1 Conceptes generals

1.1 Definicions

Sovint no coneixem el *valor exacte* $\bar{x}$ d'una magnitud i hem de treballar amb un valor aproximat x. Aleshores definim:

- *Error absolut de* x: $e_a(x) = x - \bar{x}$.
- *Error relatiu de* x: $e_r(x) = \dfrac{e_a(x)}{\bar{x}}$, i a la pràctica es pren $e_r(x) \simeq \dfrac{e_a(x)}{x}$ si $x, \bar{x} \neq 0$.

En general només coneixem una fita d'aquests errors: $|e_a(x)| \leq \varepsilon_a(x)$, $|e_r(x)| \leq \varepsilon_r(x)$.

Proposició. *Si* ε_a, ε_r *són fites de l'error absolut i relatiu, respectivament, llavors* $|\bar{x} - x| \leq \varepsilon_a(x)$ *i* $|\bar{x} - x| \leq |x|\varepsilon_r(x)$. *Emprarem les notacions* $\bar{x} = x \pm \varepsilon_a(x)$ *i* $\bar{x} = x(1 \pm \varepsilon_r(x))$.

Demostració. Ambdúes desigualtats es dedueixen immediatament de la definició. □

1.2 Fonts d'error

Des del punt de vista numèric ens centrarem en tres *fonts d'error*: errors en les dades d'entrada, errors d'arrodoniment durant els càlculs, i errors de truncament del mètode emprat.

1.2.1 Errors en les dades d'entrada

Els errors en les dades poden ser deguts en primera instància a mesuraments inexactes, doncs els aparells que ens proporciona la tecnologia no tenen una precisió infinita. D'altra banda quan treballem numèricament amb ordinadors o calculadores ens veiem obligats a emprar un nombre finit de xifres per cada dada.

Sabem que donada una base natural $b > 1$, qualsevol nombre real positiu es pot escriure de forma única de la següent manera, que anomenem *representació digital de* x *en base* b:

$$x = c_n b^n + c_{n-1} b^{n-1} + \cdots + c_1 b + c_0 + a_1 b^{-1} + a_2 b^{-2} + a_3 b^{-3} + \cdots$$

on $c_i, a_i \in \mathbb{Z}$, $0 \leq c_i, a_i < b$, i els a_i no són tots iguals a 9 d'un lloc en endavant. Habitualment escrivim aquest nombre com $(c_n c_{n-1} \ldots c_0.a_1 a_2 \ldots)_b$. Suposem que treballem

amb una màquina que només permet representar nombres amb t xifres en base b. Llavors tindrem la següent *representació en punt flotant* de x que notarem $\mathrm{fl}(x)$:

$$\mathrm{fl}(x) = s \cdot m \cdot b^q,$$

on $s = \pm 1$, $q \in \mathbb{Z}$, $m = (a_1.a_2a_3 \ldots a_t)_b$, $0 \leq a_i < b$ per $i = 1, \ldots, t$ i $a_1 \neq 0$. El nombre m s'anomena *mantissa*, q *exponent* i s és el *signe*. Un cas particular d'aquesta representació és la *notació científica*, on es pren la base decimal $b = 10$. Per exemple, el nombre d'Avogadro escrit en notació científica amb 4 xifres significatives és $6.022 \cdot 10^{23}$.

Per passar de x a $\mathrm{fl}(x)$ ho podem fer per *tall*, simplement suprimint els dígits de x a partir de a_t, o per *arrodoniment*, de manera que $|\mathrm{fl}(x) - x|$ sigui mínim. Això equival a tallar en a_t si $0 \leq a_{t+1} < [b/2]$, o a tallar en a_t i augmentar a_t en una unitat si $[b/2] \leq a_{t+1} < b$. En aquest darrer cas, si $a_t + 1 = b$ posem $a_t = 0$ i augmentem en una unitat a_{t-1}. Si tornem a tenir $a_{t-1} + 1 = b$ fem el mateix, i així successivament fins tenir un $a_i + 1 < b$ o bé acabar amb $m = (1.00 \ldots 0)_b$ i augmentant una unitat l'exponent.

Exemple. El nombre $x = -99.962$ en notació científica és $x = -9.9962 \cdot 10^1$. Tallat a 3 xifres és $\mathrm{fl}_T(x) = -9.99 \cdot 10^1$, i arrodonit a 3 xifres, $\mathrm{fl}_A(x) = -1.00 \cdot 10^2$.

Proposició. *Les fites de l'error relatiu comès per* $\mathrm{fl}(x)$ *en cada cas són:*

- *Per tall:* $\varepsilon_r(\mathrm{fl}_T(x)) = b^{-t+1}$.
- *Per arrodoniment:* $\varepsilon_r(\mathrm{fl}_A(x)) = \dfrac{1}{2} b^{-t+1}$.

Demostració. Tenim $x = \pm(a_1.a_2 \ldots a_t a_{t+1} \ldots)_b \cdot b^q$ i $\mathrm{fl}_T(x) = \pm(a_1.a_2 \ldots a_t)_b \cdot b^q$. D'aquí tenim, per una banda, que $|x - \mathrm{fl}_T(x)| = (0.0 \ldots 0 a_{t+1} \ldots)_b \cdot b^q = (a_{t+1}.a_{t+2} \ldots)_b \cdot b^{q-t}$, i per tant $|x - \mathrm{fl}_T(x)| < b \cdot b^{q-t} = b^{q-t+1}$. D'altra banda $|\mathrm{fl}_T(x)| \geq b^q$, i combinant ambdúes fites tenim

$$|e_r(\mathrm{fl}_T(x))| = \left| \frac{e_a(\mathrm{fl}_T(x))}{\mathrm{fl}_T(x)} \right| = \frac{|x - \mathrm{fl}_T(x)|}{|\mathrm{fl}_T(x)|} \leq \frac{b^{q-t+1}}{b^q} = b^{-t+1}.$$

La fita en el cas de l'arrodoniment s'aconsegueix de la mateixa manera, notant ara que en ser mínim $|x - \mathrm{fl}_A(x)|$ tenim una fita millor $|x - \mathrm{fl}_A(x)| < b/2 \cdot b^{q-t} = b^{q-t+1}/2$ i per tant

$$|e_r(\mathrm{fl}_A(x))| = \frac{|x - \mathrm{fl}_A(x)|}{|\mathrm{fl}_A(x)|} \leq \frac{b^{q-t+1}}{2b^q} = \frac{1}{2} b^{-t+1}. \qquad \square$$

1.2.2 Errors d'arrodoniment durant els càlculs

L'error en un resultat no solament pot provenir dels erros de les dades d'entrada, sinó també dels errors d'arrodoniment durant els càlculs intermedis. Així per exemple si tenim

$x = 1.111 \cdot 10^{-1}$ i treballem amb una màquina que treballa amb 4 xifres decimals arrodonint, quan volem calcular x^2 tenim que $x^2 = 1.234321 \cdot 10^{-2}$ i $\mathrm{fl}_A(x^2) = 1.234 \cdot 10^{-2}$. Cometem doncs un $e_a = 3.21 \cdot 10^{-6}$ i un $e_r = 3.21 \cdot 10^{-6}/(1.234 \cdot 10^{-2}) = 2.601 \cdot 10^{-4}$.

Si no es diu el contrari es suposa que la fita de l'error relatiu en cada operació aritmètica $+, -, \cdot, /$ és la mateixa que en la representació de les dades de partida.

1.2.3 Error de truncament del mètode emprat

Quan resolem un problema per mètodes numèrics (per exemple quan aproximem una integral per una suma finita o una derivada per un quocient incremental), encara que partim d'unes dades exactes i fem les operacions exactament obtenim només una aproximació numèrica del resultat. Aquest error produït depèn del mètode que haguem emprat, i rep el nom d'*error de truncament*.

És important conèixer una expressió que ens doni una fita d'aquest error per cada mètode numèric que emprem.

A més d'aquest error de truncament també caldrà tenir en compte els errors d'arrodoniment de cadascuna de les operacions que comporti el mètode.

Caldrà evitar, per exemple, els mètodes que comportin *cancel·lacions* de xifres en restar dues quantitats molt properes, atès que es produeixen error relatius grans. Quan ens trobem en aquest cas caldrà cercar expressions algebraiques equivalents que ho evitin. En general s'han d'evitar els algorismes *numèricament inestables*, això és, aquells que produeixen variacions grans en el resultat final davant de variacions petites en les dades inicials. Veiem-ne un exemple: suposem que volem trobar un mètode per calcular les integrals

$$I_n = \int_0^1 \frac{x^n}{a+x}\, dx$$

per a cada n. Tenint en compte que

$$I_n = \int_0^1 \frac{x^{n-1}(x+a-a)}{a+x}\, dx = \frac{1}{n} - aI_{n-1}$$

podem emprar l'algorisme iteratiu $I_0 = \ln|(1+a)/a|$, $I_n = 1/n - aI_{n-1}$. Aleshores, amb les notacions del principi, tenim la següent expressió per l'error absolut:

$$e_n = I_n - \bar{I_n} = \left(\frac{1}{n} - aI_{n-1}\right) - \left(\frac{1}{n} - a\bar{I_{n-1}}\right) = -a\left(I_{n-1} - \bar{I_{n-1}}\right) = -ae_{n-1}$$

de manera que $e_n = (-a)^n e_0$, és a dir, a cada iteració l'error relatiu es multiplica per a, i si $|a| > 1$ aquest mètode no és convenient perquè propaga l'error molt depressa.

2 Estimació i fitació d'errors

L'objectiu de qualsevol estudi d'errors és mirar de conèixer l'efecte que té sobre el resultat final d'un problema numèric cadascun dels diferents tipus d'errors que hi poden tenir lloc. Els tipus d'errors que distingirem seran els tres que hem vist abans: els de les dades, els d'arrodoniment en els càlculs intermedis i els de truncament del mètode emprat. Cadascun d'aquests errors contribuïrà en l'error total del resultat final.

2.1 Propagació dels errors de les dades

Proposició. *Donades dues dades x, y afectades d'error, amb fites $\varepsilon_a(x), \varepsilon_a(y)$ per l'error absolut i fites $\varepsilon_r(x), \varepsilon_r(y)$ pel relatiu, tenim les següents fites per les operacions elementals:*

$$\begin{aligned}\varepsilon_a(x \pm y) &= \varepsilon_a(x) + \varepsilon_a(y)\\ \varepsilon_r(xy) = \varepsilon_r(x/y) &\simeq \varepsilon_r(x) + \varepsilon_r(y)\end{aligned}$$

Demostració. Amb les notacions habituals tenim $x = \bar{x} + e_a(x)$, $y = \bar{y} + e_a(y)$, per tant $x \pm y = \bar{x} \pm \bar{y} + (e_a(x) \pm e_a(y))$ i llavors tenim la fita

$$|e_a(x \pm y)| = |e_a(x) \pm e_a(y)| \leq |e_a(x)| + |e_a(y)| = \varepsilon_a(x) + \varepsilon_a(y).$$

Considerant ara els errors relatius tenim $x = \bar{x}(1 + e_r(x))$, $y = \bar{y}(1 + e_r(y))$, de manera que $xy = \bar{x}\bar{y}(1 + e_r(x) + e_r(y) + e_r(x)e_r(y)) \simeq \bar{x}\bar{y}(1 + e_r(x) + e_r(y))$ ja que el producte $e_r(x)e_r(y)$ és despreciable respecte de la resta de termes si considerem que els errors relatius són petits. Així doncs tenim la fita aproximada

$$|e_r(xy)| \simeq |e_r(x) + e_r(y)| \leq |e_r(x)| + |e_r(y)| = \varepsilon_r(x) + \varepsilon_r(y).$$

Procedim de forma similar pel quocient, de nou tenint en compte que podem despreciar els termes $e_r^2(y)$ i $e_r(x)e_r(y)$:

$$\begin{aligned}\frac{x}{y} &= \frac{\bar{x}}{\bar{y}}\frac{1 + e_r(x)}{1 + e_r(y)} = \frac{\bar{x}}{\bar{y}}\frac{(1 + e_r(x))(1 - e_r(y))}{1 - e_r^2(y)} \simeq \frac{\bar{x}}{\bar{y}}(1 + e_r(x))(1 - e_r(y)) =\\ &= \frac{\bar{x}}{\bar{y}}(1 + e_r(x) - e_r(y) - e_r(x)e_r(y)) \simeq \frac{\bar{x}}{\bar{y}}(1 + e_r(x) - e_r(y)),\end{aligned}$$

i aleshores tenim la fita aproximada

$$|e_r(x/y)| \simeq |e_r(x) - e_r(y)| \leq |e_r(x)| + |e_r(y)| = \varepsilon_r(x) + \varepsilon_r(y). \qquad \square$$

També és natural preguntar-nos com es veu afectat l'error en una dada quan hi apliquem una funció:

Teorema (fórmula aproximada de propagació de l'error maximal). *Sigui f una funció derivable amb continuïtat, x una dada inicial afectada d'error, i $y = f(x)$. Aleshores tenim la següent fita aproximada per l'error absolut:*

$$\varepsilon_a(y) \simeq |f'(x)|\, \varepsilon_a(x).$$

DEMOSTRACIÓ. El resultat és conseqüència directa del teorema del valor mitjà per funcions d'una variable derivables amb continuïtat. En efecte, si $e_a(x) = x - \bar{x}$ llavors

$$e_a(y) = y - \bar{y} = f(x) - f(\bar{x}) = f'(\xi)(x - \bar{x}) = f'(\xi)e_a(x)$$

per algun ξ comprès entre x i $\bar{x}$. Per raons pràctiques podem aproximar $f'(\xi) \simeq f'(x)$ i ja tenim la fita que cercàvem: $|e_a(y)| \simeq |f'(x)||e_a(x)| \leq |f'(x)|\varepsilon_a(x)$. □

EXEMPLE. Volem calcular l'expressió $\ln(1 - 1/e)$ partint d'una aproximació del nombre e, i dubtem entre calcular-la tal i com està escrita o bé fent servir l'expressió equivalent $\ln(e - 1) - 1$. Considerant la fórmula de propagació de l'error amb les funcions $f_1(x) = \ln(1 - 1/x)$ i $f_2(x) = \ln(x - 1) - 1$, en calculem les seves derivades:

$$f_1'(x) = \frac{1}{x(x-1)} \quad , \quad f_2'(x) = \frac{1}{x-1},$$

i com $f_1'(x) < f_2'(x)$ per a x proper a e és més convenient emprar la primera expressió.

Per al problema numèric més general, que consisteix a calcular un resultat $y = f(x_1, \ldots, x_n)$ a partir d'unes dades inicials $x_1, \ldots, x_n$ disposem de la següent fórmula:

Teorema. *Sigui $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ una funció diferenciable, $x = (x_1, \ldots, x_n)$ unes dades inicials afectades d'error, i $y = f(x) = f(x_1, \ldots, x_n)$. Aleshores tenim la següent fita aproximada de propagació de l'error:*

$$\varepsilon_a(y) \simeq \sum_{i=1}^{n} \left| \frac{\partial f(x)}{\partial x_i} \right| \varepsilon_a(x_i).$$

DEMOSTRACIÓ. Posant $e_a(x_i) = x_i - \bar{x}_i$ usem el desenvolupament de Taylor de f al voltant de $x = (x_1, \ldots, x_n)$:

$$-e_a(y) = \bar{y} - y = f(\bar{x}) - f(x) = \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}\, e_a(x_i) + \mathcal{O}(e_a^2),$$

on $\mathcal{O}(e_a^2)$ són sumes de productes de la forma $e_a(x_i)e_a(x_j)$ que podem despreciar, i aleshores obtenim la fita aproximada

$$|e_a(y)| \simeq \left| \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}\, e_a(x_i) \right| \leq \sum_{i=1}^{n} \left| \frac{\partial f(x)}{\partial x_i} \right| \varepsilon_a(x_i). \qquad \square$$

2.2 Propagació dels errors en els càlculs

Partint de dades inicials exactes, per culpa d'acumulacions d'errors en les operacions obtenim un resultat afectat d'error. Per estudiar-lo apliquem el mètode de l'*anàlisi de l'error cap enrere*, que consisteix a trobar les modificacions que caldria fer a les dades d'entrada de forma que, suposant que no hi haguessin errors en les operacions, s'obtingués el mateix resultat.

Aquest estudi es fa emprant successivament la fórmula $\mathrm{fl}(x \star y) = (x \star y)(1 + \delta_\star)$, on $\star$ representa una operació aritmètica elemental i $\delta_\star$ és l'error relatiu en l'operació $\star$ que suposarem fitat, $|\delta_\star| \leq \varepsilon_\star$. De la mateixa manera, si f és una funció que intervé en els càlculs, tindrem $\mathrm{fl}(f(x)) = f(x)(1 + \delta_f)$, amb $|\delta_f| \leq \varepsilon_f$.

Finalment s'escriu una expressió del resultat final que permeti imputar els errors dels càlculs a les dades inicials, i un cop fet això podem aplicar les fórmules de propagació de l'error vistes en el cas anterior, considerant que els càlculs ja es fan sense error. La següent figura mostra gràficament el procés.

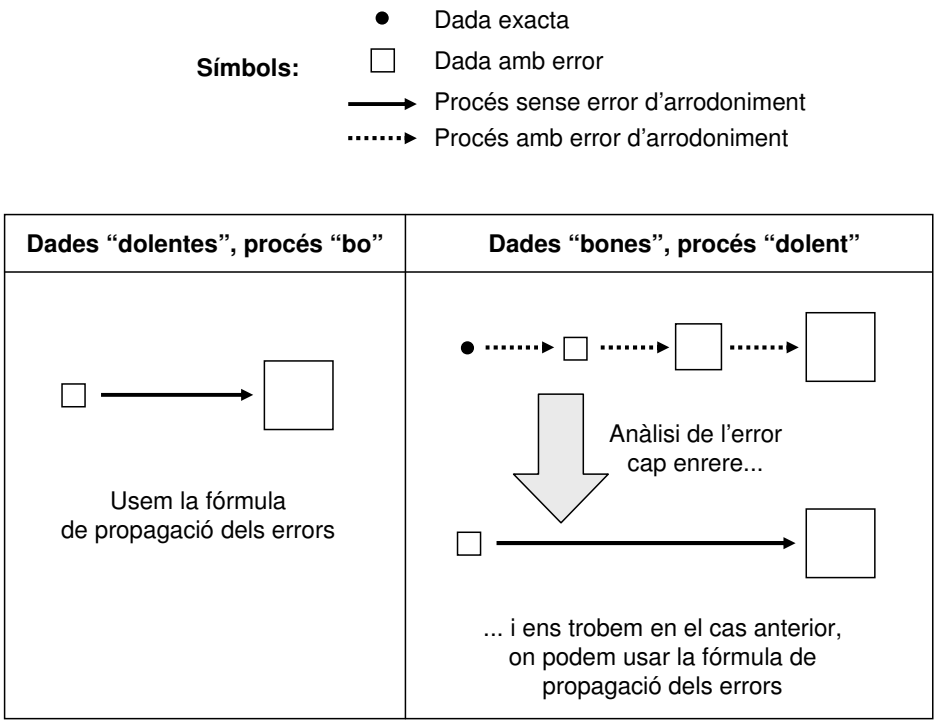

Figura 1: Quadre resum del tractament d'errors.

2.3 Errors de truncament

L'error de truncament depèn del mètode numèric emprat, i el seu estudi es fa de manera específica per a cada mètode. Per exemple, la fórmula dels trapezis d'integració numèrica té una fita de l'error diferent de la de la fórmula de Simpson. És una part fonamental de l'exposició de cada mètode l'estudi i la fita d'aquests errors.

3 Tractament estadístic de l'error

La fórmula de propagació d'errors és només una fórmula de primer ordre per a la fitació de l'error. El seu ús ens proporciona fites totalment rigoroses però sovint desproporcionades i poc realistes quan el nombre de càlculs és gran. Per a estimacions més properes a la realitat molts cops s'analitza l'error des d'un punt de vista *estadístic*, suposant que els errors en les dades d'entrada són *variables aleatòries independents* amb una certa *funció de densitat* donada.

Per exemple, sigui e l'error en arrodonir un nombre a d xifres decimals. Aleshores e pren valors sobre $[-\varepsilon, \varepsilon]$, on $\varepsilon = \frac{1}{2}10^{1-d}$. Si suposem que cada valor sobre aquest interval és igualment probable, la *funció densitat* ρ de e correspon a una distribució uniforme (això és, una funció que val $\frac{1}{2\varepsilon}$ sobre $[-\varepsilon, \varepsilon]$ i 0 fora d'aquest interval). La seva *funció distribució* ve donada per

$$F(x) = \int_{-\infty}^{x} \rho(t)dt$$

i ens dóna la probabilitat que e prengui valors més petits o iguals que x. Es tracta d'una recta de pendent $\frac{1}{2\varepsilon}$ sobre $[-\varepsilon, \varepsilon]$. La següent figura mostra el gràfic d'ambdúes funcions:

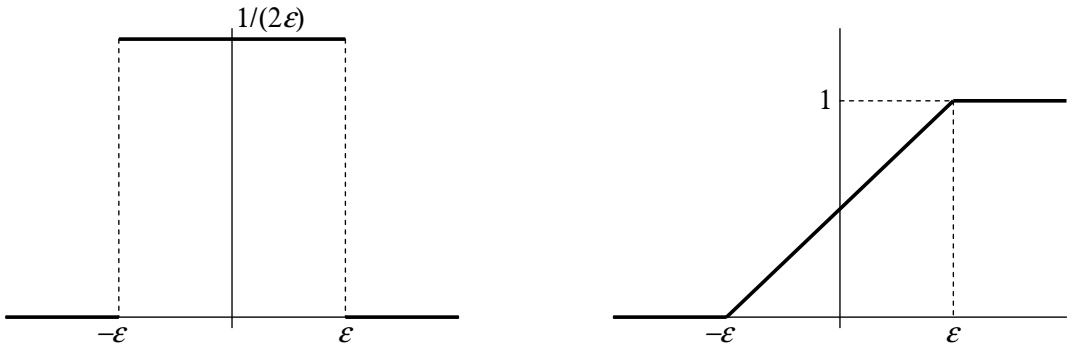

Figura 2: Funcions densitat i distribució de la distribució uniforme.

Llavors tenim les següents magnituds estadístiques de e (mitjana μ, variància σ^2 i desviació estàndard σ):

$$\mu = \int_{-\infty}^{+\infty} x\rho(x)\,dx = 0 \;\Rightarrow\; \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2\rho(x)\,dx = \frac{\varepsilon^2}{3} \;\Rightarrow\; \sigma = \frac{\varepsilon}{\sqrt{3}}.$$

El que és més habitual trobar, però, a la pràctica, són variables aleatòries e amb una *distribució normal*, això és, amb una funció densitat de probabilitat igual a

$$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

on μ és la mitjana i σ^2 la variància. La següent figura en mostra una representació gràfica:

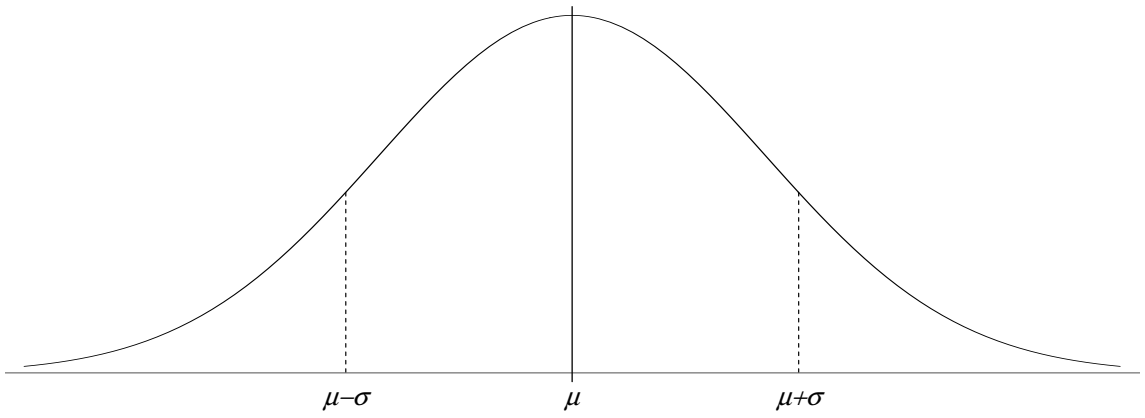

Figura 3: Funció densitat de la distribució normal.

La probabilitat $P(\delta)$ que e prengui valors entre $\mu - \delta$ i $\mu + \delta$ és llavors

$$P(\delta) = \int_{\mu-\delta}^{\mu+\delta} \rho(t)\,dt = \frac{2}{\sqrt{\pi}} \int_0^{\frac{\delta}{\sqrt{2}\sigma}} e^{-t^2} dt.$$

Per posar uns exemples, $Q(\delta) = 1 - P(\delta)$ val aproximadament 0.32, 0.045 i 0.0027 quan δ val σ, 2σ i 3σ respectivament, cosa que significa que si e representa l'error d'una dada amb mitjana $\mu = 0$ (per exemple) i desviació típica σ, l'afirmació que la magnitud de l'error sigui inferior a σ, 2σ i 3σ és falsa només en el 32%, 5% i 0.3% dels casos, respectivament. Això ens mostra com podem aconseguir estimar l'error d'una forma molt més realista.