	2,7	0,934	40
0,3	0,160	53	2,8	0,944	39
0,4	0,213	53	2,9	0,950	9
0,5	0,264	54	3,0	0,957	7
0,6	0,314	50	3,1	0,963	6
0,7	0,363	49	3,2	0,969	6
0,8	0,411	48	3,3	0,974	5
0,9	0,456	45	3,4	0,978	4
1,0	0,500	44	3,5	0,982	4
1,1	0,542	42	3,6	0,985	3
1,2	0,582	40	3,7	0,987	2
1,3	0,619	37	3,8	0,990	3
1,4	0,655	36	3,9	0,991	1
1,5	0,688	33	4,0	0,993	2
1,6	0,719	31	4,1	0,994	1
1,7	0,748	29	4,2	0,995	4
1,8	0,775	27	4,3	0,996	4
1,9	0,800	25	4,4	0,997	4
2,0	0,823	23	4,5	0,998	4
2,1	0,843	20	4,6	0,998	0
2,2	0,862	19	4,7	0,998	0
2,3	0,879	17	4,8	0,999	4
2,4	0,895	16	4,9	0,999	0
2,5	0,908	13	5,0	0,999	0

Cette table est indépendante de la précision des observations : elle donne la probabilité que l'erreur, pour une espèce quelconque d'observations, ne dépasse pas une certaine valeur exprimée en fonction de l'erreur probable.

Elle montre que, sur 1000 erreurs, il en reste 54 au-dessous de 0,1 de l'erreur probable ; 107 au-dessous de 0,2, etc. En d'autres termes, on peut parier 54 contre 946 que l'erreur que l'on commettra, dans une espèce quelconque d'observations, sera moindre que 0,1 de l'erreur probable ; 107 contre 895 qu'elle sera moindre que 0,2 de l'erreur probable, etc.

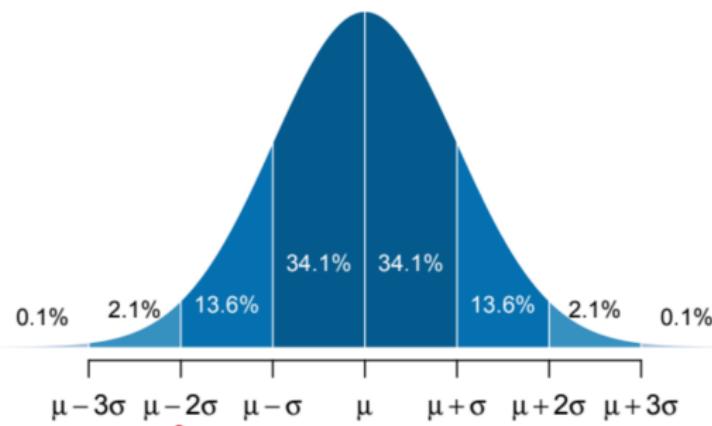
# Gaussian distribution

Loi normale / Gaussienne  $\mathcal{N}(\mu, \sigma^2)$

$X \sim \mathcal{N}(\mu, \sigma^2)$ , with density on  $\mathbb{R}$ , for  $\mu \in \mathbb{R}$  and  $\sigma \in \mathbb{R}_{+*}$

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

$\mathbb{E}(X) = \mu$  and  $\text{Var}(X) = \sigma^2$  (or  $\sigma$  is the standard deviation)



# Central Limit Theorem

Let  $X_i \sim \mathcal{B}(p)$ ,

$$\mathbb{P}(X_i = 0) = 1 - p \text{ and } \mathbb{P}(X_i = 1) = p.$$

then  $X = X_1 + \cdots + X_n \sim \mathcal{B}(n, p)$  (binomial distribution), for  $k = 0, 1, \dots, n$ ,

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

then, when  $n$  is large enough

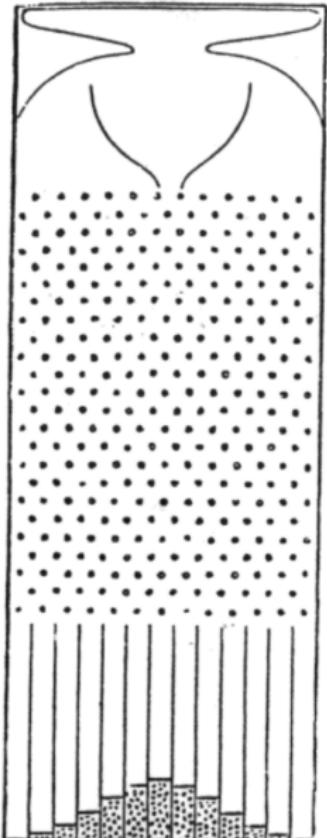
$$X \simeq \mathcal{N}(np, np(1-p))$$

or

$$\bar{X} = \frac{X}{n} \simeq \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

(picture Quincunx, or Galton's box)

FIG. 7.



# Central Limit Theorem

If  $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$  and  $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$  are independent,

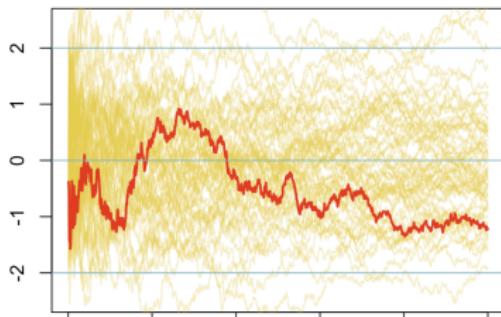
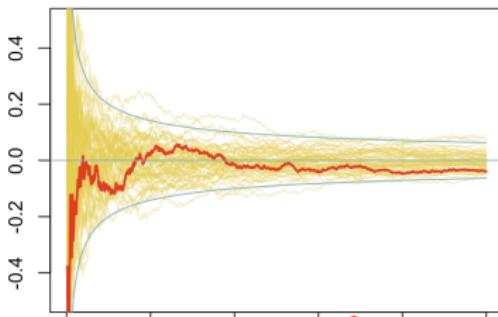
$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

## Central Limit Theorem

Suppose  $\{X_1, \dots, X_n, \dots\}$  is a sequence of i.i.d. random variables with  $\mathbb{E}[X_i] = \mu$  and  $\text{Var}[X_i] = \sigma^2 < \infty$ , then, if

$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$  as  $n$  goes to infinity,

$$\sqrt{n}(\bar{X}_n - \mu) \rightarrow \mathcal{N}(0, \sigma^2).$$



## Gaussian (multivariate) distribution

$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , with density

$$f_{\mathbf{X}}(x_1, \dots, x_d) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

where  $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$  and  $\text{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$ .

Estimates are  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$  and  $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$

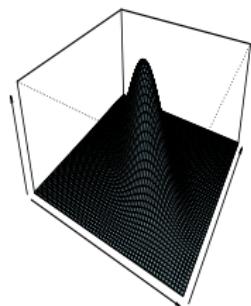
In dimension 2,  $f(x, y)$  is proportional to

$$\exp\left(-\frac{1}{2(1-\rho^2)} \left[ \frac{(x - \mu_X)^2}{\sigma_X^2} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x - \mu_X)(y - \mu_Y)}{\sigma_X \sigma_Y} \right]\right)$$

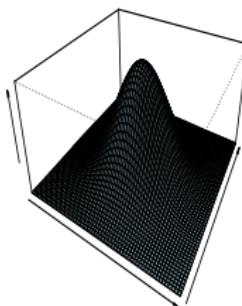
levels curves (isodensities) are ellipses.

# Gaussian (multivariate) distribution

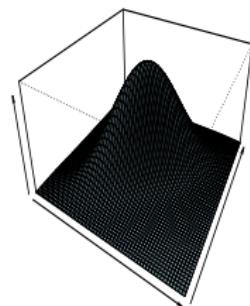
Densité du vecteur Gaussien,  $r=0.7$



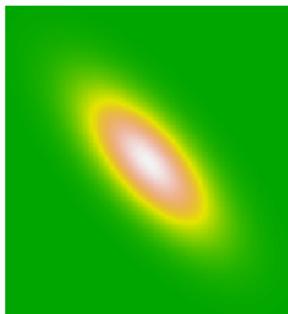
Densité du vecteur Gaussien,  $r=0.0$



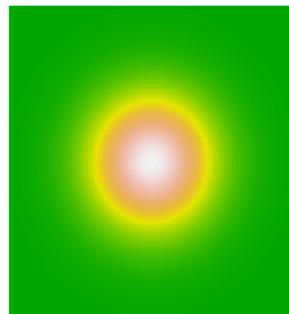
Densité du vecteur Gaussien,  $r=-0.7$



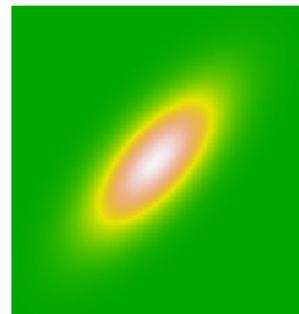
Courbes de niveau du vecteur Gaussien,  $r=-0.7$



Courbes de niveau du vecteur Gaussien,  $r=0.0$



Courbes de niveau du vecteur Gaussien,  $r=0.7$



## Random Vectors

Soit  $\mathbf{X}$  un vecteur aléatoire de dimension  $d$

- ▶ L'espérance de  $\mathbf{X}$ , notée  $\mathbb{E}(\mathbf{X})$  est définie (si elle existe) par le vecteur de dimension  $d$   $\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \dots, \mathbb{E}(X_d))^\top$ .
- ▶ La matrice de covariance (appelée aussi matrice de variance-covariance de  $\mathbf{X}$ ) est définie (si elle existe) par la matrice de taille  $(d, d)$

$$\text{Var}(\mathbf{X}) = \mathbb{E} \left( (\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^\top \right).$$

Ainsi le terme  $ij$  de cette matrice représente la covariance entre  $X_i$  et  $X_j$ ,

$$\text{Cov}(X_i, X_j) = \mathbb{E} [(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))].$$

## Random Vectors

Soit  $\mathbf{X}$  un vecteur aléatoire de dimension  $d$ , de moyenne  $\mu$  et de matrice de covariance  $\Sigma$ .

Soient  $\mathbf{A}$  et  $\mathbf{B}$  deux matrices réelles de taille  $(d, p)$  et  $(d, q)$  et enfin soit  $\mathbf{a} \in \mathbb{R}^p$  alors

- ▶  $\text{Var}(\mathbf{X}) = \mathbb{E}((\mathbf{X} - \mu)(\mathbf{X} - \mu)^\top) = \mathbb{E}(\mathbf{X}\mathbf{X}^\top) - \mu\mu^\top.$
- ▶  $\mathbb{E}(\mathbf{A}^\top \mathbf{X} + \mathbf{a}) = \mathbf{A}^\top \mu + \mathbf{a}.$
- ▶  $\text{Var}(\mathbf{A}^\top \mathbf{X} + \mathbf{a}) = \mathbf{A}^\top \Sigma \mathbf{A}.$
- ▶  $\text{Cov}(\mathbf{A}^\top \mathbf{X}, \mathbf{B}^\top \mathbf{X}) = \mathbf{A}^\top \Sigma \mathbf{B}.$

# The Gaussian Distribution

A **Gaussian variable**, with distribution  $\mathcal{N}(\mu, \sigma^2)$ , where  $\mu \in \mathbb{R}$  and  $\sigma > 0$ , has density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \text{ for all } x \in \mathbb{R}.$$

Then  $\mathbb{E}(X) = \mu$  and  $\text{Var}(X) = \sigma^2$ .

Observe that if  $Z \sim \mathcal{N}(0, 1)$ ,  $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$ .

The **Gaussian vector**  $\mathcal{N}(\mu, \Sigma)$ :  $\mathbf{X} = (X_1, \dots, X_n)$  is a Gaussian vector with mean  $\mathbb{E}(\mathbf{X}) = \mu$  and covariance matrix

$\text{Var}(\mathbf{X}) = \Sigma = \mathbb{E}\left((\mathbf{X} - \mu)(\mathbf{X} - \mu)^\top\right)$  non-degenerated ( $\Sigma$  is invertible) if its density is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu)\right), \quad \mathbf{x} \in \mathbb{R}^n,$$

see **multivariate Gaussian distribution**

# The Gaussian Distribution

If  $\mathbf{X}$  is a Gaussian vector, then for any  $i$ ,  $X_i$  has a (univariate) Gaussian distribution, but its converse is not necessarily true.

Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random vector with mean  $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$  and with covariance matrix  $\boldsymbol{\Sigma}$ , if  $\mathbf{A}$  is a  $k \times n$  matrix, and  $\mathbf{b} \in \mathbb{R}^k$ , then  $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$  is a Gaussian vector  $\mathbb{R}^k$ , with distribution  $\mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$ .

Observe that if  $(X_1, X_2)$  is a Gaussian vector,  $X_1$  and  $X_2$  are independent if and only if

$$\text{Cov}(X_1, X_2) = \mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))) = 0.$$

# The Gaussian Distribution

Let  $\mathbf{Z} = (Y, X) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{Y,Y} & \Sigma_{Y,X} \\ \Sigma_{X,Y} & \Sigma_{X,X} \end{pmatrix}$$

Then

$$Y|X = x \sim \mathcal{N}(\mu_{Y|x}, \sigma_{Y|x}^2) \text{ where } \begin{cases} \mu_{Y|x} = \mu_Y + \Sigma_{Y,X}\Sigma_{X,X}^{-1}(x - \mu_X) \\ \sigma_{Y|x}^2 = \Sigma_{Y,Y} - \Sigma_{Y,X}\Sigma_{X,X}^{-1}\Sigma_{X,Y} \end{cases}$$

Hence,  $\mathbb{E}[Y|X = x] = \mu_{Y|x}$  is linear in  $x$ , with slope

$$\text{Corr}(X, Y)\sqrt{\Sigma_{Y,Y}\Sigma_{X,X}^{-1}}$$

and  $\text{Var}[Y|X = x] = \sigma_{Y|x}^2$  is constant

fw (furthermore  $\text{Var}[Y|X = x] \leq \text{Var}[Y]$ )

# Chi-Square Distribution

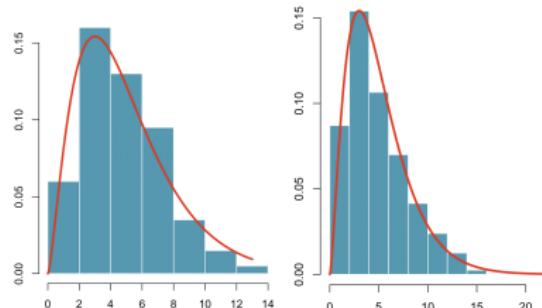
Chi-deux  $\chi^2(\nu)$

The chi-squared distribution  $\chi^2(\nu)$ , with  $\nu \in \mathbb{N}^*$  has density

$$x \mapsto \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \text{ where } x \in [0; +\infty),$$

where  $\Gamma$  denotes the Gamma function ( $\Gamma(n+1) = n!$ ).

Observe that  $\mathbb{E}(X) = \nu$  et  $\text{Var}(X) = 2\nu$ .  $\nu$  are the degrees of freedom, see chi-squared distribution



# Chi-Square Distribution

Chi-deux  $\chi^2(\nu)$

If  $X_1, \dots, X_\nu \sim \mathcal{N}(0, 1)$  are independent variables, then

$$Y = \sum_{i=1}^{\nu} X_i^2 \sim \chi^2(\nu), \text{ when } \nu \in \mathbb{N}.$$

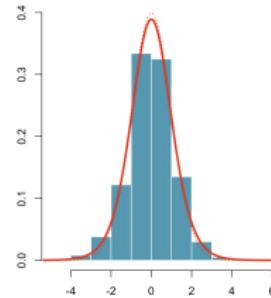
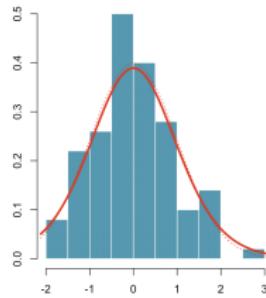
This is a particular case of the Gamma distribution,

$$X \sim \mathcal{G}\left(\frac{k}{2}, \frac{1}{2}\right), \text{ see see } \text{gamma distribution}$$

# Student's $t$ Distribution

Student  $t$   $\mathcal{St}(\nu)$

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}, \text{ on } \mathbb{R}$$



Student  $t$   $\mathcal{St}(\nu)$

$$\mathbb{E}(X) = 0 \text{ and } \text{Var}(X) = \frac{\nu}{\nu - 2} \text{ when } \nu > 2.$$

## Student's $t$ Distribution

Student  $t$   $\mathcal{St}(\nu)$

If  $X \sim \mathcal{N}(0, 1)$  and  $Y \sim \chi^2(\nu)$  are independents, then

$$T = \frac{X}{\sqrt{Y/\nu}} \sim \mathcal{St}(\nu).$$

see Student's  $t$

Let  $X_1, \dots, X_n$  be  $\mathcal{N}(\mu, \sigma^2)$  independent random variables. Let

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n} \text{ and } S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Then  $\frac{(n-1)S_n^2}{\sigma^2}$  has a  $\chi^2(n-1)$  distribution, and furthermore

$$T = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \sim \mathcal{St}(n-1).$$

## Fisher's $F$ Distribution

Loi de Fisher  $\mathcal{F}(d_1, d_2)$

$$f(x) = \frac{1}{x B(d_1/2, d_2/2)} \left( \frac{d_1 x}{d_1 x + d_2} \right)^{d_1/2} \left( 1 - \frac{d_1 x}{d_1 x + d_2} \right)^{d_2/2}$$

for  $x \geq 0$  and  $d_1, d_2 \in \mathbb{N}$ , where  $B$  denotes the Beta function.

Loi de Fisher  $\mathcal{F}(d_1, d_2)$

$$\mathbb{E}(X) = \frac{d_2}{d_2 - 2} \text{ when } d_2 > 2$$

$$\text{Var}(X) = \frac{2 d_2^2 (d_1 + d_2 - 2)}{d_1 (d_2 - 2)^2 (d_2 - 4)} \text{ when } d_2 > 4.$$

## Fisher's $F$ Distribution

If  $X \sim \mathcal{F}(\nu_1, \nu_2)$ , then  $\frac{1}{X} \sim \mathcal{F}(\nu_2, \nu_1)$ .

Loi de Fisher  $\mathcal{F}(d_1, d_2)$

If  $X_1 \sim \chi^2(\nu_1)$  and  $X_2 \sim \chi^2(\nu_2)$  are independent

$$Y = \frac{X_1/\nu_1}{X_2/\nu_2} \sim \mathcal{F}(\nu_1, \nu_2)$$

see [Fisher's  \$\mathcal{F}\$](#)  on wikipedia

## Fisher's $F$ Distribution

On peut montrer que si  $X \sim Std(\nu)$ , alors  $X^2 \sim \mathcal{F}(1, \nu)$ . Ou dit autrement si  $F_{1-p}$  est le quantile de niveau  $1 - p$  de la loi  $\mathcal{F}(1, \nu)$ ,  $F_{1-p} = t_{1-p/2}^2$  où  $t_{1-p}$  est le quantile de niveau  $1 - p$  de la loi  $Std(\nu)$ .

La loi  $\mathcal{F}(1, \nu)$  a pour densité

$$f(u) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)} \nu^{\nu/2} u^{-1/2} (\nu + u)^{-(\nu+1)/2} \text{ sur } \mathbb{R}_+$$

$$f(u) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} u^{-1/2} \left(1 + \frac{u}{\nu}\right)^{-(\nu+1)/2} \text{ sur } \mathbb{R}_+$$

aussi

$$\int_0^{F_{1-p}} f(u) du = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \int_0^{F_{1-p}} u^{-1/2} \left(1 + \frac{u}{\nu}\right)^{-(\nu+1)/2} du = 1 - p$$

## Fisher's $F$ Distribution

Faisons le changement de variable,  $t = \sqrt{u}$ ,

$$2 \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \int_0^{\sqrt{F_{1-p}}} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} dt = 1 - p$$

on reconnaît une intégrale associée à la loi de Student.

Si  $T \sim Std(\nu)$ , on a écrit  $\mathbb{P}(T \in [0, \sqrt{F_{1-p}}])$ ,

$$2\mathbb{P}(T \in [0, \sqrt{F_{1-p}}]) = 1 - p \text{ i.e. } \frac{1 - p}{2} = \mathbb{P}(T \leq \sqrt{F_{1-p}}) - \underbrace{\mathbb{P}[T \leq 0]}_{=1/2}$$

$$\mathbb{P}(T \leq \sqrt{F_{1-p}}) = 1 - \frac{p}{2} \text{ mais on sait que } \mathbb{P}(T \leq t_{1-p/2}) = 1 - \frac{p}{2}$$

donc  $F_{1-p} = t_{1-p/2}^2$ .

```
1 > qf(.95, 1, 10)
2 [1] 4.964603
3 > qt(.975, 10)^2
4 [1] 4.964603
```