Estatística e tratamento de dados

Fenómenos aleatórios, como lançar um dado, ou o número de desintegrações duma fonte radioactiva em cada unidade de tempo, flutuam de ensaio para ensaio. São caracterizados por uma variável aleatória x, que pode tomar um conjunto de valores discretos ou contínuos e por uma distribuição de frequências na ocorrência de cada valor possível de x, $P(x)$.

Por exemplo, no caso do dado x é uma variável aleatória discreta tal que $P(x) = 1/6$ é a distribuição de frequências de x.

Se x for contínua, a distribuição $P(x)$ será uma densidade de probabilidade, tal que a probabilidade de x se encontrar entre x e $x + dx$ é $P(x)\,dx$.

Como x deve tomar um dos seus possíveis valores, a soma das frequências (ou o integral da densidade de probabilidade) estendido a todo o seu domínio dá a unidade:

variável discreta

$$\sum_{i=1}^{N} P(x_i)$$

variável contínua

$$\int_{x_{min}}^{x_{max}} P(x)\,dx$$

Para caracterizar uma distribuição completamente seria preciso um grande número de ensaios ($N \to \infty$).

→ ver fig.

Em vez disso podem avaliar-se certos parâmetros que caracterizam uma distribuição, como sejam a sua **média** e a sua **dispersão** (ou largura), utilizando N ensaios.

• Média

$$\bar{x} = \sum_{i=1}^{N} x_i \, P(x_i) \qquad \text{ou} \qquad \bar{x} = \int_{x_{min}}^{x_{max}} x \, P(x) \, dx$$

É o melhor estimador do verdadeiro valor central da distribuição, X: $\bar{x} \to X$ quando $N \to \infty$.

Se calcularmos o valor médio dos desvios Δx em relação à média, obtemos:

$$\Delta x \equiv x - \bar{x}$$

$$\Rightarrow \quad \overline{\Delta x} = \overline{x - \bar{x}} = \bar{x} - \bar{x} = 0$$

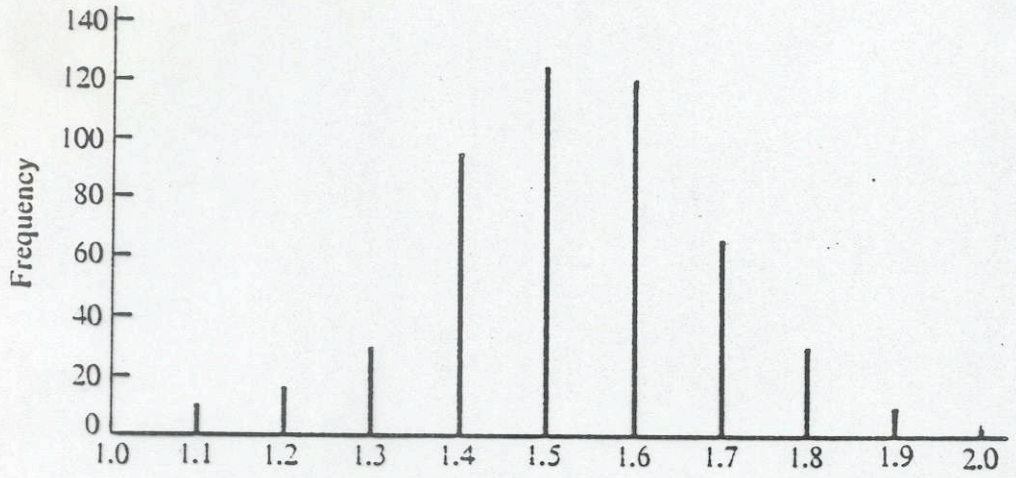
Ou seja, o desvio médio é nulo.

• Desvio padrão

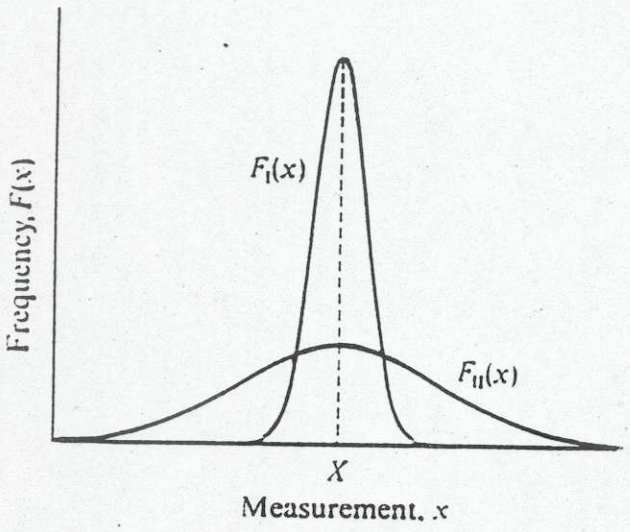
A largura duma distribuição não pode, pois, ser caracterizada pelo desvio médio. Mas se definirmos um desvio quadrático médio já é uma quantidade positiva:

$$\sigma^2 \equiv \overline{(\Delta x)^2} = \sum_{i=1}^{N} (x_i - \bar{x})^2 \, P(x_i) \geq 0$$

→ ver fig.



Amostra de distribuição discreta



Distribuições contínuas, com diferentes dispersões, normalizadas.

Note-se que temos

$$\sigma^2 \equiv \overline{(x-\bar{x})^2} = \overline{x^2 - 2x\bar{x} - \bar{x}^2} = \overline{x^2} - 2\bar{x}\bar{x} + \bar{x}^2$$

$$= \overline{x^2} - \bar{x}^2 \geqslant 0$$

⇒ Quanto mais disseminados forem os valores de x_i, maior a dispersão da distribuição.

Para conhecermos com exactidão a distribuição seria preciso conhecermos, por exemplo, todos os desvios da forma $(\Delta x)^n$. Na prática é muita vezes suficiente conhecer-se a média $\bar{x}$ e o desvio padrão σ ($\sigma = \sqrt{\text{variância}} \equiv \sqrt{\sigma^2}$).

Distribuições de probabilidade mais usadas

• Distribuição binomial

É usada em situações em que cada ensaio independente, repetido N vezes, só tem 2 possibilidades (sim ou não, cara ou coroa, dentro ou fora ...)

- p é a probabilidade de cada ensaio ter sucesso
- $q \equiv 1-p$ probabilidade de cada ensaio falhar

⇒ a probabilidade de uma dada sequência de r sucessos e $N-r$ falhanços será:

$$\underbrace{p \cdots p}_{r \text{ factores}} \underbrace{q \cdots q}_{N-r \text{ factores}} = p^r (1-p)^{N-r}$$

Mas há muitas maneiras diferentes de se obterem r sucessos e $N-r$ falhanços com N ensaios:

$$^N C_r = {}^N C_{N-r}$$

$$\therefore \quad P(r) = {}^{N}C_{r}\, p^{r}(1-p)^{N-r} = \frac{N!}{r!(N-r)!}\, p^{r}(1-p)^{N-r}$$

- Repare-se que a distribuição $P(r)$ está normalizada à unidade (binómio de Newton):

$$\sum_{r=0}^{N} P(r) = \sum_{r=0}^{N} {}^{N}C_{r}\, p^{r}(1-p)^{N-r} = [p + (1-p)]^{N} = 1^{N} = 1$$

- A sua média, ou seja, o valor médio dos sucessos vale:

$$\bar{r} = \sum_{r=0}^{N} r\, P(r) = N p \quad ,$$

o que é intuitivo dado termos N ensaios independentes, cada qual com a probabilidade p de sucesso. → ver fig.

Quer dizer, da observação de uma distribuição de N ensaios com valor médio $\bar{r}$ podemos inferir a sua probabilidade $p = \bar{r} / N$.

- A sua variância vale:

$$\sigma^{2} \equiv \sum_{r=0}^{N} (r-\bar{r})^{2} P(r) \equiv \bar{r^{2}} - \bar{r}^{2} = N p (1-p)$$

Relacionando média e variâucia vem:

$$\sigma^{2} = N p (1-p) = \bar{r}(1-p)$$

$$\therefore \quad \sigma^{2} \leq \bar{r}$$ → ver fig.

O desvio padrão é $\sigma \equiv \sqrt{\sigma^{2}} = \sqrt{N p (1-p)}$

$\Rightarrow$ A largura relativa da distribuição será:

$$\frac{\sigma}{\bar{r}} \equiv \frac{\sqrt{N p (1-p)}}{N p} = \sqrt{\frac{1-p}{p}} \cdot \frac{1}{\sqrt{N}}$$

No caso importante $p = 1-p = 1/2$ tem-se $\frac{\sigma}{\bar{r}} = \frac{1}{\sqrt{N}}$

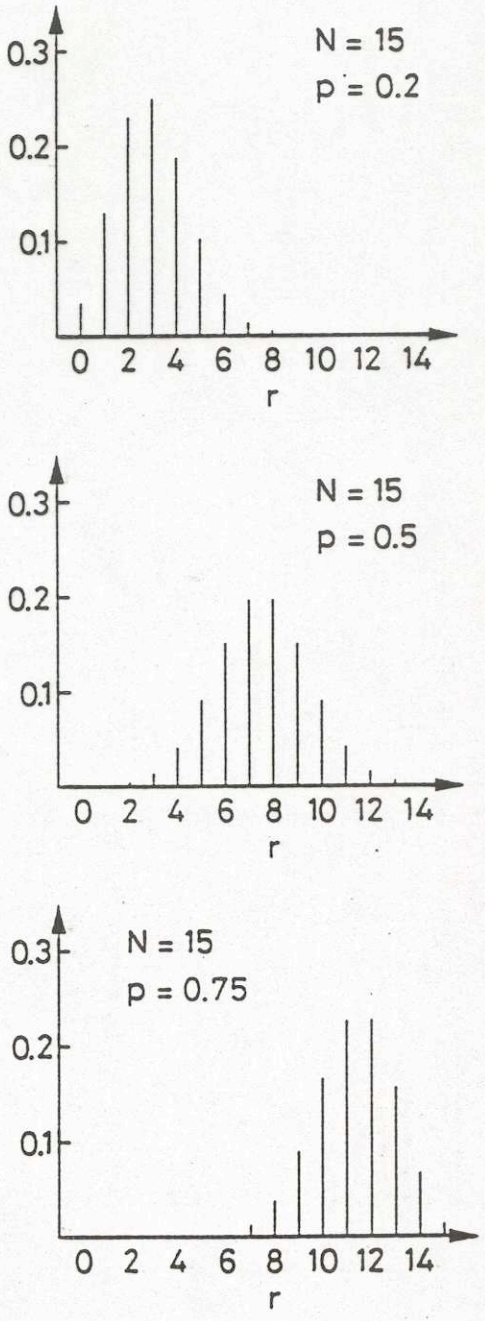


Fig. 4.1. Binomial distribution for various values of N and p

$\therefore \quad \bar{r} \uparrow\uparrow N \; ; \quad \sigma \uparrow\uparrow \sqrt{N} \; ; \quad \sigma/\bar{r} \searrow\searrow 1/\sqrt{N} \; .$

Nota sobre cálculo de $\bar{x}$ e σ

- Para demonstrar que

$$\bar{r} = \sum_r r\, P(r) = \sum_r r\, {}^{N}C_r\, p^r (1-p)^{N-r} = Np$$

faça-se $r\, p^r = p \frac{\partial}{\partial p}(p^r)$ e aplique-se a fórmula do binómio.

- Para provar que $\sigma^2 = Np(1-p)$, sabendo já que $\bar{r}^2 = N^2 p^2$, faça-se para o cálculo do outro termo: $\sum_r r^2\, {}^{N}C_r\, p^r (1-p)^{N-r}$, a substituição $r^2 p^r = r\left(p \frac{\partial}{\partial p}\right)(p^r) = \left(p \frac{\partial}{\partial p}\right)^2 (p^r)$.

- Para grandes valores de N a distribuição binomial é de difícil tratamento, pelo que é frequente serem usadas as seguintes distribuições limites:
  - $N \to \infty$ e $p \to 0$ com $\mu = Np =$ constante: distr. Binomial → distr. Poisson
  - $N \to \infty$ e $p =$ constante: distr. Binomial → distr. Gauss (ou normal)

Distribuição de Poisson

A probabilidade de se obterem r eventos se em média se obtém μ é:

$$P(r) = \frac{\mu^r}{r!} e^{-\mu}$$

Trata-se de uma distribuição discreta definida

para $r = \{0, \infty\}$. Exemplo típico é o decaimento radioactivo: a probabilidade de um núcleo não decair (sobreviver) num período de tempo t é

$$P(r=0) = \frac{\mu^0}{0!} e^{-\mu} = e^{-\lambda t},$$

com $\mu = \lambda t$ e em que $\lambda \ll$. Como há muitos núcleos, mesmo numa amostra ínfima, vem $\mu \equiv Np = N\lambda t$ finito, isto é, $\mu \not\to 0$ apesar de $\lambda \ll$.

A probabilidade de um ou mais núcleos decaírem é

$$P(r \geq 1) = P(1) + P(2) + \dots = 1 - P(0) = 1 - e^{-\lambda t}$$

Como só a média μ aparece na expressão da distr. Poisson, o conhecimento de N ou p não é necessário.

$\Rightarrow$ caso normal em processos radioactivos ou em colisões com feixes de partículas, onde as taxas de contagem são mais fàcilmente obtidas que o número de núcleos ou de partículas do feixe.

- **Nota sobre limite Binomial → Poisson**

  - $\lim_{N\to\infty} \frac{N!}{(N-r)!} = \lim_{N\to\infty} N(N-1)(N-2)\cdots(N-r+1) \simeq N^r$

  - $\lim_{N\to\infty} (1-p)^{N-r} = \lim_{N\to\infty} \sum_{K=0}^{N-r} \frac{(N-r)!}{K!(N-r-K)!} (-p)^K$

$$= \sum_{K=0}^{\infty} \frac{(N-r)^K}{K!} (-p)^K \simeq \sum_{K=0}^{\infty} \frac{(-Np)^K}{K!} = e^{-Np}$$

$$\therefore P(r) = \lim_{\substack{p\to 0 \\ N\to\infty}} \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r} = \frac{N^r p^r}{r!} e^{-Np} = \frac{\mu^r}{r!} e^{-\mu}$$

- Norma: $\sum_r P(r) = \sum_r \frac{\mu^r}{r!} e^{-\mu} = e^{-\mu} \sum_r \frac{\mu^r}{r!} = e^{-\mu} e^{\mu} = 1$

- Média: $\bar{r} \equiv \sum_r r P(r) = \sum_r r \frac{\mu^r}{r!} e^{-\mu} = e^{-\mu} \sum_r \frac{\mu^r}{(r-1)!}$

$$= e^{-\mu} \mu \cdot \sum_r \frac{\mu^{r-1}}{(r-1)!} = e^{-\mu} \mu e^{\mu} = \mu$$

- Variância: $\sigma^2 \equiv \sum_r (r - \bar{r})^2 P(r) = \overline{r^2} - \bar{r}^2 = (\mu^2 + \mu) - \mu^2 = \mu$

$\therefore$ desvio padrão: $\sigma = \sqrt{\mu}$

$\Rightarrow$ Esta é a base para os valores da estatística de contagens e suas flutuações:

$n \pm \sqrt{n}$ → ver fig.

A distr. de Poisson não é simétrica

$\Rightarrow$ valor máximo $\neq$ média.

Mas, à medida que $\mu >>$, a distr. Poisson torna-se cada vez mais simétrica e tende para a distr. de Gauss.

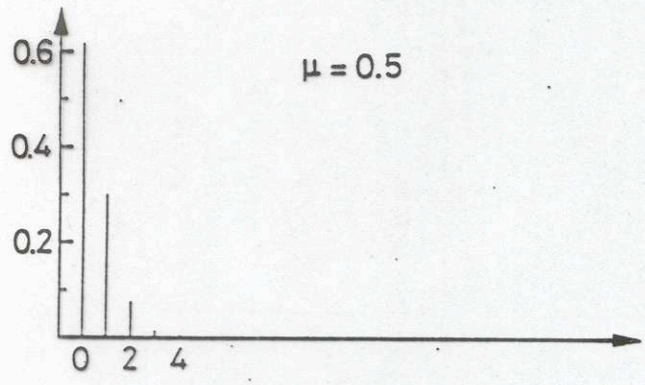
• Distribuição de Gauss (ou normal)

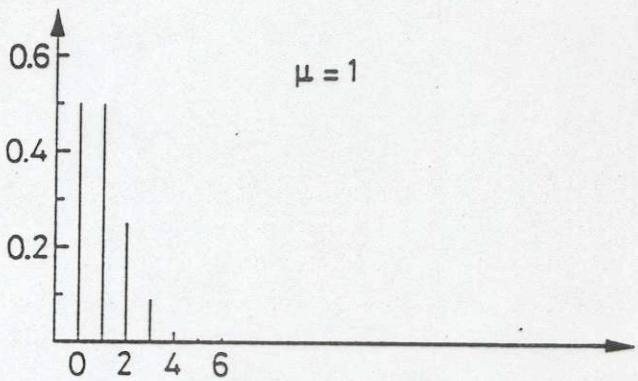
Quando $N \to \infty$ com p constante, quer dizer, quando $Np \to \infty$ (na verdade $\mu \equiv Np \sim 10$ já basta!) estamos no limite da distr. Gauss:

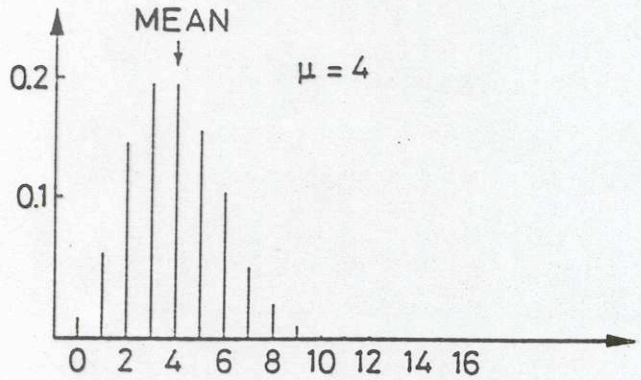
$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp -\frac{(x-\mu)^2}{2\sigma^2} .$$

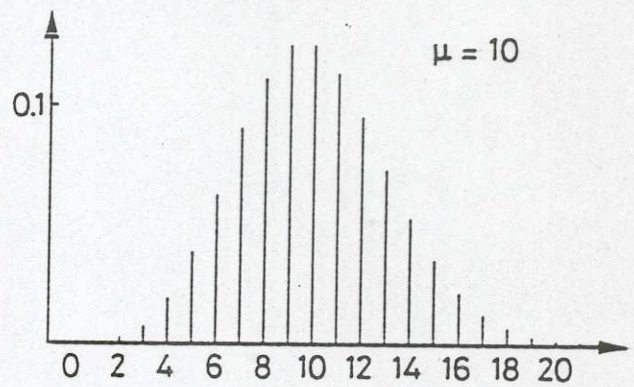
É uma distribuição contínua com $-\infty < x < \infty$, simétrica em relação ao valor médio μ. σ é o desvio padrão

A maior parte dos erros instrumentais têm









Distribuições de Poisson
com médias μ crescentes:

distrib. Poisson $\xrightarrow[\mu >>]{}$ distrib. Gauss

distribuição gaussiana. Na medida de comprimentos, tempos, temperaturas, tensões, correntes, etc, os ensaios seguem a distr. normal.

Na estatística de contagens, além do desvio padrão σ, usa-se a largura do pico a meia altura (FWHM = full width half maximum), tal que:

$$FWHM = 2.35\,\sigma$$ → ver figs.

O significado estatístico de σ é tal que o integral da densidade de probabilidade $P(x)$ entre:

$\bar{x} - \sigma \longleftrightarrow \bar{x} + \sigma$: 68.3%

$\bar{x} - 2\sigma \longleftrightarrow \bar{x} + 2\sigma$: 95.5% → ver figs.

$\bar{x} - 3\sigma \longleftrightarrow \bar{x} + 3\sigma$: 99.7%

Ou seja, se um resultado é dado na forma $x \pm \sigma$, isso quer dizer que em 100 novos ensaios da variável x só 68 deverão cair dentro das barras de erro $\Rightarrow$ $\sim 1/3$ cai fora.

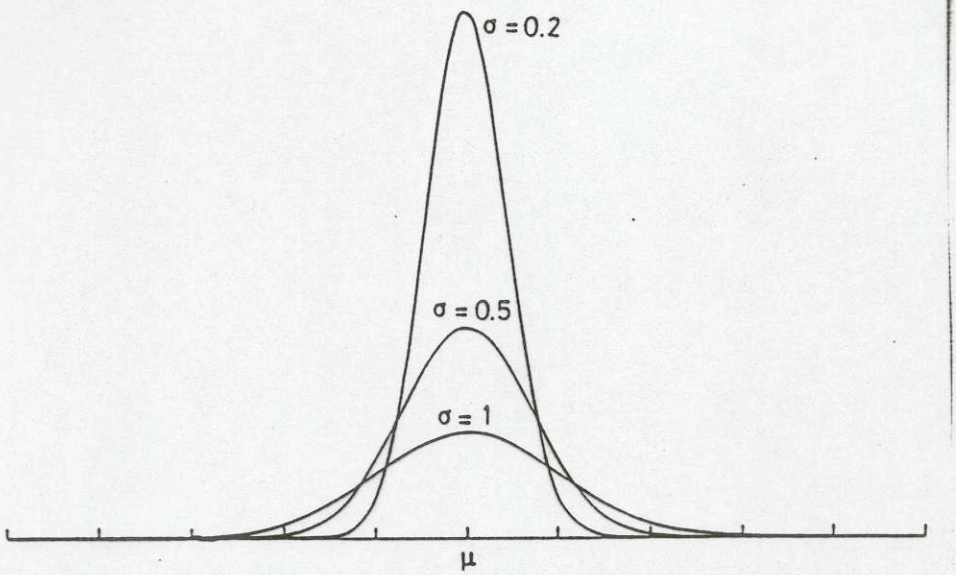


Fig. 4.3. The Gaussian distribution for various σ. The standard deviation determines the width of the distribution

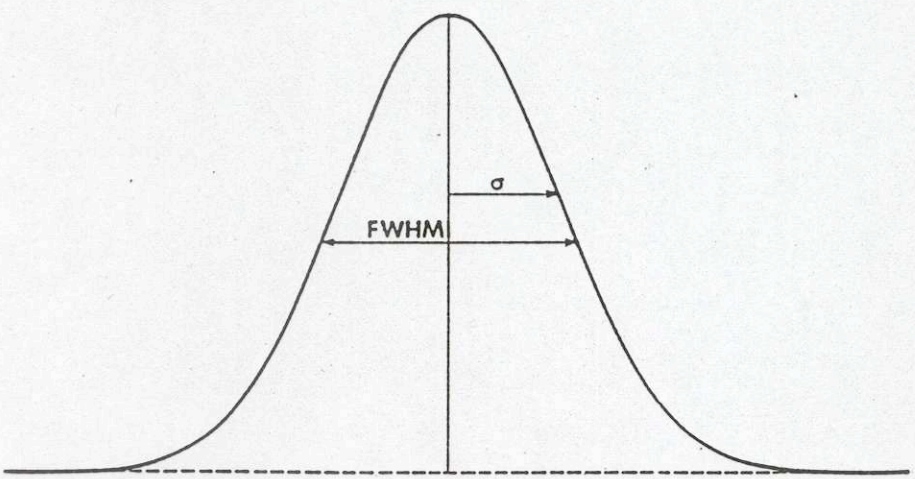
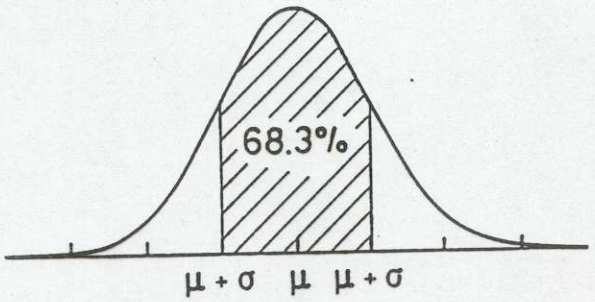
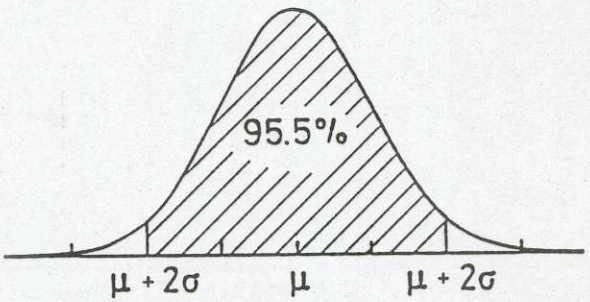
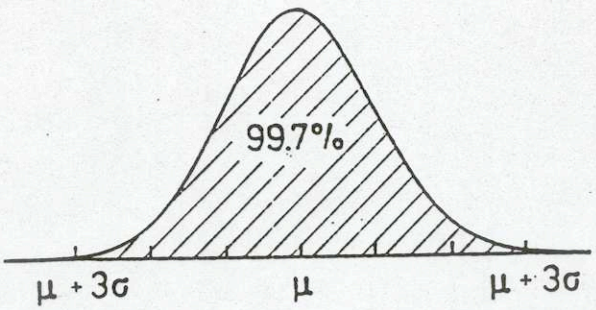


Fig. 4.4. Relation between the standard deviation σ and the full width at half-maximum (FWHM)







Distribuição de Gauss: significado estatístico do integral de $\Pi(x)$ em torno da média μ para os limites $\mu \pm \sigma$, $\mu \pm 2\sigma$ e $\mu \pm 3\sigma$.

• Distribuição do χ^2

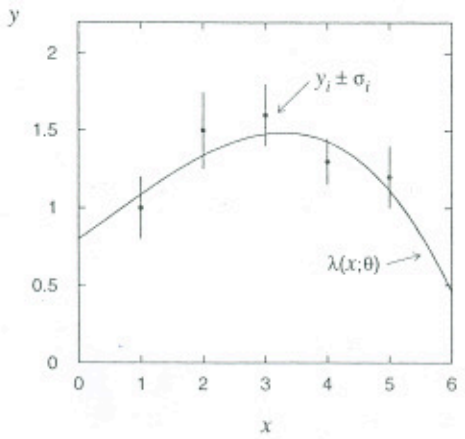
Quando se compara uma fórmula teórica $y = f(x)$ com dados experimentais costumam usar-se estimadores da qualidade do ajuste.

Um deles é o χ^2:

$$\chi^2 \equiv \sum_{i=1}^{n} \frac{(y_i - y_{teorico})^2}{\sigma_i^2}$$

Quer dizer, é a diferença quadrática entre valores experimentais e teóricos, para cada ponto experimental (n pontos ao todo), normalizada ao respectivo erro experimental e somada para todos.

Quanto menor for o χ^2, melhor o ajuste. Mas, devido ao significado estatístico de σ, espera-se: $\chi^2/\nu \sim 1$ (ν = nº graus de liberdade = n − nº parâmetros da eqº ajuste)



Ingredients of the least squares problem: N values $y_1, \ldots, y_N$ are measured with errors $\sigma_1, \ldots, \sigma_N$ at the values of x given without error by $x_1, \ldots, x_N$. The true value λ_i of y_i is assumed to be given by a function $\lambda_i = \lambda(x_i;\theta)$. The value of θ is adjusted to minimize the value of χ^2 given by equation (7.3).

- **Aplicação ao ajuste dos mínimos quadráticos linear**

Se a expressão teórica se puder exprimir na forma $y_{teor} = a x + b$, o estimador χ^2 escreve-se:

$$\chi^2 = \sum_i^n \frac{(y_i - a x_i - b)^2}{\sigma_i^2}$$

Pretende minimizar-se a função χ^2 em ordem aos coeficientes $\underline{a}$, e $\underline{b}$, simultâneamente:

$$\begin{cases} \frac{\partial \chi^2}{\partial a} = \text{- - -} = 0 \\ \frac{\partial \chi^2}{\partial b} = \text{- - -} = 0 \end{cases} \Rightarrow$$

de modo a extrair-se os valores de $\underline{a}$ e $\underline{b}$ e estimarem-se os erros respectivos quadràticamente

Calculando, obtém-se:

$$\begin{cases} a = \frac{c_1 c_5 - c_3 c_4}{\Delta} \\ b = \frac{c_2 c_4 - c_1 c_3}{\Delta} \end{cases} , \quad \begin{cases} \sigma_a^2 = \frac{c_5}{\Delta} \\ \sigma_b^2 = \frac{c_2}{\Delta} \end{cases}$$

com: $\Delta = c_2 c_5 - c_3^2$ e:

$$\begin{cases} c_1 = \sum_1^n \frac{x_i y_i}{\sigma_i^2} \\ c_2 = \sum_1^n \frac{x_i^2}{\sigma_i^2} \\ c_3 = \sum_1^n \frac{x_i}{\sigma_i^2} \\ c_4 = \sum_1^n \frac{y_i}{\sigma_i^2} \end{cases} \qquad c_5 = \sum_1^n \frac{1}{\sigma_i^2}$$

Análise dimensional

Ex.: $[x] = [y] = L = [\sigma_x] = [\sigma_y]$

$\therefore [b] = L \; ; \; [a] = 1$

Logo: $[\sigma_b^2] = \frac{[c_2]}{[\Delta]}$

$L^2 = \frac{1}{L^{-2}}$

Estimação dos parâmetros de uma distribuição

Para determinar os parâmetros duma distribuição desconhecida faz-se uma amostragem, isto é, obtém-se um conjunto de dados representativo.

Como já se disse, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, média da amostra, é o melhor estimador da verdadeira média X da distribuição.

Mas, como a amostra é finita, se efectuarmos outra amostragem obteremos um $\bar{x}$ diferente. Pretende saber-se qual a precisão de $\bar{x}$, ou seja, qual a sua variância:

$$\sigma^2(\bar{x}) = \overline{(\bar{x} - X)^2}$$

Ora $\overline{(\bar{x}-X)^2} = \overline{\left(\frac{1}{n}\sum_i^n x_i - X\right)^2} = \frac{1}{n^2}\overline{\left(\sum_i^n x_i - nX\right)^2} =$

$$= \frac{1}{n^2}\overline{\left[\sum_i^n (x_i - X)\right]^2}$$

Como $\left[\sum_i (x_i - X)\right]^2 = \sum_i (x_i - X)^2 + \sum_i \sum_{j \neq i} (x_i - X)(x_j - X)$,

vem:

$$\sigma^2(\bar{x}) = \frac{1}{n^2}\overline{\sum_i (x_i - X)^2} = \frac{1}{n^2}\sum_i \overline{(x_i - X)^2} = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n},$$

em que os termos cruzados têm valor médio nulo.

$\Rightarrow$ O desvio padrão do valor médio é:

$$\sigma(\bar{x}) = \sigma / \sqrt{n}$$

A precisão do valor médio $\bar{x}$ é a largura da distribuição dividida pelo tamanho da amostra:

$n \uparrow\uparrow \Rightarrow \bar{x}$ mais fiável

• Propagação de erros

Dada uma quantidade z calculável a partir de variáveis x e y directamente medidas, isto é, tal que $z = f(x, y)$, pretende calcular-se o erro associado a z, σ_z. Quanto ao valor médio de z, $\bar{z}$, ele é obtido directamente por aplicação da função f:

$$\bar{z} = f(\bar{x}, \bar{y}) \ .$$

O desvio de z em relação à média $z - \bar{z}$ é, em 1.ª ordem,

$$z - \bar{z} = (x - \bar{x}) \left(\frac{\partial f}{\partial x}\right)_{\bar{x}} + (y - \bar{y}) \left(\frac{\partial f}{\partial y}\right)_{\bar{y}} \ ,$$

em que as derivadas parciais são calculadas nos pontos médios respectivos.

Como $\sigma_z^2 = \overline{(z - \bar{z})^2}$, quadrando e extraindo a média, vem:

$$\overline{(z - \bar{z})^2} = \overline{(x - \bar{x})^2} \left(\frac{\partial f}{\partial x}\right)^2 + \overline{(y - \bar{y})^2} \left(\frac{\partial f}{\partial y}\right)^2 + 2 \overline{(x - \bar{x})(y - \bar{y})} \frac{\partial f}{\partial x} \frac{\partial f}{\partial y}$$

ou:

$$\sigma_z^2 = \left(\frac{\partial f}{\partial x}\right)_{\bar{x}}^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)_{\bar{y}}^2 \sigma_y^2 + 2 \, \mathrm{cov}(x, y) \left(\frac{\partial f}{\partial x}\right)_{\bar{x}} \left(\frac{\partial f}{\partial y}\right)_{\bar{y}} \ ,$$

em que a covariância de x e y, $\mathrm{cov}(x, y)$ é dada por:

$$\mathrm{cov}(x, y) = \overline{(x - \bar{x})(y - \bar{y})} \ ,$$

e é nula caso x e y sejam independentes. Caso sejam proporcionais, $\mathrm{cov}(x, y) = \sigma_x \sigma_y$.

Combinação de diferentes resultados experimentais

Se possuirmos N diferentes amostras $x_1, x_2, \dots, x_N$ da mesma distribuição, de desvios padrão $\sigma_1, \sigma_2, \dots, \sigma_N$ (os métodos experimentais podem ser diferentes), para obtermos a média das médias das amostras deveremos usar a

média ponderada : $$\bar{x} = \frac{\sum_i^N x_i/\sigma_i^2}{\sum_i^N 1/\sigma_i^2}$$

e a sua

variância : $$\sigma^2(\bar{x}) = \frac{1}{\sum_i^N 1/\sigma_i^2}$$

⇒ Deve dar-se menos peso às amostras com maiores dispersões (medidas com instrumentos menos precisos).

A variância $\sigma^2(\bar{x})$ obtém-se directamente da média ponderada, aplicando a esta a expressão quadrática de propagação dos erros.

Se todas as amostras têm a mesma dispersão σ, vem

$$\bar{x} = \frac{\sum_i^N x_i}{N}$$

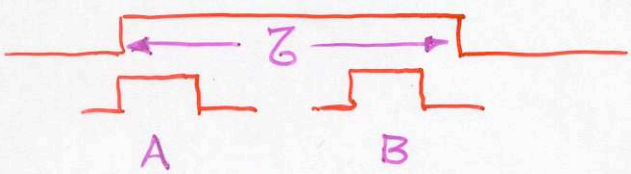
$$\sigma^2(\bar{x}) = \frac{\sigma^2}{N} \quad \text{ou} \quad \sigma(\bar{x}) = \frac{\sigma}{\sqrt{N}}$$

Aplicação da distribuição de Poisson: coincidências fortuitas

Sejam N_A e N_B o nº de eventos de dois sinais aleatórios e independentes observados no tempo de aquisição T.

Pretende obter-se no tempo T o nº de sinais em coincidência $N_C = N_A * N_B$.

A probabilidade de 1 ou mais eventos de A (ou B) estarem contidos num intervalo de tempo τ (porta electrónica da unidade de coincidências) é:



$$P_A(r \geq 1) = \sum_{r=1}^{\infty} P_A(r) = 1 - P_A(r=0)$$

$$= 1 - \frac{\mu^0}{0!} e^{-\mu} = 1 - e^{-\mu} \quad ,$$

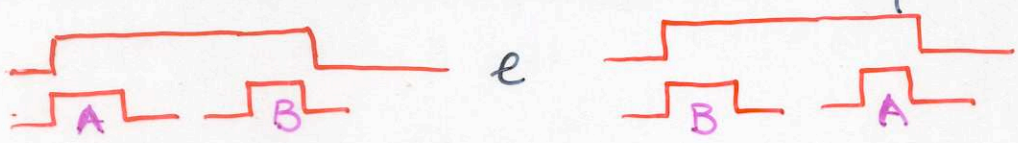
com $\mu = \lambda \tau \equiv$ média dos eventos em $\tau = \frac{N_A}{T} \tau$.

Como $\mu << 1$, vem $P_A(r \geq 1) \simeq 1 - (1-\mu) = \mu$.

A probabilidade conjunta em τ será o produto das probabilidades independentes:

$$P_C = 2 \cdot P_A(r \geq 1) \cdot P_B(r \geq 1) = 2 N_A N_B \tau^2 / T^2$$

Factor 2: Duas maneiras de abrir a porta electrónica:



Como $N_C = P_C \cdot T/\tau$, vem: $\boxed{N_C = 2 N_A N_B \tau / T}$.