Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

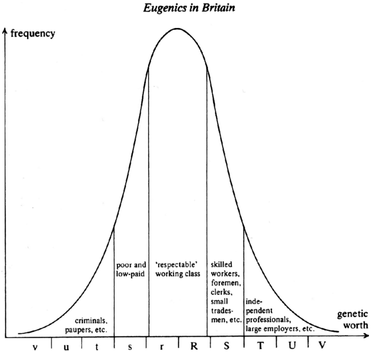
08 - La loi normale et les lois dérivées

été 2022

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 $\Phi(1.96) = 97.5\%$,
$\Phi^{-1}(0.975) = 1.96$
$\Phi^{-1}(0.025) = -1.96$

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

Table n° 3.

VALEURS DE L'INTÉGRALE DÉFINIE $P_2 = \frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2} dt$, POUR DES VALEURS DE t EXPRIMÉES EN FONCTION DE ρ PRIS POUR UNITÉ.

$\frac{t}{\rho}$	$\frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2} dt.$	Différences	$\frac{t}{\rho}$	$\frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2} dt.$	Différences
0,0	0,000	54	2,5	0,908	13
0.1	0,054	53	2,6	0,921	10
0,2	0,107	53	2,7	0,934	10
0.3	0,160	53	2,8	0,944	9
0.4	0,213	51	2,9	0,950	7
0.5	0,264	50	3.0	0,957	6
0,6	0,314	49	3,1	0,963	6
0.7	0.363	48	3,2	0,969	5
0.8	0,411	45	3.3	0,974	4
0.9	0.456	44	3,4	0,978	4
1.0	0,500	42	3,5	0,982	3
1.1	0.542	40	3,6	0,985	2
1,2	0.582	37	3,7	0,987	3
1.3	0,619	36	3.8	0,990	1
1,4	0,655	33	3,9	0,991	2
1,5	0,688	31	4,0	0,993	1
1.6	0.719	29	4,1	0,994	1
1,7	0,748	27	4,2	0,995	1
1,8	0,775	25	4,3	0,996	1
1,9	0.800	23	4,4	0,997	1
2,0	0,823	20	4.5	0,998	0
2,1	0,843	19	4,6	0,998	0
2,2	0,862	17	4.7	0,998	1
2,3	0,879	16	4.8	0,999	0
2,4	0,895	13	4,9	0,999	0
2.5	0.908		5,0	0,999	

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.

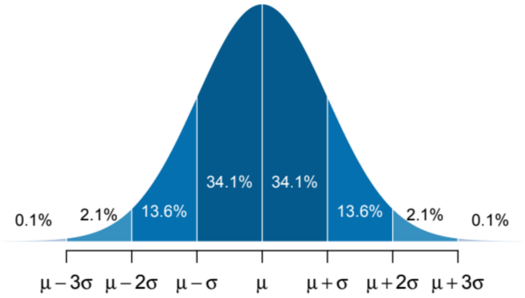
Elle montre que, sur 1000 erreurs, il en reste 54 au-dessous de 0,1 de l'erreur probable ; 107 au-dessous de 0,2, etc. En d'autres termes, on peut parier 54 contre 946 que l'erreur que l'on commettra, dans une espèce quelconque d'observations, sera moindre que 0,1 de l'erreur probable ; 107 contre 893 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 $\sigma\in\mathbb{R}_{+\star}$

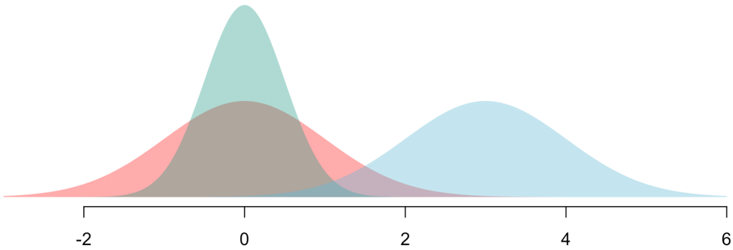
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$



Gaussian distribution

Si $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

Sur le dessin ci-dessous, il y a les densités de trois lois normales, $\mathcal{N}(0,1)$, $\mathcal{N}(0,0.5)$, $\mathcal{N}(3,1)$.



Loi normale / centréee-réduite

Si $X \sim \mathcal{N}(\mu, \sigma^2)$ alors $Z = \dfrac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$.

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 $k = 0, 1, \cdots, n$,

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

then, when n is large enough

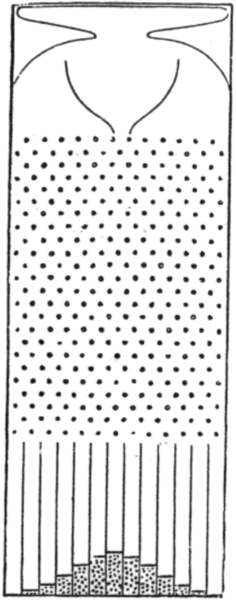
$$X \simeq \mathcal{N}(np, np(1-p))$$

or

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

(picture Quincunx, or Galton's box)

FIG. 7.



Central Limit Theorem

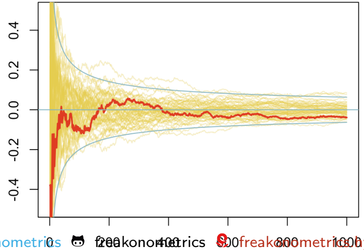
If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent,

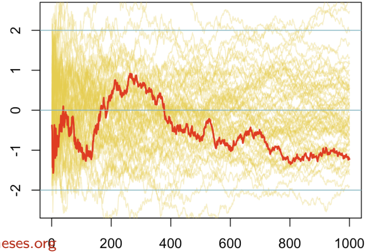
$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

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

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \to \mathcal{N}\left(0, \sigma^2\right).$$





Chi-Square Distribution

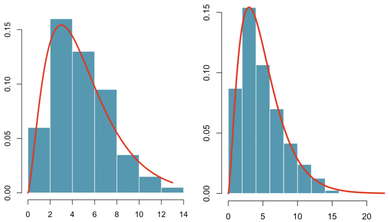
Chi-deux $\chi^2(\nu)$

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

$\mathbb{E}(X) = \nu$ et $\mathrm{Var}(X) = 2\nu$, cf chi-squared distribution



Chi-Square Distribution

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

If $X_1, \cdots, X_\nu \sim \mathcal{N}(0,1)$ are independent variables, then
$Y = \sum_{i=1}^{\nu} X_i^2 \sim \chi^2(\nu)$, when $\nu \in \mathbb{N}_*$.

Somme de Chi-deux $\chi^2(\nu)$ indépendantes

Si $X \sim \chi^2(\mu)$ et $Y \sim \chi^2(\nu)$ sont indépendantes,

$$X + Y \sim \chi^2(\mu + \nu)$$

Chi-Square Distribution ★★★

Chi-deux $\chi^2(\nu-1)$

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

Then ${S_n}^2 = \frac{1}{\sigma^2}\sum_{i=1}^n\left(X_i-\overline{X}\right)^2$ has a $\chi^2(\nu-1)$ distribution.

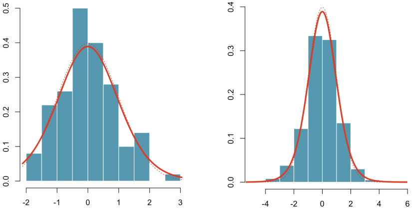
Preuve (heuristique):

$$\frac{1}{\sigma^2}\sum_{i=1}^n(X_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^n\left(X_i-\overline{X}\right)^2 + \underbrace{\frac{1}{\sigma^2}\left(\overline{X}-\mu\right)^2}_{\sim\chi^2(1)} \sim \chi^2(\nu)$$

Student's t Distribution

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

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



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

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

Student's t Distribution

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

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

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

see Student's t

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

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

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

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

Fisher's F Distribution ★★★

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

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

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

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

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

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

Fisher's F Distribution ★★★

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

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 $X \sim \mathcal{S}td(\nu)$, 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 $\mathcal{S}td(\nu)$.

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 $T\sim\mathcal{S}td(\nu)$, 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 $F_{1-p}=t^2_{1-p/2}$.

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