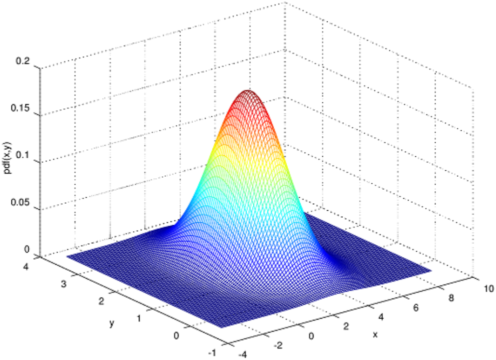
Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#431 Gaussian distribution

été 2020

Karl Friedrich Gauss & the Gaussian distribution

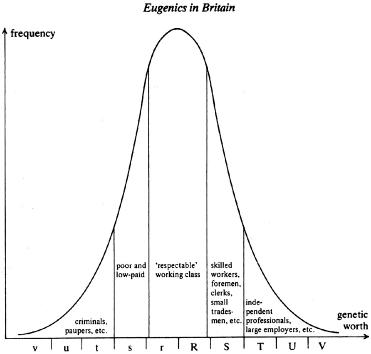


Gaussian distribution

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)

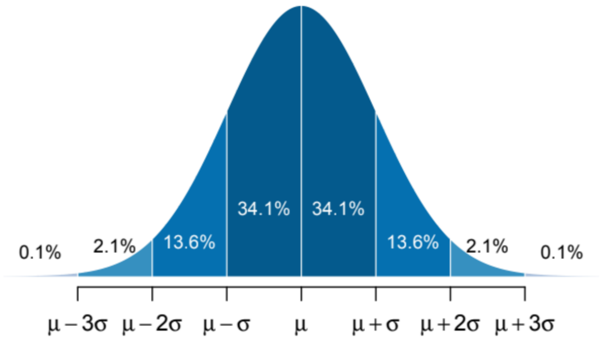


Galton needed to revolutionize this branch of mathematics, error theory and the use of the Gauss distribution as a distribution of errors from a mean value. A new statistical paradigm was needed, The Structure of Scientific Revolutions, Kuhn 1970.

Gaussian distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$, with density $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right]$

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



Observe that $\dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ (standard score, or normalizing)

Gaussian Tables

In many applications we should solve

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

no simple analytical formula...
Need for a standard normal table
Hence $F(1.64) = 95\%$
and $F(1.96) = 97.5\%$.

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	0,931	10
0.3	0,160	53	2.8	0,941	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.

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

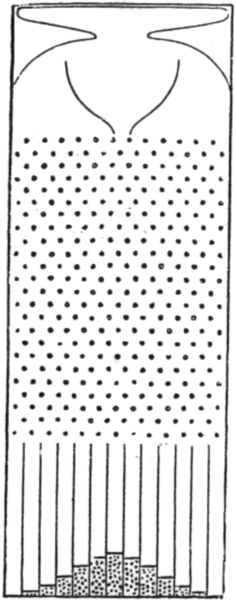
$$X \simeq \mathcal{N}(np, np(1-p))$$

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.



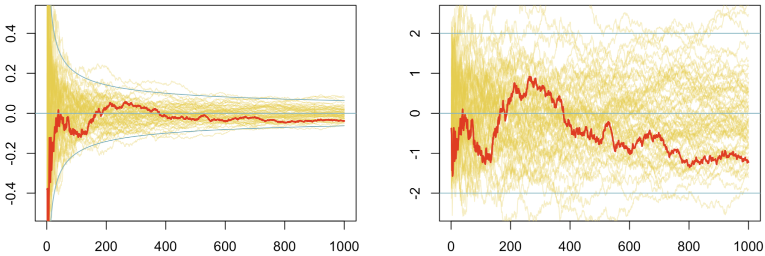
Central Limit Theorem

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 $\{X_1, \ldots, X_n, \ldots\}$ 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 $\overline{X}_n = X_1 + \ldots + X_n$ as n goes to infinity, $\sqrt{n}(\overline{X}_n - \mu)$ converges toward a $\mathcal{N}(0, \sigma^2)$ distribution

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



Gaussian (multivariate) distribution

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

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

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

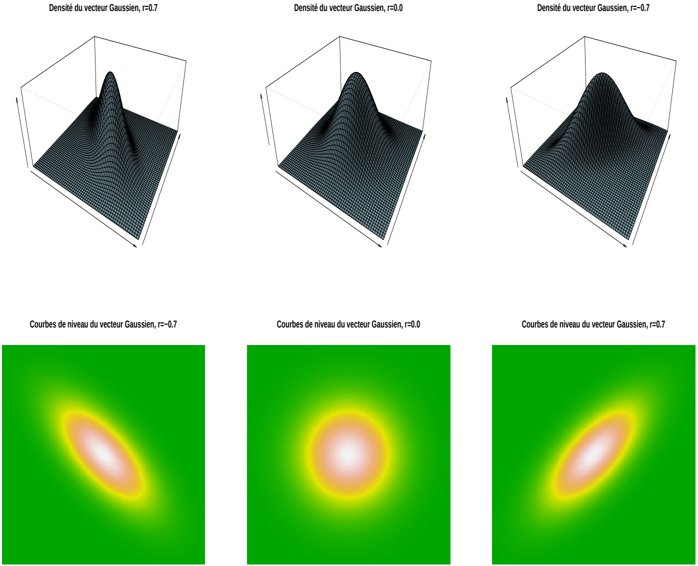
Estimates are $\overline{x} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top$

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



Chi-Square

If $Z_1, \cdots, Z_k$ are independent $\mathcal{N}(0,1)$ variables,

$$Q = \sum_{i=1}^{k} Z_i^2, \sim \chi_k^2$$

(see wikipedia)