Estadística descriptiva

1. Conceptes generals
2. Representacions gràfiques
3. Mesures de tendència central
4. Mesures de dispersió
5. Mesures de posició
6. Covariància i correlació
7. Matriu de covariància

L'estadística és una part essencial en totes les ciències, ja que constitueix una ferramenta que permet el tractament de les incerteses inherents a totes les mesures experimentals i ajuda a extraure conclusions dels resultats. Des del punt de vista experimental és també una ferramenta imprescindible de disseny i planificació.

Aquest primer tema, però, està dedicat únicament a les diferents maneres de presentar les dades o el conjunt de resultats, sense fer cap anàlisi profunda ni cap inferència: és l'anomenada ***estadística descriptiva****.*

Bibliografia*:*

- *Problemes d'estadística amb aplicació a l'enginyeria* (Pujol, Gibergans, García)
- *Statistics. A guide to the use of Statistical Methods in the Physical Sciences* (Barlow)
- *An introduction to Error Analysis* (John R. Taylor)

Conceptes generals

- L'estadística és la ciència que tracta de l'**anàlisi de dades**.
- Aquestes dades provenen d'una **població** (o d'un subconjunt o **mostra**) que és un conjunt de referència sobre el qual es fan les observacions o mesures.
- Una població està formada per **individus**, que és la unitat estadística.
- S'estudien diferents caràcters de la població; un caràcter és una propietat inherent de l'individu.
- Es defineixen tres tipus de caràcters o variables estadístiques:
  - **Variable qualitativa o atribut**: no prenen valors numèrics sinó que descriuen qualitats (color dels ulls, sexe...).
  - **Variable quantitativa discreta**: pren únicament valors enters i correspon generalment a comptar el nombre de vegades que succeeix un esdeveniment (nombre de desintegracions, nombre naixements...).
  - **Variable quantitativa contínua**: pren valors en un interval i correspon a mesurar magnituds reals (el pes, la longitud...). Aquests valors no poden tenir una precisió infinita.

Conceptes generals

- Interval de classe
  - Cada un dels intervals en què es poden agrupar les dades d'una variable estadística.
  - En el cas de dades d'una variable contínua, l'agrupació en un mateix interval de diferents dades, encara que pròximes, implica una perduda de precisió.
  - Es denomina **marca de classe** el punt mitjà de l'interval.

Exemple 1: El resultat del llançament de 20 monedes a l'aire és:

{Head, Tail, H, H, T, H, T, H, H, H, T, T, H, T, T, H, T, H, H, T}

Resultats que poden agrupar-se en dos intervals H i T, i escriuríem {11H,9T}

Exemple 2: El resultat de mesurar 10 vegades la llargària d'una taula és, en cm,

{26.4, 23.9, 25.1, 24.6, 22.7, 23.8, 25.1, 23.9, 25.3, 25.4}

Resultats que poden agrupar-se en 6 intervals de la següent manera:

22 a 23	23 a 24	24 a 25	25 a 26	26 a 27	27 a 28
1	3	1	4	1	0

Conceptes generals

Exemple 2 amb MATLAB

Script

```
%INTERVALS DE CLASSE

%Introduïm les dades
x=[26.4 23.9 25.1 24.6 22.7 23.8 25.1 23.9 25.3 25.4];
%Definim els intervals (valor inicial del primer interval...
%:amplària:valor inicial de l'últim interval
binranges=22:1:27;
marca=binranges+0.5;
%Comptem els valors en cada un dels intervals
n=histc(x,binranges);
%Eixida dels resultats
z=[binranges;marca;n];
fprintf('Interval   Marca de classe  Freqüència\n');
fprintf('  %3.1f          %3.1f             %2d\n',z);
```

Resultat

Interval	Marca de classe	Freqüència
22.0	22.5	1
23.0	23.5	3
24.0	24.5	1
25.0	25.5	4
26.0	26.5	1
27.0	27.5	0

Límit inferior de l'interval

Conceptes generals

- Classes de freqüències: definicions
  - **Freqüència absoluta n_i**: S'anomena freqüència absoluta d'un valor d'una variable el nombre de vegades que es repeteix aquest valor.
    - Si N és el nombre total de dades i n_i el nombre d'elements de l'interval i-èsim, es verifica que

$$n_i \leq N \quad \text{i} \quad \sum n_i = N$$

  - **Freqüència relativa f_i**: És la relació existent entre la freqüència absoluta i el nombre total de dades.

$$f_i = \frac{n_i}{N}$$

    - La freqüència relativa verifica que $\sum f_i = 1$

  - Si es treballa amb intervals de classe, les freqüències absolutes i relatives es defineixen en cada interval com el nombre d'elements que hi pertanyen.

  - **Freqüència acumulada:** Si ordenem les dades en ordre creixent o decreixent, la freqüència acumulada d'un valor donat és la suma de freqüències fins aquest valor.

Representacions gràfiques

- La transcripció de les dades en una gràfica permet veure ràpidament el contingut de la taula estadística.
- Tipus de gràfics:
  - Diagrama de barres
  - Histograma
  - Polígon de freqüències
  - Gràfic de sectors

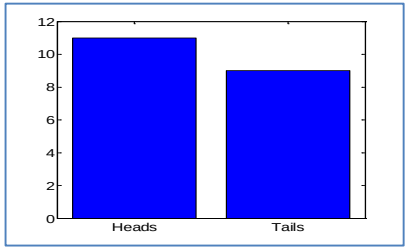
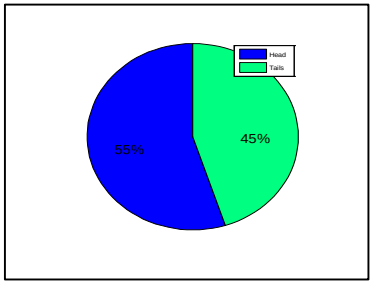


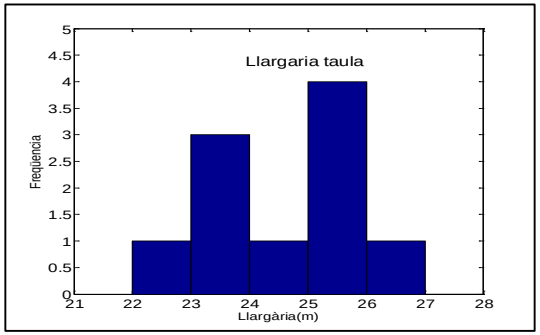
Diagrama de barres exemple 1. El número representat és proporcional a la longitud de la barra

```
>>x=[11,9];
>> bar(x)
```



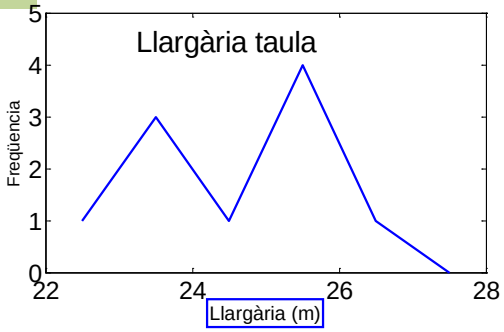
Gràfic de sectors exemple 1

```
>> x=[11 9];
>>pie(x),legend('Head','Tails')
```



Histograma de l'exemple 2. El número representat és proporcional a l'àrea de la barra.

```
>> x=[26.4 23.9 25.1 24.6 22.7 23.8 25.1 23.9 25.3 25.4];
>> hist(x,22.5:1:27.5)
>>axis([21 28 0 5])
```




```
>> x=[22.5:1:27.5];
>> y=[1 3 1 4 1 0];
>> plot(x,y)
>> axis([22 28 0 5])
```

Polígon de freqüències exemple 2.

Representacions gràfiques

- Quan treballem amb dades numèriques hem de triar l'amplària de l'interval, la qual cosa s'ha de meditar:
  - Si l'interval és molt menut, tindrem pocs valors en cada interval, i el nombre total estarà dominat per les fluctuacions estadístiques.
  - Si l'interval és massa gran, perdrem els detalls de les possibles variacions.
  - Si fóra possible, hauríem de tenir almenys 10 valors per interval.

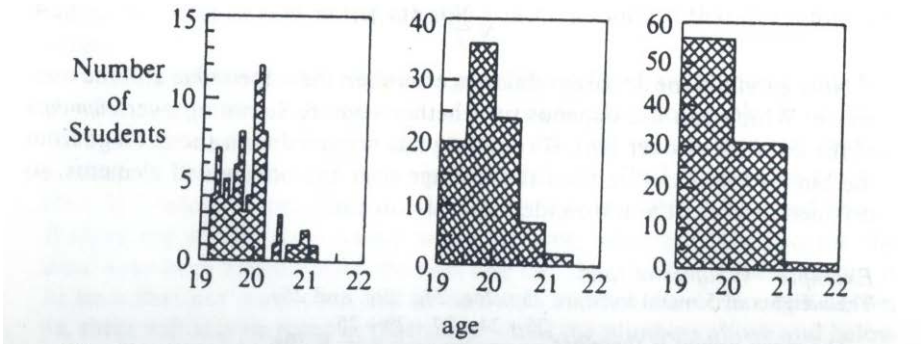


Fig. 2.2. The ages (in years) of a group of second year students, showing the effects of choosing different bin sizes for the same data.

Mesures de tendència central

- Les mesures de centralització tenen per finalitat caracteritzar la informació continguda en la taula o sèrie estadística amb un número.

- Si tenim N valors {x_1, x_2, ..., x_N}, definim la **mitjana aritmètica** $\overline{x}$ d'aquesta sèrie com la suma de tots els valors dividida pel seu nombre total:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{(mitjana aritmètica)}$$

>> mean(x)

- **Mitjana aritmètica ponderada:**
  - W_i representa el pes de la variable x_i.

$$\overline{x} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_N x_N}{w_1 + w_2 + \ldots + w_N} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

- De la mateixa manera definim el **valor mitjà** de qualsevol funció *f(x)* com a:

$$\overline{f} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

Mesures de tendència central

- Per a taules de freqüències, suposant que hi ha *k* valors diferents, es calcula com

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{k} n_j x_j = \sum_{j=1}^{k} f_j x_j \qquad \overline{f} = \frac{1}{N}\sum_{j=1}^{k} n_j f\left(x_j\right) = \sum_{j=1}^{k} f_j f\left(x_j\right)$$

- Si es treballa directament amb les dades agrupades en intervals de classe, es pren x_j igual a la marca de classe. Hem de notar que si utilitzem aquesta última expressió, com que estem substituint els valors reals per les marques de classe, perdem precisió.
- **Desviació d_i** del valor x_i respecte de la mitjana: $d_i = x_i - \overline{x}$
- Propietat de la mitjana aritmètica $\sum_i d_i = \sum_{i=1}^{N}(x_i - \overline{x}) = 0$
  - Demostració:

$$\sum_{i=1}^{N}(x_i - \overline{x}) = (\sum_{i=1}^{N} x_i) - (\sum_{i=1}^{N} \overline{x}) = (N\ \overline{x}) - (\overline{x}\sum_{I=1}^{N} 1) = N\ \overline{x} - \overline{x}N = 0$$

Mesures de tendència central

- **Mitjana geomètrica G**: és l'arrel N-èsima del producte dels N valors x

$$G = \sqrt[N]{x_1\ x_2\ \ldots\ x_N}$$

Tenim un país on hi ha tres habitants, amb rendes 1, 3 i 9. Posem que la felicitat es mesura pel **logaritme** de la renda. Podem preguntar-nos ara quina renda igualitària donaria la mateixa felicitat que aquest repartiment tan desigual. Podem estar temptats a fer la mitjana aritmètica entre 1, 3 i 9 i dir que serà 4.33, però hauríem calculat la renda mitjana, no la felicitat mitjana. Proveu que la mitjana geomètrica, de les rendes és la renda que proporciona felicitat mitjana, en el nostre cas, G=3.

- **Mitjana harmònica H:** És l'invers de la mitjana aritmètica dels seus inversos.

$$H = \frac{N}{\frac{1}{x_1} + \ldots + \frac{1}{x_N}}$$

Si un vehicle es desplaça una certa distància a una velocitat v_1 (per exemple, 60 km/h) i després la mateixa distància de nou a una velocitat v_2 (per exemple, 40 km/h), llavors la seua velocitat mitjana és la mitjana harmònica de v_1 i v_2 (48 km/h).

- **Mitjana quadràtica:** És l'arrel quadrada de la mitjana aritmètica dels quadrats dels valors *x.*

$$C = \sqrt{\frac{x_1^{\,2} + \ldots + x_N^{\,2}}{N}} \quad \text{(Root mean squared o r.m.s.)}$$

- **Propietat:**

$$H \leq G \leq \overline{x} \leq C$$

Mesures de tendència central

- ***Exemple:*** Calculeu la mitjana aritmètica, geomètrica, harmònica i quadràtica del següent conjunt de dades obtingudes pesant cinc làmines metàl·liques:

  25, 24, 27, 29 i 25 (g)

Solució

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{5}\left(25+24+27+29+25\right)\text{g} = 26\text{ g}$$

$$G = \sqrt[n]{x_1\ x_2\ \ldots\ x_N} = \sqrt[5]{25\times 24\times 27\times 29\times 25} = 25.94\text{ g}$$

$$H = \frac{N}{\frac{1}{x_1}+\ldots+\frac{1}{x_N}} = \frac{5}{\frac{1}{25}+\frac{1}{24}+\frac{1}{27}+\frac{1}{29}+\frac{1}{25}} = 25.88\text{ g}$$

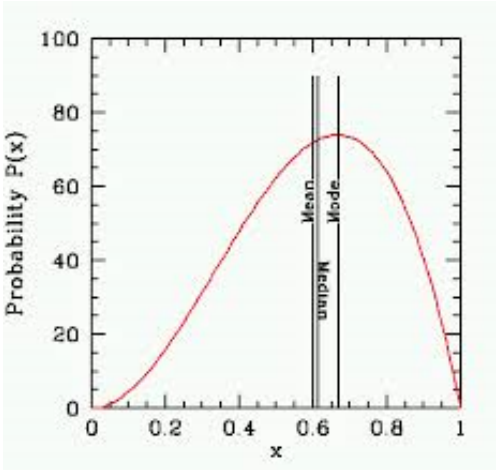
$$C = \sqrt{\overline{\left(x^2\right)}} = \sqrt{\frac{x_1^{\,2}+\ldots+x_N^{\,2}}{N}} = \sqrt{\frac{25^2+24^2+27^2+29^2+25^2}{5}} = 26.06\text{ g}$$

Mesures de tendència central

- **Mediana:** És el valor de l'element central de la sèrie estadística, ordenada en sentit creixent o decreixent. Es denota per *Me.* No més del 50% dels valors són menors a la mediana i no més del 50% dels valors són majors. Si N és parell, hi ha dos elements centrals. En aquest cas, la mediana és la mitjana d'aquests dos elements centrals.

- **Moda:** Es correspon al valor més freqüent de la sèrie estadística i es denota amb *Mo*.

  – En l'exemple anterior: *Me*= *Mo*= 25g



- Poden haver-hi diverses modes; aleshores la distribució de les dades s'anomena **distribució multimodal** (si la variable ve donada en intervals de classe, parlem d'interval modal, que correspon a intervals amb major freqüència).

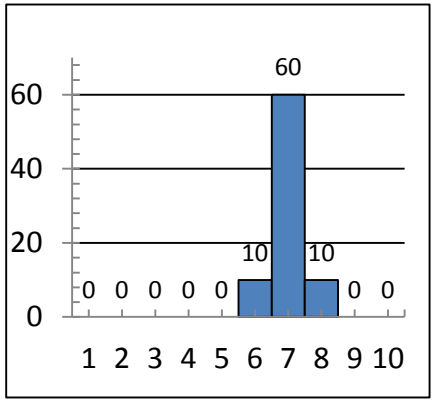
Mesures de dispersió

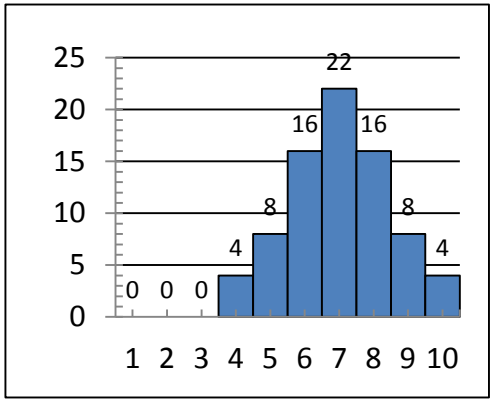
- Es defineix la **variància** d'un conjunt de dades x_i com la mitjana dels quadrats de les desviacions respecte de la mitjana aritmètica:

$$V(x) \equiv \frac{1}{N}\sum_i (x_i - \overline{x})^2$$

>> cov(x,1)

- Expressa *la dispersió* de les dades respecte de la mitjana (en unitats de x^2).





Histograma que mostra les notes obtingudes en dos exàmens diferents del mateix grup de 80 alumnes. Encara que la mitjana és 7.0 en els dos casos, les notes estan distribuïdes de formes molt diferents. En el primer cas la variància és 0.25 i en el segon és 2.1.

- **Variància d'una funció *f***

$$V(f) \equiv \frac{1}{N}\sum_i (f(x_i) - \overline{f})^2$$

Mesures de dispersió

- Propietat: la variància és la diferència entre el quadrat de la mitjana quadràtica (mitjana dels quadrats) i el quadrat de la mitjana aritmètica.

$$V(x) = \frac{1}{N}\sum_i x_i^{\,2} - \overline{x}^2 \equiv \overline{\left(x^2\right)} - \overline{x}^{\,2}$$

  - Demostració

$$V(x) = \frac{1}{N}\sum_i (x_i - \overline{x})^2 = \frac{1}{N}\sum_i (x_i^{\,2} - 2x_i\overline{x} + \overline{x}^2)$$

$$V(x) = \frac{1}{N}\sum_i x_i^{\,2} - \frac{1}{N}\sum_i 2x_i\overline{x} + \frac{1}{N}\sum_i \overline{x}^2$$

$$V(x) = \frac{1}{N}\sum_i x_i^{\,2} - \frac{1}{N}2\overline{x}\sum_i x_i + \frac{1}{N}\overline{x}^2\sum_i 1$$

$$V(x) = \overline{x^2} - 2\overline{x}^2 + \overline{x}^2 = \overline{x^2} - \overline{x}^2 = \frac{1}{N}\sum_i x_i^{\,2} - \left(\frac{1}{N}\sum_i x_i\right)^2$$

- Corol·lari: $V(ax+b) = a^2V(x)$

  Ajuda: Demostreu primer que si $y_i = ax_i + b$ tindrem $\overline{y} = a\overline{x} + b$ i després calculeu V(y) aplicant-hi la propietat anterior.

Mesures de dispersió

- **Desviació típica o estàndard:** És la mitjana quadràtica de les desviacions, és a dir, l'arrel positiva de la variància

$$\sigma = +\sqrt{Var(x)} = +\sqrt{\frac{1}{N}\sum_i (x_i - \overline{x})^2}$$

(Root mean squared deviation)

>> std(x,1)

$$\sigma = \sqrt{\overline{x^2} - \overline{x}^2} = \sqrt{\frac{1}{N}\sum_i x_i^{\,2} - \left(\frac{1}{N}\sum_i x_i\right)^2}$$

- Representa la *diferència típica* entre qualsevol punt i la mitjana (la major part dels punts estan a 1 o 2 σ de la mitjana).

- Com que les unitats de la desviació típica són les de *x*, és preferible utilitzar σ en lloc de V(x) (encara que els estadístics prefereixen treballar amb V).

- Les desviacions típiques dels histogrames dels exemples anteriors són 0.50 punts i 1.45 punts respectivament.

Mesures de dispersió

- **Una altra definició de la desviació típica o estàndard:**

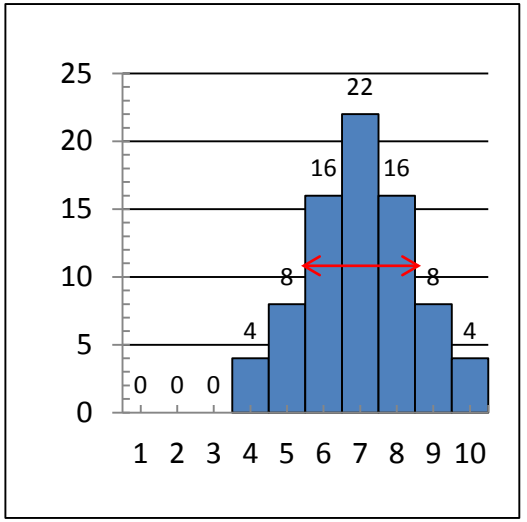
$$s = \sqrt{\frac{1}{N-1}\sum_i (x_i - \overline{x})^2}$$

(sample standard deviation)

>> std(x)

- **Desviació absoluta mitja:** $\frac{1}{N}\sum_{i=1}^{N}\left|x_i - \overline{x}\right|$

- **FWHM** (Full width at half maximum): és una mesura de l'amplària del pic central al mig del màxim del pic.
  (per una gaussiana FWHM=2.35σ).



FWHM=3 punts (fletxa vermella)

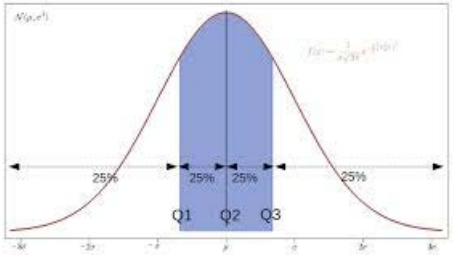
- **Recorregut:** És la diferència entre el valor màxim i el valor mínim de la variable X.

$$R = màx(X) - \min(X)$$

- **Rang:** És l'interval que té per extrems el mínim i el màxim de la variable.

Mesures de posició

- **Quartil:** Els quartils divideixen la distribució en quatre parts iguals ordenades en forma creixent.
  - Quartil inferior o primer quartil Q_1: és el punt amb el 25% de dades per davall d'aquest punt.
  - Segon quartil Q_2: És la **mediana**, ja que el 50% dels valors són iguals o menors a Q_2.
  - Quartil superior o tercer quartil Q_3: El 75% dels valors són iguals o inferiors a aquest.



- **Rang interquartílic:** mesura de la dispersió de les dades.

$$RI = Q_3 - Q_1$$

- **Decil:** Els valors que divideixen les dades en 10 parts iguals es denominen decils.

- **Percentil:** Els valors que divideixen les dades en 100 parts iguals es denominen percentils. El percentil k-èsim P_k és un valor tal que el k% de les observacions són menors o iguals a ell.

Covariància i correlació entre variables

- **Covariància:** Si tenim una mostra de dades agrupades en N **parelles** (x_i,y_i) corresponents a dues variables estadístiques diferents X i Y, definim la covariància entre aquestes dues variables com a:

* Vegeu la demostració pàg. 14

$$\mathrm{cov}(x,y) \equiv \frac{1}{N}\sum_i (x_i-\bar{x})(y_i-\bar{y}) \overset{*}{=} \overline{xy} - \bar{x}\,\bar{y}$$

```
>> Mcov=cov(x_col,y_col,1);
>> covxy=Mcov(1,2)
```

- Si X i Y són independents, les desviacions al voltant de les mitjanes donen valors positius i negatius de forma independent i de tal manera que la suma neta es fa zero. Si valors grans de X estan associats a valors grans de Y, el signe de les desviacions tendirà a ser el mateix, donant una covariància positiva. Si per contra X grans van associats a Y petits, donaran una covariància negativa.

- **Correlació:** Es defineix *el coeficient de correlació lineal* entre dues variables X i Y com la quantitat adimensional

$$\rho = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$$

$$\rho = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2}\sqrt{\sum_i (y_i-\bar{y})^2}}$$

```
>>Mcorr=corrcoef(x_col,y_col)
>>corrxy=Mcorr(1,2)
```

Covariància i correlació

- **Propietat:** $-1 \le \rho \le +1$

  – **Demostració:**

$$x'_i = x_i - \bar{x}; \quad y'_i = y_i - \bar{y}$$

$$\sum (y'_i - cx'_i)^2 \ge 0 \Rightarrow \sum \left(y'^2_i + c^2 x'^2_i - 2c x'_i y'_i \right) \ge 0$$

$$\text{Si } \; c = \frac{\sum x'_i y'_i}{\sum x'^2_i}$$

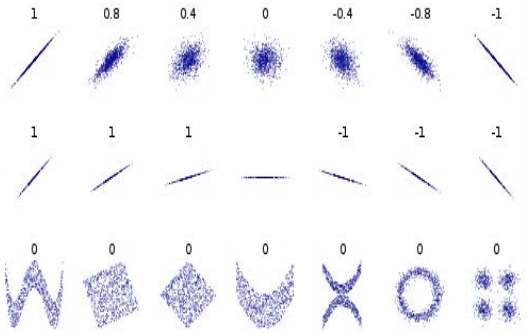
$$\sum y'^2_i + \left(\frac{\sum x'_i y'_i}{\sum x'^2_i} \right)^2 \sum x'^2_i - 2 \frac{\sum x'_i y'_i}{\sum x'^2_i} \sum x'_i y'_i \; \ge 0$$

$$\sum y'^2_i - \frac{(\sum x'_i y'_i)^2}{\sum x'^2_i} \ge 0$$

$$\sum y'^2_i \sum x'^2_i \ge (\sum x'_i y'_i)^2$$

$$\rho^2 = \frac{\left(\sum x'_i y'_i \right)^2}{\sum x'^2_i \sum y'^2_i} \le 1 \Rightarrow -1 \le \rho \le +1$$

La covariància i la correlació són mesures de relació lineal. El signe de les dues indica la direcció de l'associació: valors positius evidencien relacions positives i viceversa. El coeficient de correlació té un avantatge addicional: indica també el grau o força de la relació lineal. Els seus valors estan acotats en l'interval [-1, 1]: com més gran és el valor absolut de la correlació, més gran és el grau d'associació lineal entre les variables. En l'extrem, si la relació és perfectament lineal, el coeficient de correlació és igual a 1 en valor absolut. Finalment, si la correlació és propera a zero, podem dir que no hi ha prova de relació lineal, encara que podria haver-hi un altre tipus de relació entre les variables.



Covariància i correlació

Exemple: Notes d'un grup de Física en diferents matèries.

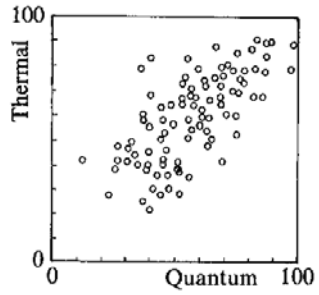


Fig. 2.5 Marks in quantum mechanics and thermal physics.

Els estudiants que aproven la quàntica, en general, aproven també la termo i viceversa (coef. correl. = 0.7). Correlació forta.

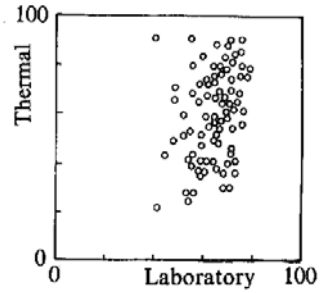


Fig. 2.6 Marks in laboratory and thermal physics.

Els estudiants que aproven el laboratori no necessàriament aproven la termo (coef. correl. = 0.3). Correlació dèbil.

Covariància i correlació

- **Significat quantitatiu del coeficient de correlació**
  - Diem que dues variables x i y estan no correlacionades quan el seu coeficient de correlació, en el límit d'infinites mesures, és zero.
  - Si el que tenim és un nombre finit de punts, és molt improbable que siga exactament zero, i , de fet, podem calcular la probabilitat de tenir un coeficient igual o més gran que un determinat valor ρ_0 quan tenim un conjunt de *N* punts no correlacionats i la designem per:

$$\mathrm{Prob}_{\mathrm{N}}\left(|\rho| \geq \rho_0\right)$$

  - Si aquest valor és menor del 5%, direm que les variables estan *significativament correlacionades* (o correlacionades amb un 95% de seguretat o nivell de confiança).
  - Si és menor de l'1%, direm que estan *molt significativament correlacionades.*

>>[Mcorr,ProbN]=corrcoef(x_{col},y_{col})

Table C. The percentage probability $Prob_N(|r| \geq r_o)$ that N measurements of two uncorrelated variables give a correlation coefficient with $|r| \geq r_o$, as a function of N and r_o. (Blanks indicate probabilities less than 0.05%.)

	r_o										
N	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
3	100	94	87	81	74	67	59	51	41	29	0
4	100	90	80	70	60	50	40	30	20	10	0
5	100	87	75	62	50	39	28	19	10	3.7	0
6	100	85	70	56	43	31	21	12	5.6	1.4	0
7	100	83	67	51	37	25	15	8.0	3.1	0.6	0
8	100	81	63	47	33	21	12	5.3	1.7	0.2	0
9	100	80	61	43	29	17	8.8	3.6	1.0	0.1	0
10	100	78	58	40	25	14	6.7	2.4	0.5		0
11	100	77	56	37	22	12	5.1	1.6	0.3		0
12	100	76	53	34	20	9.8	3.9	1.1	0.2		0
13	100	75	51	32	18	8.2	3.0	0.8	0.1		0
14	100	73	49	30	16	6.9	2.3	0.5	0.1		0
15	100	72	47	28	14	5.8	1.8	0.4			0
16	100	71	46	26	12	4.9	1.4	0.3			0
17	100	70	44	24	11	4.1	1.1	0.2			0
18	100	69	43	23	10	3.5	0.8	0.1			0
19	100	68	41	21	9.0	2.9	0.7	0.1			0
20	100	67	40	20	8.1	2.5	0.5	0.1			0
25	100	63	34	15	4.8	1.1	0.2				0
30	100	60	29	11	2.9	0.5					0
35	100	57	25	8.0	1.7	0.2					0
40	100	54	22	6.0	1.1	0.1					0
45	100	51	19	4.5	0.6						0
	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	
50	100	73	49	30	16	8.0	3.4	1.3	0.4	0.1	
60	100	70	45	25	13	5.4	2.0	0.6	0.2		
70	100	68	41	22	9.7	3.7	1.2	0.3	0.1		
80	100	66	38	18	7.5	2.5	0.7	0.1			
90	100	64	35	16	5.9	1.7	0.4	0.1			
100	100	62	32	14	4.6	1.2	0.2				

Matriu de covariància

- Quan tenim un nombre *n* de variables estadístiques $x_{(1)}, x_{(2)}, \ldots x_{(n)}$ tenim una covariància per cadascuna de les diferents parelles:

$$\mathrm{cov}(x_{(i)}, x_{(j)}) = \overline{x_{(i)} x_{(j)}} - \overline{x}_{(i)} \overline{x}_{(j)}$$

- Agrupem les diferents covariàncies en una matriu *n* x *n* simètrica amb elements

$$V_{ij} = \mathrm{cov}(x_{(i)}, x_{(j)})$$

- Aquesta matriu es diu **matriu de covariància** o també *matriu variància* o *matriu error*.
- Els elements diagonals es corresponen amb les variàncies de cada variable.

$$V_{ii} = \sigma_i^{\,2}$$

- La matriu té la forma:

$$\mathbf{V}_x = \begin{pmatrix} \sigma_1^{\,2} & \cdots & \mathrm{cov}_{1n} \\ \vdots & \ddots & \vdots \\ \mathrm{cov}_{n1} & \cdots & \sigma_n^{\,2} \end{pmatrix}$$

>> cov(X_{col},1)

Distribucions de probabilitat

1. Propietats generals de les distribucions
2. La distribució de Bernoulli
3. La distribució binomial
4. La distribució de Poisson
5. La distribució normal o gaussiana
6. Altres distribucions

Les prediccions de les lleis bàsiques de la ciència són modificades per les distribucions estadístiques a causa del nombre finit de dades, les inexactituds experimentals... Estem obligats a conèixer com aquestes distribucions donen origen a les dades mesurades per a comprendre les lleis de la natura.

Bibliografia*:*

- *Problemes d'estadística amb aplicació a l'enginyeria* (Pujol, Gibergans, García)
- *Statistics. A guide to the use of Statistical Methods in the Physical Sciences* (Barlow)
- *Data reduction and error analysis for the physical sciences* (Bevington, Robinson)

Propietats generals de les distribucions

Exemple 1: Llançament de quatre monedes.

- Per cada moneda la probabilitat de traure cara és ½
  (per tant, ½ de probabilitat de traure creu).
- Probabilitat d'obtenir 4 cares: $P(4) = P(CCCC) = ½ \times ½ \times ½ \times ½ = 1/16$
- Probabilitat d'obtenir 3 cares: $P(3) = P(CCCX) + P(CCXC) + P(CXCC) + P(XCCC) = 4 \times 1/16 = ¼$
- Probabilitat d'obtenir 2 cares:

$$P(2) = P(CCXX) + P(CXCX) + P(CXXC) + P(XCCX) + P(XCXC) + P(XXCC) = 6 \times 1/16 = 3/8$$

- Probabilitat d'obtenir 1 cara (igual a la probabilitat d'obtenir 3 creus): $P(1) = ¼$
- Probabilitat d'obtenir 0 cares (igual a probabilitat d'obtenir 4 creus): $P(0) = 1/16$

- Observem que $\sum_{\nu} P(\nu) = P(0) + P(1) + P(2) + P(3) + P(4) = 1$

- Si ν representa el nombre de cares (ν=0,1,2,3,4), tenim una col·lecció de probabilitats P(ν) que ens dóna la probabilitat de traure ν cares quan llancem una moneda. Tenim, doncs, una *distribució de probabilitat.*

Propietats generals de les distribucions

Exemple 1: Llançament de quatre monedes (n vegades): taula de resultats experimentals.

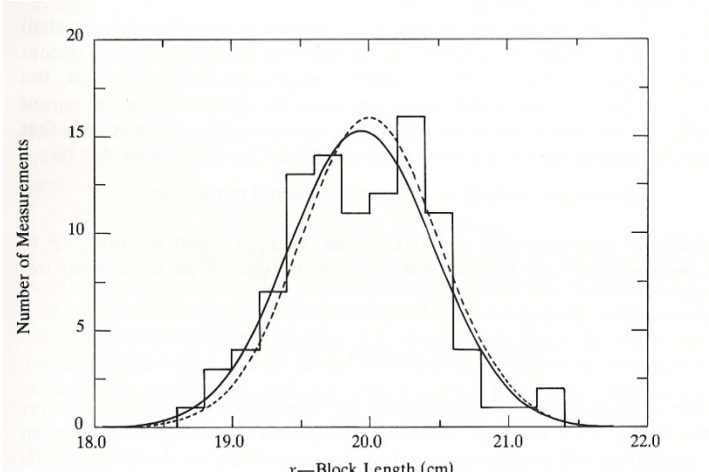
Nombre de cares	ν=4	ν=3	ν=2	ν=1	ν=0
Predicció teòrica (N=16) Dades	1 2	4 7	6 2	4 4	1 1
Predicció teòrica (N=160) Dades	10 10	40 40	60 61	40 38	10 11
Predicció teòrica (N=1600) Dades	100 125	400 403	600 567	400 409	100 96
Predicció teòrica (N=16000) Dades	1000 1009	4000 3946	6000 5992	4000 4047	1000 1006

- L'acord arriba a ser millor i millor a mesura que el nombre d'assaigs augmenta i els efectes aleatoris desapareixen
- La teoria prediu un conjunt de probabilitats. Les dades observades no coincideixen completament, però quan el nombre de dades N augmenta, les fluctuacions es cancel·len, i les freqüències tendeixen a les probabilitats quan N tendeix a infinit (aquesta és la *llei dels grans nombres*).

Propietats generals de les distribucions

Exemple 2

- Histograma de freqüències que representa els resultats de mesurar 100 blocs de fusta de longituds aleatòries entre 18 i 22 cm.
- La corba contínua estimada a partir d'aquestes dades és una gaussiana de mitjana 19.9 cm i desviació típica 0.52 cm.
- La corba discontínua representa la distribució original amb mitjana 20.0 cm i desviació típica 0.50 cm.



La probabilitat de trobar blocs en un interval donat de llargària és:

$$P\left[L_1 < L < L_2\right] = \int_{L_1}^{L_2} P(x)dx$$

P(x) és l'anomenada *densitat de probabilitat* i representa:

$$P(x) = \lim_{\Delta x \to 0} \frac{(\text{Resultats entre } x \text{ y } x + \Delta x)}{\Delta x}$$

Propietats generals de les distribucions

- Moltes **variables aleatòries** associades a experiments estadístics tenen propietats similars i es poden descriure amb una mateixa distribució de probabilitat, que poden ser:
  - **Discretes**: uniforme discreta, de Bernoulli, binomial, de Poisson, geomètrica...
  - **Contínues**: uniforme contínua, exponencial, gaussiana, de Lorentz...

- Aquestes distribucions teòriques s'anomenen **distribucions d'origen** (*parent distributions* o *limiting distributions*) en contraposició a les distribucions de dades obtingudes experimentalment, anomenades **distribució de la mostra** (*sample distributions).*

- Si poguérem fer un nombre N infinit de mesures, la distribució de la mostra coincidiria exactament amb la distribució original que determina les dades estadístiques obtingudes experimentalment (*llei dels grans nombres*).

$$(\text{paràmetre d'origen}) = \lim_{N \to \infty} (\text{paràmetre experimental})$$

Propietats generals de les distribucions

- **Distribució (o funció) probabilitat P(x) (p.d.f.)**
  1. **Variables aleatòries discretes** (associades amb experiments en els quals es compta el nombre de vegades que succeeix un esdeveniment).
     - A cada valor possible x_i de la variable X l'aplicació P(X) li associa un valor $P(x_i)$ que representa la probabilitat que la variable prenga aquest valor x_i.
     - Propietats:

$$0 \le P(x_i) \le 1$$

$$\sum_i P(x_i) = 1$$

  2. **Variables aleatòries contínues**

S'anomena funció de densitat de probabilitat de la variable X la funció $P(x) : \mathbb{R} \to \mathbb{R}^+$ **que verifica les dues condicions següents:**

$$1.\ P(x) \ge 0, \quad \forall x \in \mathbb{R} \qquad\qquad 2.\ \int_{-\infty}^{+\infty} P(x)dx = 1$$

Unitats de P: x^{-1}

A més a més es verifica que

$$P[a < X < b] = \int_a^b P(x)dx$$

En el límit d'un gran nombre d'observacions, **la fracció *dN*** d'observacions de la variable *X* que pren valors entre *x* i *x+dx* és

$$dN = P(x)dx$$

Propietats generals de les distribucions

- **Valors esperats**

  1. Distribucions *P(x)* discretes (P adimensional**)**: Si n és el nombre de valors possibles de l'observable X, el valor esperat de X és:

$$E[X] = \langle X \rangle = \mu = \sum_{j=1}^{n} \left[x_j P(x_j) \right]$$

  P. e. en l'experiment de les quatre monedes:

$$E[X = \text{Nombre de cares}] = 0 \times \frac{1}{16} + 1 \times \frac{1}{4} + 2 \times \frac{3}{8} + 3 \times \frac{1}{4} + 4 \times \frac{1}{16} = 2$$

– Hi ha un paral·lelisme entre el valor esperat i la mitjana d'un conjunt de dades: el primer és una suma sobre una distribució de probabilitat teòrica, i la segona és una suma sobre dades reals.

2. Distribucions contínues

$$\langle X \rangle = \int_{-\infty}^{+\infty} x P(x) dx$$

(Moment de primer ordre)

– En general, el valor esperat de qualsevol funció f(x) és:

$$\langle f(x) \rangle = \sum_{j=1}^{n} \left[f(x_j) P(x_j) \right] \text{ (discreta)} \quad \langle f(x) \rangle = \int_{-\infty}^{+\infty} f(x) P(x) dx \text{ (contínua)}$$

Propietats generals de les distribucions

- **Variàncies**

1. Distribucions *P(x)* discretes

$$\left\langle (x-\mu)^2 \right\rangle = V(X) = \sigma^2 = \sum_{j=1}^{n} \left[(x_j - \mu)^2 P(x_j) \right]$$

En l'experiment de les quatre monedes

$$V(X) = (0-2)^2 \times \frac{1}{16} + (1-2)^2 \frac{1}{4} +$$
$$+(2-2)^2 \times \frac{3}{8} + (3-2)^2 \times \frac{1}{4} + (4-2)^2 \times \frac{1}{16} = 1$$

Propietat: $V(X) = \left\langle x^2 \right\rangle - \mu^2$

2. Distribucions *P(x)* contínues

Propietat: $V(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 P(x) dx$ (Moment central de segon ordre)

$$V(X) = \int_{-\infty}^{+\infty} x^2 P(x) dx - \mu^2 \quad = \left\langle x^2 \right\rangle - \mu^2$$

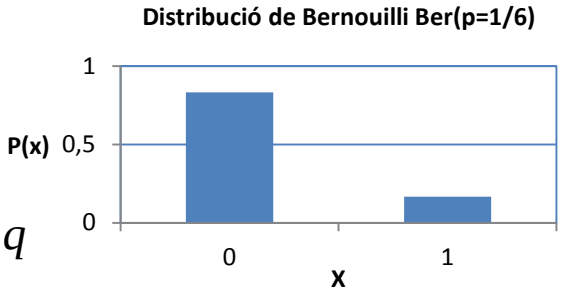
La covariància de dues variables es defineix com a:

$$\mathrm{cov}(x, y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

La distribució de Bernoulli

- Es considera **un** experiment aleatori qualsevol i en **cada** realització de l'experiment s'estudia si ocorre un esdeveniment A (èxit) o no ocorre (fracàs) (només dues opcions).
- Es defineix la variable X amb **distribució de Bernoulli Ber(p)** de la següent manera

$$\begin{cases} X(\text{èxit}) = 1 \\ X(\text{fracàs}) = 0 \end{cases}$$



- La funció de probabilitat és $\begin{cases} P[X=1] = p \\ P[X=0] = 1-p = q \end{cases}$
- Valor esperat $\mu = \sum_{j=1}^{n} \left[x_j P(x_j) \right] = \left[1 \times p + 0 \times q \right] = p$
- Variància: $V(X) = \sum_{j=1}^{n} \left[\left(x_j - \mu \right)^2 P(x_j) \right] = (1-p)^2 p + (0-p)^2 q = pq$

La distribució binomial

- Es realitzen **n experiments independents de Bernoulli** Ber(p), sent p la probabilitat d'èxit. Aleshores es compten quantes vegades s'obté èxit.
- **Distribució binomial $B_{n,p}(v)$:** La variable aleatòria X, que compta el nombre d'èxits en fer l'experiment n vegades, té una funció de probabilitat:

$$P[X = v] = \binom{n}{v} p^v q^{n-v}$$

amb $\binom{n}{v} = \frac{n!}{v!(n-v)!}$

>>% Coeficients binomials
>> nchoosek(n,v)

Els coeficients binomials apareixen en el desenvolupament binomial

$$(p+q)^n = p^n + np^{n-1}q + \ldots + q^n = \sum_{v=0}^{n} \binom{n}{v} p^v q^{n-v}$$

- És a dir, P[X=v] és la probabilitat d'obtenir *v* èxits en n experiments.

- Es compleix que:

$$\sum_v P = \sum_{v=0}^{n} \binom{n}{v} p^v (1-p)^{n-v} = (p+1-p)^n = 1$$

La distribució binomial

- **Valor esperat de la binomial** $\mu = np$

Demostració

$$\mu = \sum_{\nu=0}^{n} \nu B_{n,p}(\nu) = \sum_{\nu=0}^{n} \nu \binom{n}{\nu} p^{\nu} q^{n-\nu} = \sum_{\nu=0}^{n} \nu p^{\nu} q^{n-\nu} \frac{n!}{\nu!(n-\nu)!} = 0 + np \sum_{\nu=1}^{n} p^{\nu-1} q^{n-\nu} \frac{(n-1)!}{(\nu-1)!(n-\nu)!}$$

Fent el canvi $\nu' = \nu - 1$ i $n' = n - 1$

$$\mu = np \sum_{\nu'=0}^{n'} p^{\nu'} q^{n'-\nu'} \frac{n'!}{\nu'!(n'-\nu')!} = np(p+q)^{n'} = np(p+1-p)^{n'} = np$$

q.e.d.

La distribució binomial

- **Variància de la binomial** $\boxed{V(X) = npq}$

Demostració

$$\mathrm{V}(X) = \langle \nu^2 \rangle - \mu^2 = \langle \nu^2 \rangle - n^2 p^2$$

Calculem ara $\langle \nu^2 \rangle$ a partir de:

$$\langle \nu(\nu-1) \rangle = \sum_{\nu=0}^{n} \nu(\nu-1) p^{\nu} q^{n-\nu} \frac{n!}{\nu!(n-\nu)!} = 0 + 0 + \sum_{\nu=2}^{n} \nu(\nu-1) p^{\nu} q^{n-\nu} \frac{n!}{\nu!(n-\nu)!} =$$

$$= p^2 n(n-1) \sum_{\nu=2}^{n} p^{\nu-2} q^{n-\nu} \frac{(n-2)!}{(\nu-2)!(n-\nu)!} \qquad \text{Si fem ara} \quad \nu' = \nu - 2 \quad \text{i} \quad n' = n - 2$$

$$\langle \nu^2 \rangle - \langle \nu \rangle = p^2 n(n-1) \sum_{\nu'=0}^{n'} p^{\nu'} q^{n'-\nu'} \frac{n'!}{\nu'!(n'-\nu')!} = p^2 n(n-1)(p+q)^{n'} = p^2 n(n-1)$$

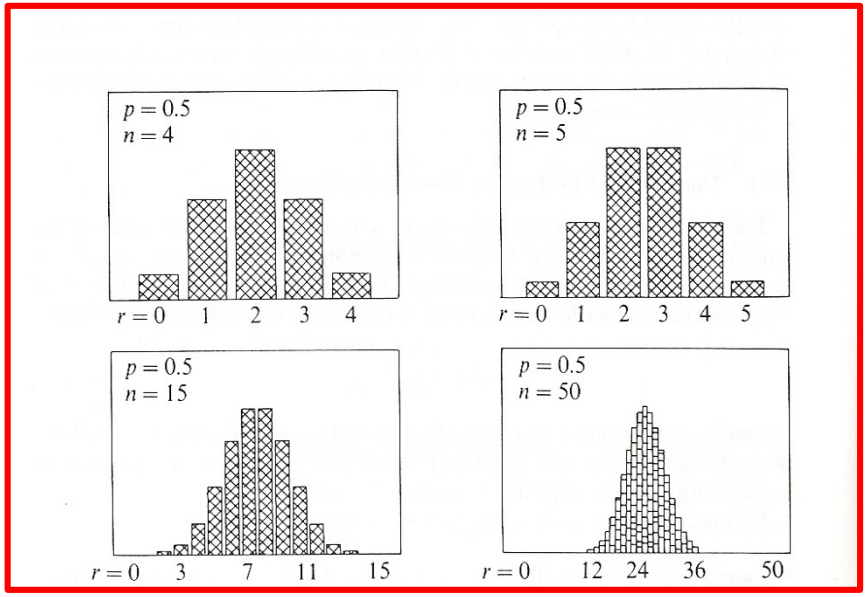
$$\langle \nu^2 \rangle = p^2 n(n-1) + \langle \nu \rangle = p^2 n(n-1) + np$$

$$V(X) = \langle \nu^2 \rangle - n^2 p^2 = \left[p^2 n(n-1) + np \right] - n^2 p^2 = np - np^2 = np(1-p) = npq$$

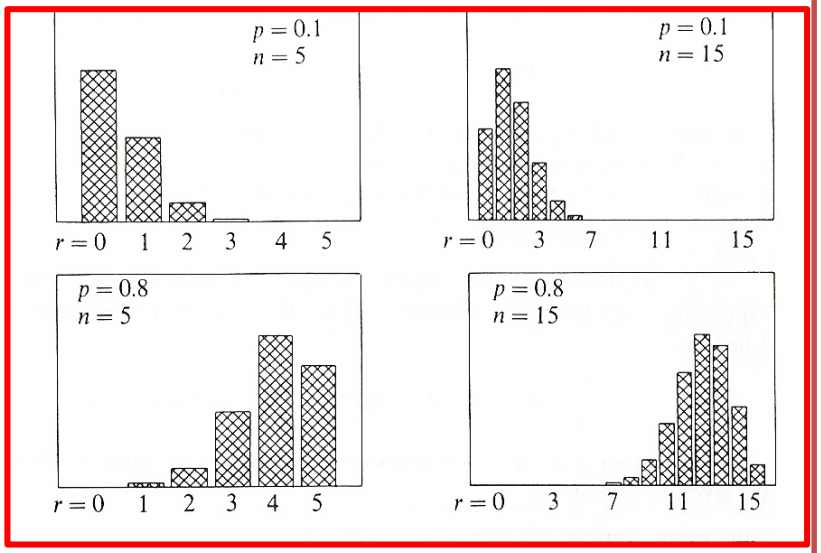
q.e.d.

La distribució binomial

Algunes binomials amb diferents valors de n i p.



Quan p=0.5 la binomial és simètrica al voltant de n/2



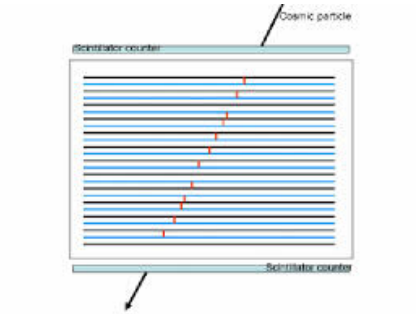
Si p≠0.5, en general, la binomial no és simètrica

– En el límit de n gran (n>30) i p no massa petit (p>0.05), la binomial es pot aproximar per una gaussiana de mitjana *np* i variància *npq,* la qual cosa ens permet obtenir les probabilitats associades a nombres grans evitant els càlculs tediosos dels coeficients binomials.

La distribució binomial

Estem intentant mesurar les traces dels raigs còsmics utilitzant càmeres d'espurnes (*spark chambers*) cada una de les quals té una eficiència del 95%. Amb el criteri que una traça està definida amb tres punts, calculeu l'eficiència en la detecció de les traces si tenim: a) tres càmeres; b) quatre càmeres.

- Cada càmera pot espurnejar o no amb probabilitats 0.95 i 0.05 respectivament.
- Siga X la variable *espurneja* quan passa la partícula.



a) Com que la traça està definida per 3 punts,
les tres càmeres han d'espurnejar simultàniament (3 de 3):

$$P = B_{3,0.95}(3) = \binom{3}{3} \times 0.95^3 \times 0.05^0 = \quad 0.86 \quad (86\%)$$

b) Ara tenim 4 càmeres: calculem la probabilitat que espurnegen totes quatre o només tres.

$$P = B_{4,0.95}(4) + B_{4,0.95}(3) = \binom{4}{4} \times 0.95^4 \times 0.05^0 + \binom{4}{3} \times 0.95^3 \times 0.05^1 = 0.81 + 0.16 = 0.986 \quad (98.6\%)$$

La distribució de Poisson

- Descriu els resultats dels experiments de recompte d'esdeveniments aleatoris, però amb una tassa mitjana definida μ.
- En general l'experiment consisteix a observar l'aparició d'un esdeveniment puntual A, en un interval continu de temps o espai. Per exemple, comptar el nombre de vehicles que passen per un lloc determinat, durant un interval de temps; o comptar el nombre de plantes d'una determinada espècie que hi ha en certa superfície de bosc, nombre de desintegracions d'una mostra radioactiva en un interval de temps...
- **Característiques d'un experiment de Poisson:**
  - **Els successos s'esdevenen aleatòriament i de manera independent en un interval continu d'espai o de temps.**
  - **Es produeix un comportament uniforme en el sentit que, a llarg termini, el nombre mitjà de vegades que passa l'esdeveniment A és constant per unitat d'observació, μ.**
  - **Si dividim l'interval en n subintervals molt menuts la possibilitat que dos esdeveniments s'esdevinguen en el mateix subinterval pot considerar-se negligible o nul·la, de manera que cada subinterval té una probabilitat $p=\mu/n$ d'incloure un esdeveniment A.**

- **Distribució de Poisson:** dóna la probabilitat d'obtenir ν esdeveniments sent el valor esperat μ.

$$P_{\mu}(X=\nu)=\frac{\mu^{\nu}}{\nu!}e^{-\mu}$$

La distribució de Poisson

- La distribució de Poisson s'obté com el límit d'una binomial quan el nombre n de vegades que es realitza l'experiment tendeix a infinit, la probabilitat d'èxit p tendeix a zero i el nombre mitjà d'èxits s'estabilitza al voltant d'un valor constant $\mu = np$ ($p=\mu/n$)

$$P[X=\nu] = \frac{n!}{\nu!(n-\nu)!} p^{\nu} q^{n-\nu} = \frac{n!}{\nu!(n-\nu)!} \frac{\mu^{\nu}}{n^{\nu}} \left(1-\frac{\mu}{n}\right)^{n-\nu}$$

$$\frac{n!}{(n-\nu)!} = n(n-1)\ldots(n-\nu+1) \xrightarrow[n\to\infty]{} n^{\nu}$$

$$\left(1-\frac{\mu}{n}\right)^{n-\nu} \xrightarrow{n\to\infty} \left(1-\frac{\mu}{n}\right)^{n} \xrightarrow{n\to\infty} e^{-\mu}$$

$$\xrightarrow[n\to\infty]{} P = \frac{\mu^{\nu}}{\nu!} e^{-\mu}$$

$$\ln\left(1-\frac{\mu}{n}\right)^{n} = n\ln\left(1-\frac{\mu}{n}\right) \xrightarrow{\infty} n\left(-\frac{\mu}{n}\right) = -\mu = \ln e^{-\mu}$$

$$\ln(1-x) = -\left(x - \frac{x^2}{2} + \frac{x^3}{2} - \ldots\right)$$

- **Valor esperat**: $\langle X \rangle = \mu$
- **Variància**: $V(X) = \mu$

- En la pràctica, si p<0.05 i *np* es manté finita, es pot aproximar la binomial per una Poisson.

La distribució de Poisson

– Demostracions:

1. Normalització:

$$\sum_{\nu=0}^{\infty} P_{\mu}(\nu) = e^{-\mu}\sum_{\nu=0}^{\infty}\frac{\mu^{\nu}}{\nu!} = e^{-\mu}\left(1+\mu+\frac{\mu^2}{2!}+\ldots\right) = e^{-\mu}e^{\mu} = 1$$

2. Valor esperat:

$$\langle\nu\rangle = \sum_{\nu=0}^{\infty}\nu\frac{\mu^{\nu}}{\nu!}e^{-\mu} = 0+\mu e^{-\mu}\sum_{\nu=1}^{\infty}\frac{\mu^{\nu-1}}{(\nu-1)!} \quad \left|\nu' = \nu-1\right| \quad = \mu e^{-\mu}\sum_{\nu'=0}^{\infty}\frac{\mu^{\nu'}}{\nu'!} = \mu e^{-\mu}e^{\mu} = \mu$$

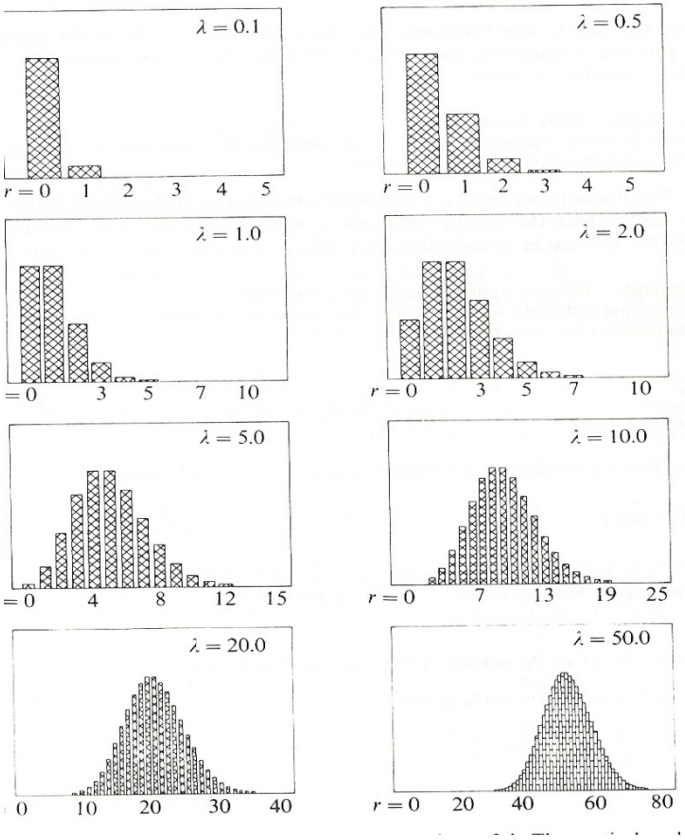
3. Variància:

$$\langle\nu(\nu-1)\rangle = \sum_{\nu=0}^{\infty}\nu(\nu-1)\frac{\mu^{\nu}}{\nu!}e^{-\mu} = 0+0+\mu^2 e^{-\mu}\sum_{\nu=2}^{\infty}\frac{\mu^{\nu-2}}{(\nu-2)!} \quad \left|\nu' = \nu-2\right| \quad =$$

$$= \mu^2 e^{-\mu}\sum_{\nu'=0}^{\infty}\frac{\mu^{\nu'}}{\nu'!} = \mu^2 e^{-\mu}e^{\mu} = \mu^2 \Rightarrow \langle\nu^2\rangle = \mu^2+\langle\nu\rangle = \mu^2+\mu$$

$$V(X) = \langle\nu^2\rangle - \langle\nu\rangle^2 = (\mu^2+\mu)-\mu^2 = \mu$$

La distribució de Poisson



Diverses distribucions de Poisson amb mitjanes λ.

Com més gran es fa la mitjana, més simètrica es fa la distribució de Poisson. De fet, es pot demostrar que per a valors grans de μ la distribució de Poisson s'aproxima a una distribució normal de mitjana μ i $\sigma = \sqrt{\mu}$.

En la pràctica, si μ>10, es pot aproximar la Poisson per una gaussiana.

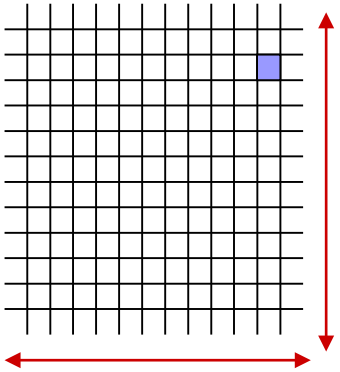
La distribució de Poisson

- **Exemple**: Segona Guerra Mundial. Bombardeig de Londres per part de l'exèrcit alemany amb bombes voladores. ¿Apuntaven o disparaven a l'atzar?

Dividim els 144 km² de Londres en 576 sectors de 0.25 km². Com que van caure n=537 bombes, **repartides aleatòriament**, esperem un nombre de bombes caigudes en cada sector de

$$\mu = \frac{537 \text{ bombes}}{576 \text{ sector}} = 0.932 \quad \frac{\text{bombes}}{\text{sector de } 0.25\text{km}^2}$$

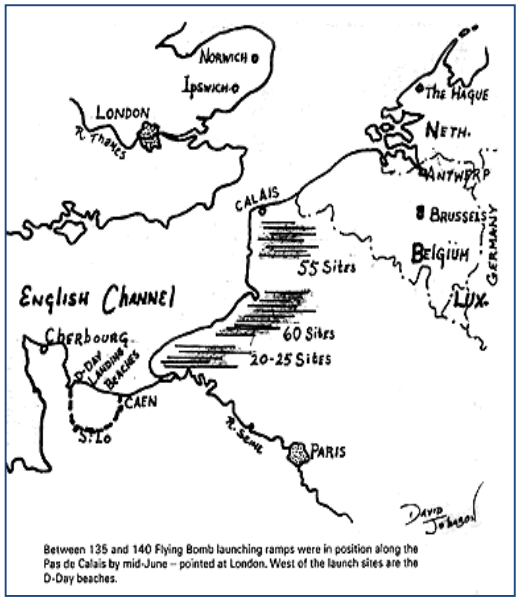
seguint una distribució Poissoniana (*n*=537, *p*=1/576=0.0017) amb aquesta mitjana μ=*np*=0.932.



Probabilitat que hagen caigut ν bombes en un sector determinat:

$$P_\mu(X = \nu) = \frac{\mu^\nu}{\nu!} e^{-\mu}$$

$$\frac{144}{0.25} = 576 \quad \text{sectors}$$

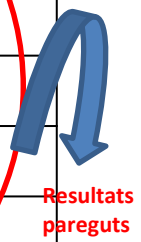


Between 135 and 140 Flying Bomb launching ramps were in position along the Pas de Calais by mid-June – pointed at London. West of the launch sites are the D-Day beaches.

La distribució de Poisson

- Calculem P per a v=0,1,2,3,4... bombes i sumem per als 576 sectors i comparem amb les bombes que van caure realment

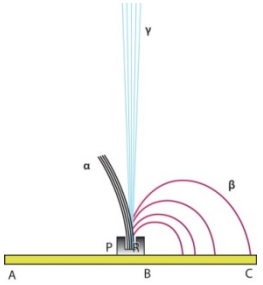
Nombre v de bombes per sector	$P_{\mu=0.932}(v)$	Nombre **esperat** de sectors amb v bombes	Nombre **observat** de sectors amb v bombes
0	39.4%	$576P_{0.932}(v=0)=226.7$	229
1	36.7%	211.4	211
2	17.1%	98.5	93
3	5.3%	30.6	35
4	1.2%	7.1	7
5 o més	0.27%	1.6	1
Total	100%	576	576



Conclusió: No apuntaven contra cap objectiu particular.

La distribució de Poisson

- **Exemple**: Comptant desintegracions radioactives.



- Considerem una font radioactiva, com **el Cs-137** amb un període de semidesintegració (*half-life*) de T=30.1 anys.
- La probabilitat per unitat de temps que es desintegre **un** nucli és la seua constant de desintegració λ, relacionada amb T per: $\lambda = p = \ln 2/T = 7.3 \times 10^{-10} s^{-1}$
- $1\mu g$ de Cs-137 conté de l'ordre de $n=10^{15}$ nuclis.
- El nombre de desintegracions per unitat de temps d'aquesta font radioactiva té una mitjana $\mu = np = 7.3 \times 10^{5} s^{-1}$, que satisfà les condicions de Poisson.
- La probabilitat d'observar ν desintegracions per segon és $P_\mu(X=\nu) = \dfrac{\mu^\nu}{\nu!} e^{-\mu}$

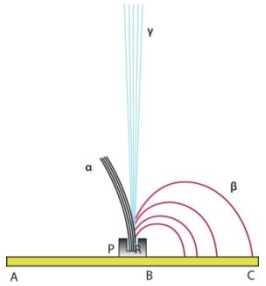
- Com que només apareix la mitjana de desintegracions per unitat de temps, no cal conèixer *n* i *p*.
- En general, si volem conèixer la probabilitat d'observar un nombre ν de desintegracions en un interval de temps t, coneguda la taxa de desintegracions *R,* podem escriure:

$$P_\mu(X=\nu) = \frac{(Rt)^\nu}{\nu!} e^{-Rt}$$

on hem substituït μ per *Rt.*

La distribució de Poisson

- **Exemple:** Una mostra de tori radioactiu emet partícules alfa al ritme d'1.5 partícules per minut. Si comptem en intervals de temps de 2 minuts, quin serà el resultat esperat i amb quina probabilitat? Quina és la probabilitat de trobar 0, 1, 2, 3, 4 partícules? I més de 4? Quina és la variància esperada?



  - Recompte esperat de partícules en 2 minuts:

  - Probabilitat

$$\mu = Rt = 1.5\text{min}^{-1} \times 2\text{min} = 3$$

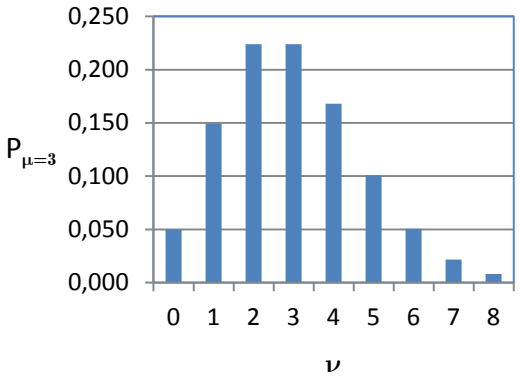
$$P_\mu(X=3) = \frac{\mu^3}{3!}e^{-\mu} = \frac{3^3}{3!}e^{-3} = 0.224 \rightarrow 22.4\%$$

ν	$P_{\mu=3}(\nu)$
0	0.050
1	0.149
2	0.224
3	0.224
4	0.168
5	0.101
6	0.050
7	0.022
8	0.008
9	0.003

$$P_\mu(X \geq 5) =$$

$$= 1 - \sum_{\nu=0}^{\nu=4} P_{\mu=3}(\nu) = 0.185$$

$$18.5\%$$



  - Variància:

$$V(X) = \mu = 3$$

La distribució normal o gaussiana

- És l'exemple més important de distribució de probabilitat associada a una variable aleatòria contínua. La seua importància rau en el fet que molts fenòmens segueixen aquest model.
- **Distribució normal**

$$G_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- **Propietats:**
  - És simètrica respecte a la recta x=μ.
  - L'eix OX és una asímptota de G (x).
  - Té el seu màxim en x=μ.
  - Té dos punts d'inflexió en x= μ-σ i x= μ+σ.
  - La moda i la mediana valen μ.

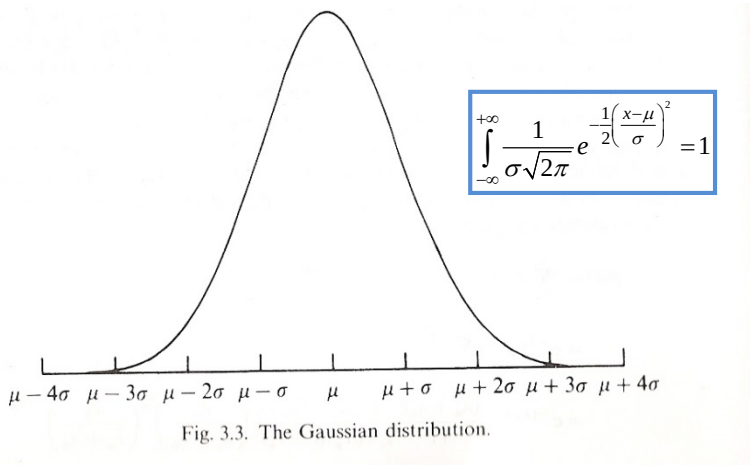
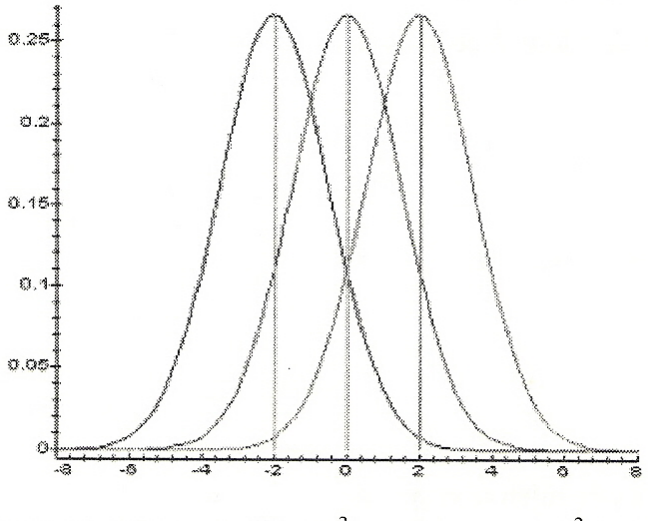


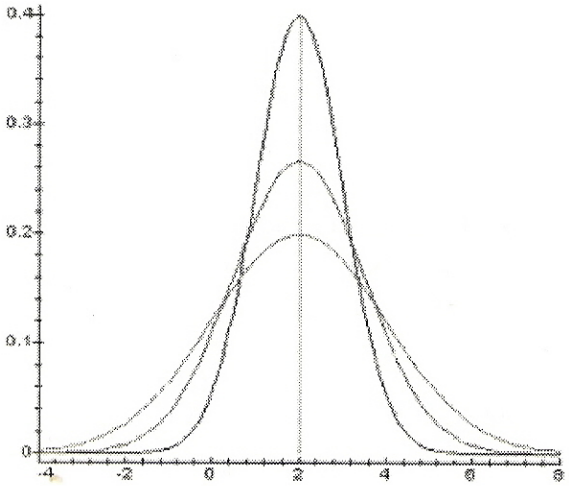
Fig. 3.3. The Gaussian distribution.

- **Valor esperat:** $\langle X \rangle = \mu$
- **Variància:** $V(X) = \sigma^2$

La distribució normal o gaussiana

- **Exemples**





Si es fixa σ i es fa variar μ la gràfica es desplaça, però no modifica la seua forma.

Si es fixa μ, la gràfica es fa més ampla com més gran es fa σ.

La distribució normal o gaussiana

- **Distribució normal tipificada:** Si tenim μ=0 i σ^2=1, la llei normal corresponent es denomina llei normal tipificada o estandarditzada i es designa per $G_{0,1}(x)$ (*Unit normal distribution)*

- **La variable aleatòria** $\boxed{t = \frac{x-\mu}{\sigma}}$

segueix una normal tipificada. En efecte,

$$\frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi}}\int_{t_a}^{t_b} e^{-t^2/2} dt = \int_{t_a}^{t_b} G_{0,1}(t)\, dt$$

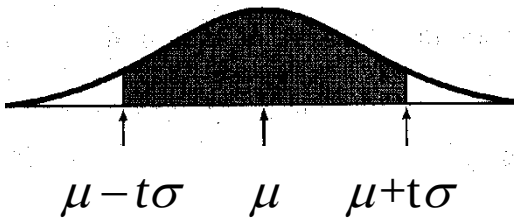
$$t = \frac{x-\mu}{\sigma} \Rightarrow dt = \frac{dx}{\sigma}$$
$$t_a = \frac{a-\mu}{\sigma}$$
$$t_b = \frac{b-\mu}{\sigma}$$

i, per tant,

$$P\left[a < X < b\right] = P\left[t_a = \frac{a-\mu}{\sigma} < t < t_b = \frac{b-\mu}{\sigma}\right] = \int_{t_a}^{t_b} G_{0,1}(t)dt$$

La distribució normal o gaussiana

- **Funció error**



$$\operatorname{erf}(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{\frac{-t^2}{2}} dt$$

$$\operatorname{er}(t) = P\left[\mu - t\sigma < X < \mu + t\sigma\right]$$

$$t = \frac{X - \mu}{\sigma}$$

Integral gaussiana de doble cua (valors de erf(*t*))

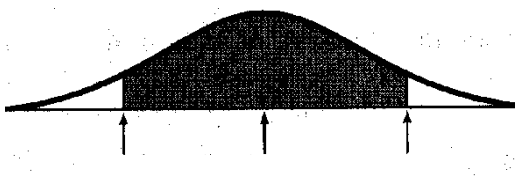
t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.80	1.60	2.39	3.19	3.99	4.78	5.58	6.38	7.17
0.1	7.97	8.76	9.55	10.34	11.13	11.92	12.71	13.50	14.28	15.07
0.2	15.85	16.63	17.41	18.19	18.97	19.74	20.51	21.28	22.05	22.82
0.3	23.58	24.34	25.10	25.86	26.61	27.37	28.12	28.86	29.61	30.35
0.4	31.08	31.82	32.55	33.28	34.01	34.73	35.45	36.16	36.88	37.59
0.5	38.29	38.99	39.69	40.39	41.08	41.77	42.45	43.13	43.81	44.48
0.6	45.15	45.81	46.47	47.13	47.78	48.43	49.07	49.71	50.35	50.98
0.7	51.61	52.23	52.85	53.46	54.07	54.67	55.27	55.87	56.46	57.05
0.8	57.63	58.21	58.78	59.35	59.91	60.47	61.02	61.57	62.11	62.65
0.9	63.19	63.72	64.24	64.76	65.28	65.79	66.29	66.80	67.29	67.78
1.0	68.27	68.75	69.23	69.70	70.17	70.63	71.09	71.54	71.99	72.43
1.1	72.87	73.30	73.73	74.15	74.57	74.99	75.40	75.80	76.20	76.60
1.2	76.99	77.37	77.75	78.13	78.50	78.87	79.23	79.59	79.95	80.29
1.3	80.64	80.98	81.32	81.65	81.98	82.30	82.62	82.93	83.24	83.55
1.4	83.85	84.15	84.44	84.73	85.01	85.29	85.57	85.84	86.11	86.38
1.5	86.64	86.90	87.15	87.40	87.64	87.89	88.12	88.36	88.59	88.82
1.6	89.04	89.26	89.48	89.69	89.90	90.11	90.31	90.51	90.70	90.90
1.7	91.09	91.27	91.46	91.64	91.81	91.99	92.16	92.33	92.49	92.65
1.8	92.81	92.97	93.12	93.28	93.42	93.57	93.71	93.85	93.99	94.12
1.9	94.26	94.39	94.51	94.64	94.76	94.88	95.00	95.12	95.23	95.34

La distribució normal o gaussiana

- **Funció error**

$$\operatorname{erf}(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{\frac{-t^2}{2}} dt$$

$$\operatorname{er}(t) = P\left[\mu - t\sigma < X < \mu + t\sigma\right]$$

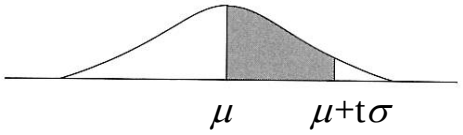


$\mu - t\sigma$ μ $\mu + t\sigma$

Integral gaussiana de doble cua (cont.)

t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.0	95.45	95.56	95.66	95.76	95.86	95.96	96.06	96.15	96.25	96.34
2.1	96.43	96.51	96.60	96.68	96.76	96.84	96.92	97.00	97.07	97.15
2.2	97.22	97.29	97.36	97.43	97.49	97.56	97.62	97.68	97.74	97.80
2.3	97.86	97.91	97.97	98.02	98.07	98.12	98.17	98.22	98.27	98.32
2.4	98.36	98.40	98.45	98.49	98.53	98.57	98.61	98.65	98.69	98.72
2.5	98.76	98.79	98.83	98.86	98.89	98.92	98.95	98.98	99.01	99.04
2.6	99.07	99.09	99.12	99.15	99.17	99.20	99.22	99.24	99.26	99.29
2.7	99.31	99.33	99.35	99.37	99.39	99.40	99.42	99.44	99.46	99.47
2.8	99.49	99.50	99.52	99.53	99.55	99.56	99.58	99.59	99.60	99.61
2.9	99.63	99.64	99.65	99.66	99.67	99.68	99.69	99.70	99.71	99.72
3.0	99.73									
3.5	99.95									
4.0	99.994									
4.5	99.9993									
5.0	99.99994									

La distribució normal o gaussiana



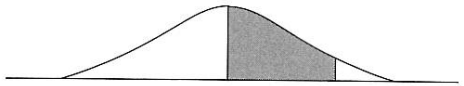
$$\frac{1}{2}\mathrm{er}(t) = \frac{1}{\sqrt{2\pi}}\int_0^{+t} e^{\frac{-t^2}{2}}\,dt =$$

$$= P(\mu < X < \mu + t\sigma)$$

$$t = \frac{X-\mu}{\sigma}$$

t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.40	0.80	1.20	1.60	1.99	2.39	2.79	3.19	3.59
0.1	3.98	4.38	4.78	5.17	5.57	5.96	6.36	6.75	7.14	7.53
0.2	7.93	8.32	8.71	9.10	9.48	9.87	10.26	10.64	11.03	11.41
0.3	11.79	12.17	12.55	12.93	13.31	13.68	14.06	14.43	14.80	15.17
0.4	15.54	15.91	16.28	16.64	17.00	17.36	17.72	18.08	18.44	18.79
0.5	19.15	19.50	19.85	20.19	20.54	20.88	21.23	21.57	21.90	22.24
0.6	22.57	22.91	23.24	23.57	23.89	24.22	24.54	24.86	25.17	25.49
0.7	25.80	26.11	26.42	26.73	27.04	27.34	27.64	27.94	28.23	28.52
0.8	28.81	29.10	29.39	29.67	29.95	30.23	30.51	30.78	31.06	31.33
0.9	31.59	31.86	32.12	32.38	32.64	32.89	33.15	33.40	33.65	33.89
1.0	34.13	34.38	34.61	34.85	35.08	35.31	35.54	35.77	35.99	36.21
1.1	36.43	36.65	36.86	37.08	37.29	37.49	37.70	37.90	38.10	38.30
1.2	38.49	38.69	38.88	39.07	39.25	39.44	39.62	39.80	39.97	40.15
1.3	40.32	40.49	40.66	40.82	40.99	41.15	41.31	41.47	41.62	41.77
1.4	41.92	42.07	42.22	42.36	42.51	42.65	42.79	42.92	43.06	43.19
1.5	43.32	43.45	43.57	43.70	43.82	43.94	44.06	44.18	44.29	44.41
1.6	44.52	44.63	44.74	44.84	44.95	45.05	45.15	45.25	45.35	45.45
1.7	45.54	45.64	45.73	45.82	45.91	45.99	46.08	46.16	46.25	46.33
1.8	46.41	46.49	46.56	46.64	46.71	46.78	46.86	46.93	46.99	47.06
1.9	47.13	47.19	47.26	47.32	47.38	47.44	47.50	47.56	47.61	47.67

La distribució normal o gaussiana



μ $\quad$ μ+tσ

$$\frac{1}{2}\text{er}(t) = \frac{1}{\sqrt{2\pi}}\int_0^{+t} e^{\frac{-t^2}{2}}\,dt =$$

$$= P(\mu < X < \mu + t\sigma)$$

$$t = \frac{X-\mu}{\sigma}$$

t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.0	47.72	47.78	47.83	47.88	47.93	47.98	48.03	48.08	48.12	48.17
2.1	48.21	48.26	48.30	48.34	48.38	48.42	48.46	48.50	48.54	48.57
2.2	48.61	48.64	48.68	48.71	48.75	48.78	48.81	48.84	48.87	48.90
2.3	48.93	48.96	49.98	49.01	49.04	49.06	49.09	49.11	49.13	49.16
2.4	49.18	49.20	49.22	49.25	49.27	49.29	49.31	49.32	49.34	49.36
2.5	49.38	49.40	49.41	49.43	49.45	49.46	49.48	49.49	49.51	49.52
2.6	49.53	49.55	49.56	49.57	49.59	49.60	49.61	49.62	49.63	49.64
2.7	49.65	49.66	49.67	49.68	49.69	49.70	49.71	49.72	49.73	49.74
2.8	49.74	49.75	49.76	49.77	49.77	49.78	49.79	49.79	49.80	49.81
2.9	49.81	49.82	49.82	49.83	49.84	49.84	49.85	49.85	49.86	49.86
3.0	49.87									
3.5	49.98									
4.0	49.997									
4.5	49.9997									
5.0	49.99997									

La distribució normal o gaussiana

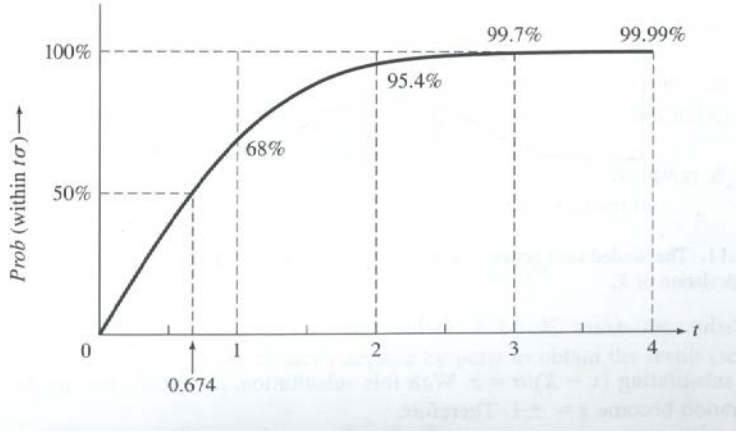
Intervals de confiança: Intervals associats a una probabilitat determinada (valor determinat de la integral de doble cua).

Nivell de confiança: Probabilitat associada a un interval determinat al voltant de la mitjana.

$$P[\mu-\sigma\le x\le\mu+\sigma] = \int_{\mu-\sigma}^{\mu+\sigma} G_{\mu,\sigma}(x)\,dx = \int_{-1}^{+1} G_{0,1}(t)\,dt = 0.6827$$

$$t_b = \frac{(\mu+\sigma)-\mu}{\sigma} = +1 \qquad t_a = \frac{(\mu-\sigma)-\mu}{\sigma} = -1$$

Nivell de confiança associat a un interval de 1σ al voltant de la mitjana: 68.3%.

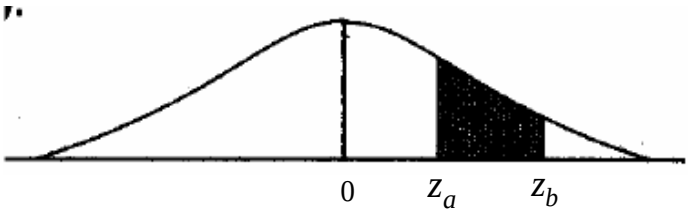


- 68.27% de les dades es troben dins de l'interval [μ-σ, μ+σ] (es troben dins amb un 68% de CL).
- 95.45% es troben a 2σ.
- 99.73% es troben a 3σ.

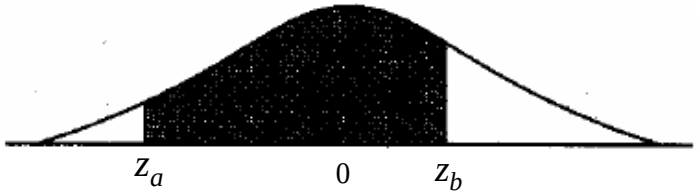
- 90% de les dades es troben dins de l'interval [μ-1.645σ, μ+1.645σ]
- 95% es troben a 1.960σ.
- 99% es troben a 2.576σ.

La distribució normal o gaussiana

Exemples: (canviem la variable *t* per *z*)

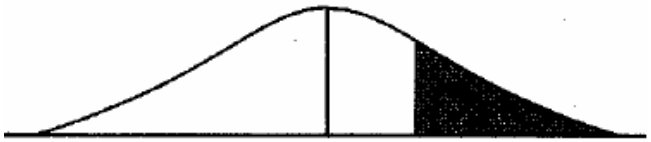


$$\int_{z_a}^{z_b} G_{0,1}(z)\,dz = \int_{0}^{z_b} G_{0,1}(z)\,dz - \int_{0}^{z_a} G_{0,1}(z)\,dz$$



$$\int_{z_a}^{z_b} G_{0,1}(z)\,dz = \int_{0}^{-z_a} G_{0,1}(z)\,dz + \int_{0}^{z_b} G_{0,1}(z)\,dz$$

$(z_a < 0\ ,\ -z_a > 0)$

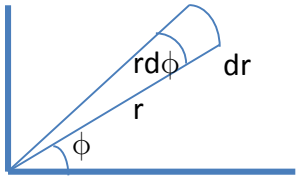


$$\int_{z_a}^{+\infty} G_{0,1}(z)\,dz = 0.5 - \int_{0}^{z_a} G_{0,1}(z)\,dz$$

$$\int_{-\infty}^{z_a} G_{0,1}(z)\,dz = 0.5 + \int_{0}^{z_a} G_{0,1}(z)\,dz$$

La distribució normal o gaussiana

- Demostrem que $$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$



Si $$I = \int_{-\infty}^{+\infty} e^{-x^2} dx \Rightarrow I^2 = \int_{-\infty}^{+\infty} e^{-x^2} dx \int_{-\infty}^{+\infty} e^{-y^2} dy = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy\ e^{-(x^2+y^2)} = \iint_R e^{-(x^2+y^2)} dxdy$$

R és tot el pla XY

Transformem a coordenades polars: $x = \rho\cos\phi \qquad y = \rho\sin\phi$

$$I^2 = \iint_{R'} e^{-\rho^2} \frac{\partial(x,y)}{\partial(\rho,\phi)} d\rho d\phi = \iint_{R'} e^{-\rho^2} \rho d\rho d\phi = \int_0^{2\pi} d\phi \int_0^{\infty} d\rho e^{-\rho^2} \rho = 2\pi\left[-\tfrac{1}{2}e^{-\rho^2}\right]_0^{\infty} = \pi \Rightarrow I = \int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$$

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \left| z = \frac{x-\mu}{\sqrt{2}\sigma} \Rightarrow dz = \frac{dx}{\sqrt{2}\sigma} \right| = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty} e^{z^2} dz = 1$$

La distribució normal o gaussiana

- Demostrem que $\langle X \rangle = \mu$

$$\langle X \rangle = \int_{-\infty}^{\infty} x\, G_{\mu,\sigma}(x)\, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

  - Fem el canvi de variable $y = x - \mu$ ➔ $dy = dx$ i $x = y + \mu$

$$\langle X \rangle = \frac{1}{\sqrt{2\pi\sigma^2}} \left\{ \int_{-\infty}^{\infty} y\, e^{-\frac{y^2}{2\sigma^2}} dy + \mu \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma^2}} dy \right\} = \frac{1}{\sqrt{2\pi\sigma^2}} \left\{ 0 + \mu\sqrt{2\pi\sigma^2} \right\} = \mu$$

  - La primera integral és zero perquè els valors positius es cancel·len amb els negatius.
  - La segona és la integral de la densitat de probabilitat i coincideix amb la constant de normalització.

$$\int_{-\infty}^{+\infty} e^{-ax^2} = \sqrt{\frac{\pi}{a}}$$

La distribució normal o gaussiana

- Demostrem que $V(X) = \sigma^2$

$$V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \, G_{\mu,\sigma}(x)\, dx$$

- Amb els canvis de variable $t = (x - \mu)/\sigma$ $\quad V(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma^2 t^2 e^{-\frac{t^2}{2}} dt$

- Integrem per parts: $\quad u = t \Rightarrow du = dt;\ dv = t\exp(-t^2/2)dt \Rightarrow v = -\exp(-t^2/2)$

$$V(X) = \frac{\sigma^2}{\sqrt{2\pi}}[te^{-\frac{t^2}{2}}]_{-\infty}^{+\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = 0 + \frac{\sigma^2}{\sqrt{2\pi}} \sqrt{2\pi} = \sigma^2$$

Altres distribucions

1. Distribució uniforme

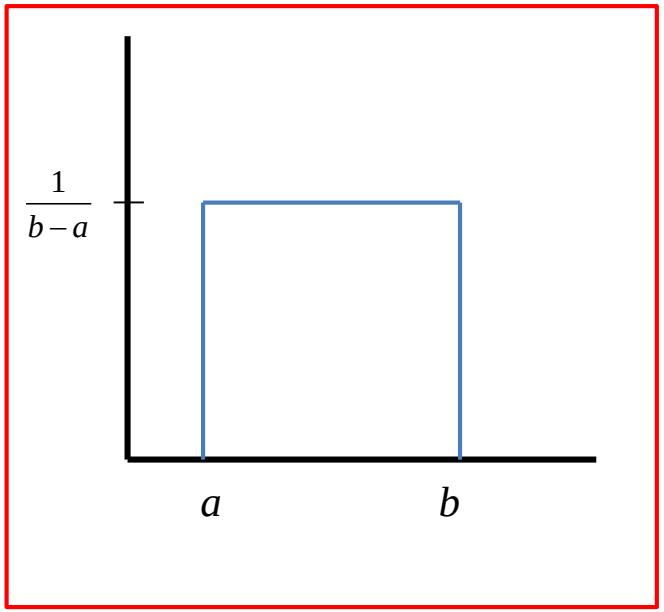
– Descriu una probabilitat constant en un interval determinat i zero fora de l'interval.

$$P(x)=\begin{cases}\dfrac{1}{b-a} & \text{si} \quad a\le x\le b\\ 0 & \text{altrament}\end{cases}$$

– Normalització: $\int_a^b P(x)=\int_a^b \frac{1}{b-a}dx=\frac{1}{b-a}\left[x\right]_a^b=1$

– Valor esperat: $\langle X\rangle=\frac{a+b}{2}$

– Variància: $V(X)=\frac{(b-a)^2}{12}$



Altres distribucions

2. La distribució de Lorentz (o distribució de Cauchy o de Breit-Wigner)

– Descriu el fenomen físic de la ressonància: oscil·lacions forçades, seccions eficaces en física nuclear, absorció de radiació en l'efecte Mössbauer...

$$P(x;\mu,\Gamma)=\frac{1}{\pi}\frac{\Gamma/2}{(x-\mu)^2+(\Gamma/2)^2}$$

– Valor esperat $\langle X\rangle=\mu$

– Variància: no definida, la integral és divergent.

– Amplària del pic:

$$\mathrm{FWHM}=\Gamma$$

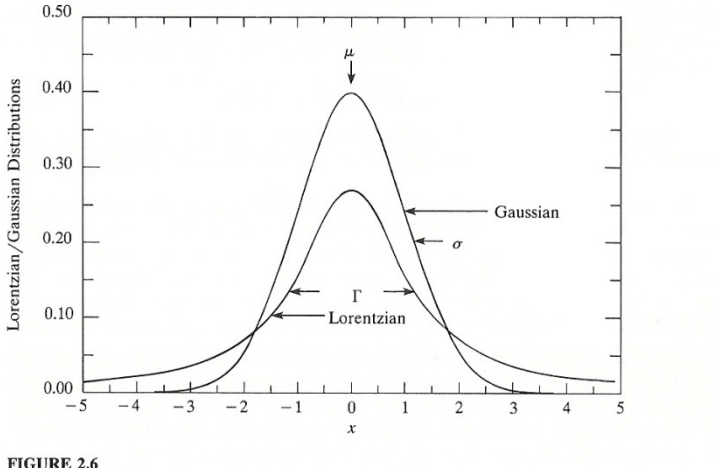


FIGURE 2.6

No disminueix a zero tan ràpidament com ho fa la gaussiana.

Anàlisi d'errors

1. Errors en els experiments de física
2. Errors aleatoris i distribucions estadístiques
3. Error típic de la mitjana
4. Propagació dels errors
5. Fórmules d'errors específiques
6. Formulació matricial de la propagació

"Everybody believes in the law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact."

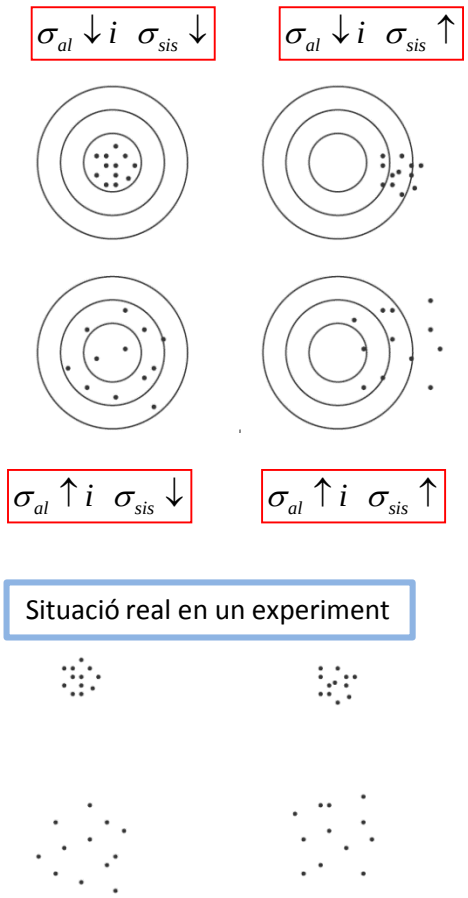
Bibliografia*:*

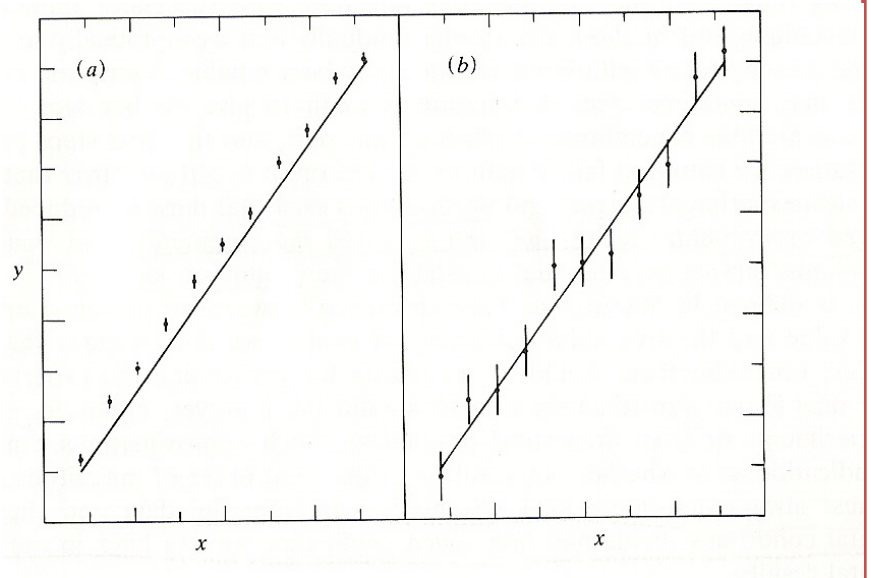
- *Statistics. A guide to the use of Statistical Methods in the Physical Sciences* (Barlow)
- *Data reduction and error analysis for the physical sciences* (Bevington, Robinson)

Errors en els experiments de física

- Els errors o les incerteses són inherents a qualsevol mesura experimental en física i han de ser reduïts millorant les tècniques experimentals o per repetició de les mesures.
- Aquests errors han de ser estimats d'una forma sistemàtica per a donar el resultat i el grau de confiança que tenim en aquests resultats.
- Els errors dels quals parlem podem classificar-los com a:
  - errors aleatoris
  - errors sistemàtics
- Els primers, els **errors aleatoris**, tenen el seu origen en les fluctuacions observades en els resultats quan repetim l'experiment diferents vegades en les mateixes condicions. S'avaluen tant millorant el mètode experimental com amb l'anàlisi estadística del conjunt de dades obtingudes per repetició de les mesures.
- Els segons, els **errors sistemàtics**, no són fàcilment detectables i no poden avaluar-se per repetició de les mesures, sinó analitzant les condicions experimentals i tècniques.
- La **precisió** (en anglès, *precision*) d'una mesura depèn del grau de minimització dels errors aleatoris.
- L'**exactitud** (en anglès, *accuracy)* de la mesura depèn del grau de minimització dels errors sistemàtics.

Errors en els experiments de física





Il·lustració de la diferència entre precisió i exactitud. En a) tenim dades precises però inexactes. En b) tenim dades exactes, però imprecises. Els valors verdaders estan representats per la recta dibuixada.

Errors aleatoris i distribucions estadístiques

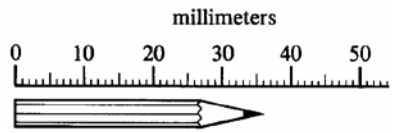
- Els errors aleatoris podem classificar-los en dos tipus:
  1. Errors aleatoris **instrumentals** (errors associats al mètode d'observació).
  2. Error aleatoris **estadístics** (errors associats a la natura aleatòria de l'observable).

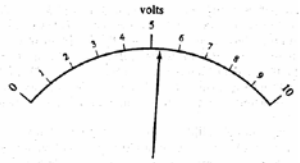
- Els primers són el resultat de les variacions de molts petits factors no controlats per l'experimentador i que poden canviar de manera impredictible d'una mesura a la següent, i donar lloc a resultats diferents. Donada la seua natura, aquest tipus d'error és estrictament inevitable, encara que podem reduir el seu valor augmentant el nombre d'observacions. Els resultats acumulats de moltes mesures subjectes a aquests **factors aleatoris** es distribueixen generalment de **forma gaussiana**.

- Els segons estan associats amb experiments de **recompte d'esdeveniments** de processos aleatoris en un interval de temps determinat. Els resultats acumulats de moltes mesures subjectes a aquestes fluctuacions estadístiques es distribueixen de **forma poissoniana**.

Errors aleatoris i distribucions estadístiques

- Els **errors instrumentals** són determinats generalment examinant els instruments i el procediment de mesura per a estimar la fiabilitat de les mesures.





La mesura és generalment expressada amb més o menys la meitat de la unitat més petita, i aquest número representa una estimació de la **desviació típica** associada a un experiment d'una única mesura.

- Si aquestes incerteses són gaussianes, el valor mesurat té una probabilitat **del 68% (nivell de confiança)** d'estar en l'interval d'1 σ al voltant del valor vertader.
- Podem utilitzar altres intervals equivalents a altres nivells de confiança (p. e. 95.5% de nivell de confiança que el valor mesurat estiga a 2 σ).

Errors aleatoris i distribucions estadístiques

- La desviació típica també pot ser obtinguda per repetició de la mesura i utilitzant

$$s = \sqrt{\frac{1}{N-1}\sum_i (x_i - \overline{x})^2}$$

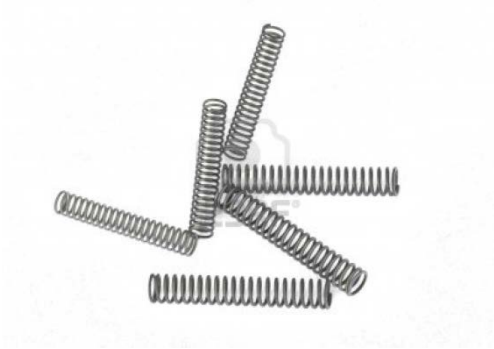
- Aquesta estimació es correspon amb l'error per a una única mesura (és a dir, si férem una única mesura, amb el mateix mètode, tindríem un 68% de probabilitat que el resultat obtingut estigués dins de l'interval [μ-1s ,μ+1s] al voltant del valor vertader).

- Ambdós mètodes haurien de donar resultats compatibles.

- Qualsevol discrepància suggeriria qualsevol problema, i hauríem de revisar rl procediment experimental.

Errors aleatoris i distribucions estadístiques

- **Exemple**: Imagina que tens una caixa de molls iguals i vols estimar les constants elàstiques de cada moll. Volem, per tant, estimar $k \pm \varepsilon_k$ per a cada un, però no volem fer mesures repetides per cada moll per a estalviar temps.



Podem seguir, la següent estratègia: Agafem el **primer moll** i mesurem **10 vegades** la seua constant elàstica (per exemple, mesurant el període de les seues oscil·lacions). Prenem el valor de *k* d'aquest primer moll com la mitjana dels valors individuals i calculem la desviació típica σ_k d'aquests valors. Aquest valor representa l'error associat a una única mesura.

Com que **la resta** de molls són molt semblants, esperem que l'error en la mesura de la resta de molls siga el mateix. Per tant, per a mesurar la resta dels molls, mesurem **una sola volta** cada constant elàstica i prenem com a error de cada un dels molls la desviació típica calculada amb el primer moll.

Per cada moll tindrem un valor $k_i \pm \sigma_k$ amb un nivell de confiança del 68%.

Errors aleatoris i distribucions estadístiques

- Els **errors aleatoris estadístics,** associats als experiments de recompte, es calculen de manera automàtica, ja que es distribueixen poissonianament i per tant la desviació típica:

$$\sigma = \sqrt{\mu}$$

- Si únicament tenim un valor x del recompte, tindrem com error d'aquesta única mesura:

$$\sigma = \sqrt{x}$$

Errors aleatoris i distribucions estadístiques

- Per què els errors aleatoris instrumentals són gaussians?

Teorema del límit central (CLT)

Considerem la suma X de N variables independents x_i obtingudes de distribucions de mitjana μ_i i variància V_i. La distribució de la variable X

a) té com valor esperat: $\langle X \rangle = \sum_i \mu_i$

a) té com variància $V(X) = \sum_i V_i$

a) tendeix a una gaussiana quan $N \to \infty$

Aquesta és la raó per la qual la gaussiana és tan important. Una quantitat produïda per l'efecte acumulatiu de moltes variables independents serà, almenys aproximadament, gaussiana, **independentment de les distribucions originals** de les variables originals.

(En les cues hi pot haver desviacions de la forma gaussiana).

Errors aleatoris i distribucions estadístiques

Demostració:

a) Mitjana:

Densitat de probabilitat conjunta

$$X = \sum_i x_i \Rightarrow \langle X \rangle = \left\langle \sum_i x_i \right\rangle = \int_S \left(\sum_i x_i \right) P(x_1, ..., x_N) dx_1 ... dx_N$$ Si són variables independents P és factoritzable i

$$\langle X \rangle = (\int_{-\infty}^{+\infty} x_1 P(x_1) dx_1 \int_{-\infty}^{+\infty} P(x_2) dx_2 ... \int_{-\infty}^{+\infty} P(x_N) dx_N) + ... + (\int_{-\infty}^{+\infty} P(x_1) dx_1 ... \int_{-\infty}^{+\infty} P(x_{N-1}) dx_{N-1} \int_{-\infty}^{+\infty} x_N P(x_N) dx_N) = \sum_i \langle x_i \rangle = \sum_i \mu_i$$

b) Variància

$$V(X) = \left\langle \left(X - \langle X \rangle \right)^2 \right\rangle = \left\langle \left(\sum_i x_i - \sum_i \mu_i \right)^2 \right\rangle = \left\langle \left[\sum_i (x_i - \mu_i) \right]^2 \right\rangle =$$

$$= \left\langle \sum_i (x_i - \mu_i)^2 \right\rangle + \left\langle \sum_i \sum_{j \neq i} (x_i - \mu_i)(x_j - \mu_j) \right\rangle =$$

$$= \sum_i \left\langle (x_i - \mu_i)^2 \right\rangle + \sum_i \sum_{j \neq i} \underbrace{\left\langle (x_i - \mu_i)(x_j - \mu_j) \right\rangle}_{\text{cov}(x_i, x_j) = 0 \text{ (si són variables independents)}} = \sum_i Var(x_i)$$

c) La prova en l'apèndix 2 del Barlow

Errors aleatoris i distribucions estadístiques

El teorema de límit central explicat amb un model senzill

TLC: *Una mesura sotmesa a molts petits errors aleatoris es distribuirà de forma normal.*

Imaginem que mesurem una quantitat continua *X* amb valor vertader μ. Suposarem que tenim n fonts independents d'errors aleatoris (paral·laxi, temps de reacció, etc.). Per a simplificar considerarem que tots són del mateix valor ε.

Cada font d'error pot augmentar o disminuir el nostre resultat una quantitat ε amb igual probabilitat p=1/2. Si tenim n fonts d'error, el nostre resultat podria estar entre [μ-nε ,μ+nε].

En una mesura particular, si tenim ν errors positius i (n-ν) negatius, el valor obtingut seria:

$$x = \mu + \nu\varepsilon - (n-\nu)\varepsilon = \mu + (2\nu - n)\varepsilon$$

On la variable discreta errors positius ν es distribueix seguint una binomial $B_{n,\frac{1}{2}}(\nu)$ amb mitjana *np=n/2* i variància *np(1-p)=n/4.*

Errors aleatoris i distribucions estadístiques

Per tant, la nostra variable $x = \mu + (2\nu - n)\varepsilon$ es distribueix simètricament al voltant de

$$\langle x \rangle = \langle \mu + (2\nu - n)\varepsilon \rangle = \mu + \left(2\langle \nu \rangle - n\right)\varepsilon = \mu + \left(2\frac{n}{2} - n\right)\varepsilon = \mu$$

Amb una desviació típica* $\sigma(x) = \left|\frac{dx}{d\nu}\right|\sigma_\nu = 2\varepsilon\sigma_\nu = \varepsilon\sqrt{n}$ (n $\to \infty$ i $\varepsilon \to 0$ de manera que $\varepsilon\sqrt{n}$ finita)

- Vegeu propagació dels errors

Quan n $\to \infty$ i $\varepsilon \to 0$ de manera que $\varepsilon\sqrt{n}$ finita X es distribueix com el límit de la binomial (distribució gaussiana) amb mitjana μ i desviació típica σ(x).

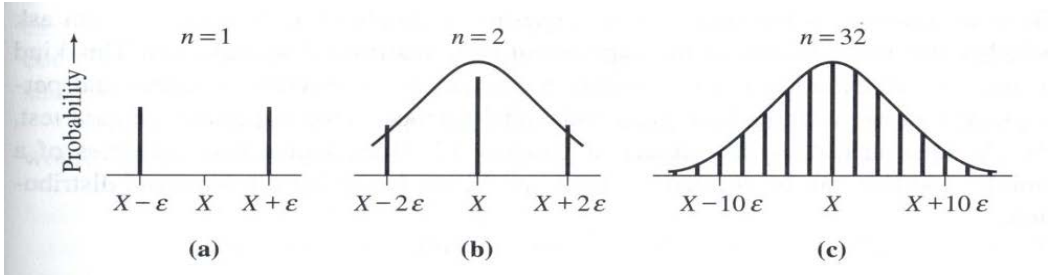


Figure 10.5. Distribution of measurements subject to n random errors of magnitude ε, for $n = 1$, 2, and 32. The continuous curves superimposed on (b) and (c) are Gaussians with the same center and width. (The vertical scales differ in the three graphs.)

Errors aleatoris i distribucions estadístiques

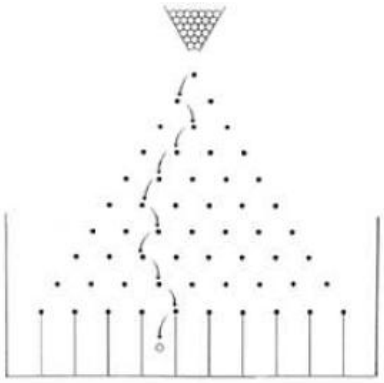
Màquina de Galton (Quincunx box)

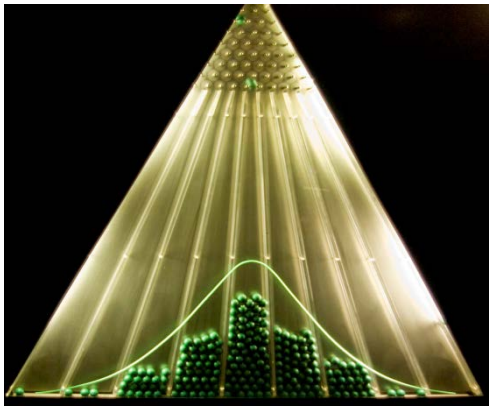
Si una boleta es desvia a la dreta v vegades, acabarà dipositada en la caixeta v-èsima comptant des de l'esquerra.

Si n és el nombre de files de claus, la probabilitat que una bola aparega en la caixeta v-èsima és igual a la probabilitat d'obtenir v desviacions cap a la dreta d'un total de n assaigs, que ve donada per la distribució binomial.

$$P[X = \nu] = \binom{n}{\nu} p^{\nu} q^{n-\nu}$$

Quan el nombre de cops és gran, tindrem una gaussiana.





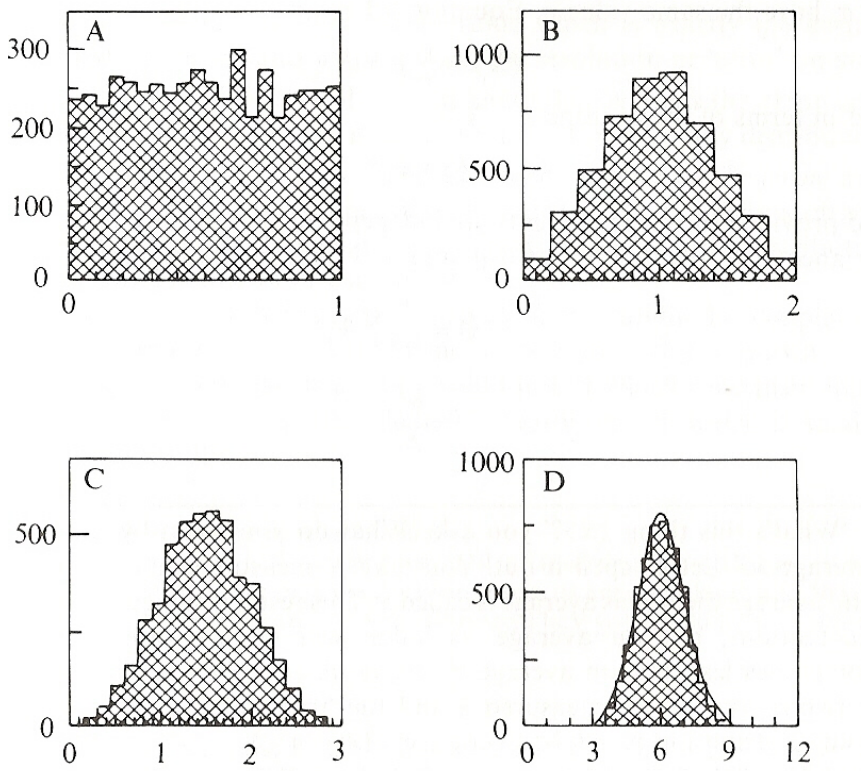
Errors aleatoris i distribucions estadístiques

A. Histograma de 5000 números obtinguts aleatòriament a partir d'una distribució uniforme entre 0 i 1. Mitjana 0.5 i variància 1/12.

B. 5000 punts obtinguts sumant números de dues distribucions uniformes del tipus A, i.e., $X=x_1+x_2$. La distribució resultant és triangular, centrada en 1.

C. Distribució obtinguda sumant tres números obtinguts de tres distribucions uniformes del tipus A. El pic està en 1.5 i la forma es corba.

D. Suma de 12 distribucions uniformes. Forma gaussiana de mitjana 6 i variància 1.



Error típic de la mitjana

- **Repetició de mesures:**
  - Suposem que mesurem el mateix observable N vegades. Podem utilitzar el *teorema del límit central* directament, per N distribucions (no necessàriament gaussianes) amb el mateix valor μ de les seues mitjanes i el mateix valor σ^2 de les variàncies. Per tant:

$$X = \sum_{i=1}^{N} x_i \Rightarrow \langle X \rangle = N\mu$$

$$\text{Com que } \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{X}{N} \quad \text{i per tant} \quad \langle \overline{x} \rangle = \left\langle \frac{X}{N} \right\rangle = \frac{1}{N}\langle X \rangle = \frac{1}{N} N\mu = \mu$$

  - A més a més, com la variància de la suma és: $\text{Var}(X)=N\sigma^2$

$$Var(\overline{x}) = Var\left(\frac{X}{N}\right) = \frac{1}{N^2} Var(X) = \frac{1}{N^2} N\sigma^2 = \frac{\sigma^2}{N} \Rightarrow \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}}$$

La desviació típica d'aquesta mitjana decreix amb $1/\sqrt{N}$!

Averaging is good for you!

Error típic de la mitjana

- **Interpretació:**

  Si prenem N mesures $x_1, x_2, \ldots, x_N$ i calculem la mitjana, tenim $\overline{x}$. Aquest resultat està sotmès a fluctuacions estadístiques, però el seu valor mitjà és μ . És a dir, $\overline{x}$ és una variable estadística descrita per una nova distribució amb valor esperat $\langle \overline{x} \rangle = \mu$ i variància $Var(\overline{x}) = \frac{\sigma^2}{N}$.

  - La quantitat $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}}$ es diu **error típic de la mitjana** *Standard deviation of the mean (SDOM)*

  En la figura es representa un histograma de 25 dades obtingudes aleatòriament a partir d'una distribució gaussiana**.** Es mostra la desviació típica de la mostra σ i l'error típic de la mitjana, que és la cinquena part de σ.

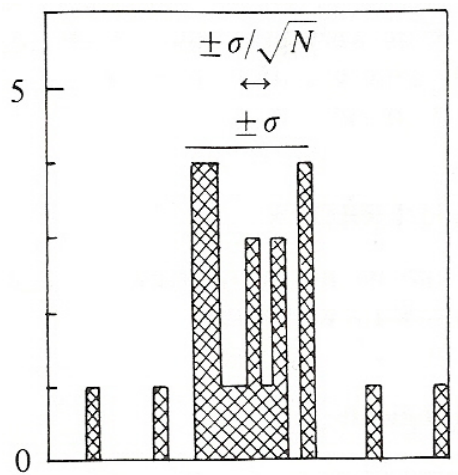


Fig. 4.2. Standard errors.

Error típic de la mitjana

- **Exemple 1: Mesures energètiques de raigs gamma.**
- La resolució energètica d'un detector de raigs γ utilitzat per a investigar la desintegració gamma d'un nucli radioactiu és de 50 KeV. Si únicament s'observa una desintegració, la incertesa serà de 50 keV. Si per contra, observem 100 desintegracions, millorem la incertesa fins a 5 keV. Per a arribar a 1keV hauríem d'observar 2500 desintegracions.

- **Exemple 2: Constant elàstica del moll**
  - Si mesurem la constant elàstica d'un moll 10 vegades i obtenim una mitjana k_1 i una desviació σ_k, el resultat de l'experiment donarà com a resultat:

$$k_1 \pm \frac{\sigma_k}{\sqrt{10}}$$

  - Si volem disminuir l'error en un factor 2, hauríem de mesurar 4 vegades més la constant del moll, és a dir, 40 vegades.

Propagació dels errors

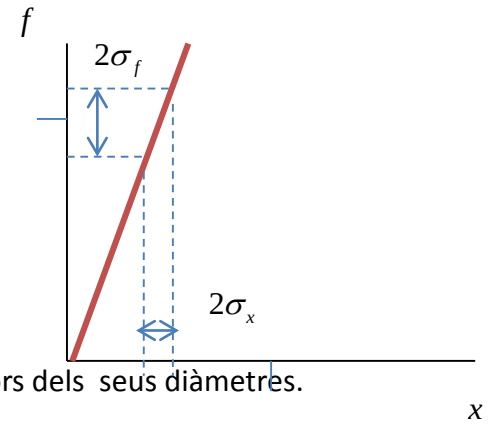
- Sovint volem determinar una magnitud f=f(X,Y,...) que és funció d'altres variables. Hem de conèixer com determinar l'error de f a partir dels errors d'aquestes variables . Aquest procés es diu **propagació d'errors.**
- **Una variable:** Suposem que f és una funció lineal de x: f=ax+b, on a i b són constants i x és una variable distribuïda amb variància V(x). La variància de f és:

$$\text{Si} \quad f = ax + b$$

$$Var(f) = \langle f^2 \rangle - \langle f \rangle^2 = \langle (ax+b)^2 \rangle - \langle ax+b \rangle^2 =$$

$$a^2 \langle x^2 \rangle + 2ab\langle x \rangle + b^2 - a^2 \langle x \rangle^2 - 2ab\langle x \rangle - b^2 =$$

$$= a^2 \left(\langle x^2 \rangle - \langle x \rangle^2 \right) = a^2 Var(x) \Rightarrow \sigma_f = |a| \sigma_x$$



- **Exemple:** Calculeu els valors dels radis de dues circumferències a partir dels valors dels seus diàmetres.

$$\phi_1 = (10.05 \pm 0.05)\text{mm}$$
$$\phi_2 = (19.95 \pm 0.05)\text{mm}$$

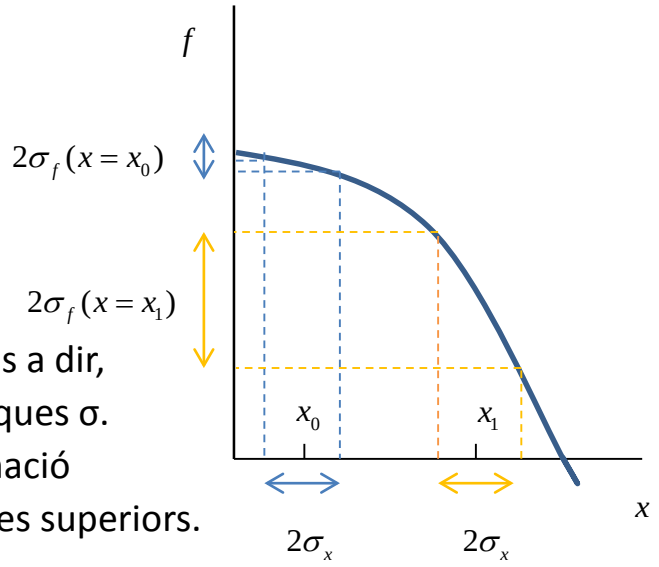
Propagació dels errors

- Considerem ara una funció f qualsevol, d'una variable aleatòria x. Si els errors de x són petits, els valors de la variable x es distribuiran al voltant del seu valor vertader. Si utilitzem l'expansió en sèrie de Taylor de la funció f, al voltant del valor vertader o valor mesurat x_0

$$f(x) \approx f(x_0) + (x - x_0)\left(\frac{df}{dx}\right)_{x=x_0}$$

- Aplicant el resultat anterior:

$$Var(f) \approx \left(\frac{df}{dx}\right)^2 Var(x) \Rightarrow \sigma_f = \left|\frac{df}{dx}\right|_{x_0} \sigma_x$$



- Aquesta expressió és vàlida quan els errors són petits, és a dir, la derivada primera no canvia molt en un interval d'unes poques σ.
- Si els errors de x són grans, no podem utilitzar l'aproximació en primer ordre de la sèrie de Taylor, i hauríem d'afegir ordres superiors.

Exemple: Calculeu les àrees dels cercles del problema anterior.

Propagació dels errors

- **Dues o més variables:** Suposem que f és una funció lineal de *x* i *y*: f=ax+by+c, on *a*, *b* i *c* són constants i x (y) és una variable distribuïda amb variància V(x) (V(y)). La variància de f és:

$$f = ax + by + c$$

$$Var(f) = \langle f^2 \rangle - \langle f \rangle^2 = \langle (ax+by+c)^2 \rangle - \langle ax+by+c \rangle^2 =$$

$$= a^2\left(\langle x^2 \rangle - \langle x \rangle^2\right) + b^2\left(\langle y^2 \rangle - \langle y \rangle^2\right) + 2ab\left(\langle xy \rangle - \langle x \rangle\langle y \rangle\right) =$$

$$= a^2 Var(x) + b^2 Var(y) + 2ab\,\mathrm{cov}(x,y)$$

- Per a funcions qualssevol de dues variables:

$$f = f(x,y) \approx f(x_0,y_0) + \left(\frac{\partial f}{\partial x}\right)_{x_0,y_0}(x-x_0) + \left(\frac{\partial f}{\partial y}\right)_{x_0,y_0}(y-y_0)$$

$$Var(f) = \left(\frac{\partial f}{\partial x}\right)^2_{x_0,y_0} Var(x) + \left(\frac{\partial f}{\partial y}\right)^2_{x_0,y_0} Var(y) + 2\left(\frac{\partial f}{\partial x}\right)_{\bar{x},\bar{y}}\left(\frac{\partial f}{\partial y}\right)_{x_0,y_0} \mathrm{cov}(x,y)$$

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2_{x_0,y_0} \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2_{x_0,y_0} \sigma_y^2 + 2\left(\frac{\partial f}{\partial x}\right)_{x_0,y_0}\left(\frac{\partial f}{\partial y}\right)_{x_0,y_0} \mathrm{cov}(x,y)$$

On l'últim terme pot ser positiu o negatiu. Les derivades calculades en els valors mesurats.

Propagació dels errors

- **Llei de propagació d'errors**

Per a qualsevol funció f de variables aleatòries x,y,z... Si es compleix que

1. Les variables són independents entre si.
2. Els errors de les distribucions de totes les variables són petits.

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial f}{\partial z}\right)^2 \sigma_z^2 + \ldots$$

- **<u>Aclariments:</u>**
- La independència de les variables x,y,z... ens assegura que les covariàncies s'anul·len, i, per tant, parlem de variables no correlacionades entre si (la major part de les mesures en experiments en física són independents). Les correlacions apareixen quan extraiem dos o més paràmetres del mateix conjunt de dades experimentals.
- Les derivades parcials s'avaluen en els valors mitjans de les variables x,y,z...
- Els errors han de ser petits per a poder aproximar la funció f de forma lineal.
- Els errors de les variables x,y,z... vénen definits a partir de les seues distribucions de probabilitat, és a dir, poden ser gaussians, poissonians, etc.

Fórmules d'errors específiques

- **Addicions i diferències simples**

$$f = x \pm a$$

$$\sigma_f^2 = \left(\frac{\partial (x \pm a)}{\partial x} \right)^2 \sigma_x^2 = \sigma_x^2 \Rightarrow \sigma_f = \sigma_x$$

- **Sumes i diferències pesades**

$$f = ax \pm by$$

$$\frac{\partial f}{\partial x} = a; \quad \frac{\partial f}{\partial y} = \pm b$$

$$\sigma_f^2 = a^2 \sigma_x^2 + b^2 \sigma_y^2 \pm 2ab \operatorname{cov}(x, y)$$

Fórmules d'errors específiques

- **Multiplicacions**

$$f = axy$$

$$\frac{\partial f}{\partial x} = ay; \quad \frac{\partial f}{\partial y} = ax$$

$$\sigma_f^2 = (ay\sigma_x)^2 + (ax\sigma_y)^2 + 2a^2 xy\,\mathrm{cov}(x,y)$$

  - En forma percentual l'expressió queda més simètrica:

$$\frac{\sigma_f^{\,2}}{f^2} = \frac{\sigma_x^{\,2}}{x^2} + \frac{\sigma_y^{\,2}}{y^2} + 2\frac{\mathrm{cov}(x,y)}{xy}$$

- **Divisions**

$$f = a\frac{x}{y}$$

$$\frac{\sigma_f^{\,2}}{f^2} = \frac{\sigma_x^{\,2}}{x^2} + \frac{\sigma_y^{\,2}}{y^2} - 2\frac{\mathrm{cov}(x,y)}{xy}$$

Fórmules d'errors específiques

- **Potències**

$$f = ax^b$$

$$\frac{\partial f}{\partial x} = abx^{b-1} = \frac{bf}{x}$$

$$\sigma_f = \left|\frac{bf}{x}\right|\sigma_x$$

$$\frac{\sigma_f}{f} = \left|b\frac{1}{x}\right|\sigma_x$$

Fórmules d'errors específiques

- **Exponencials**

$$f = ae^{\pm bx}$$

$$\frac{\partial f}{\partial x} = \pm abe^{\pm bx} = \pm bf$$

$$\sigma_f = |bf|\sigma_x$$

  - En forma percentual l'expressió queda més simètrica:

$$\frac{\sigma_f}{f} = \frac{|bf|}{f}\sigma_x = |b|\sigma_x$$

Formulació matricial de la propagació

- **Matriu de covariància:** Si tenim *n* variables aleatòries $x_{(1)}, x_{(2)}, \ldots x_{(n)}$ descrites per les seues funcions de probabilitat, definim la matriu de covariància **V** o matriu error, de manera semblant a la matriu de covariància definida per un conjunt de dades: és una matriu quadrada de *n*x*n* elements V_{ij}:

$$V_{ij} = \mathrm{cov}(x_{(i)}, x_{(j)}) = \left\langle x_{(i)} x_{(j)} \right\rangle - \mu_i \mu_j$$

- Els elements de la diagonal són les variàncies de cada variable: $V_{ii} == \left\langle x_{(i)} x_{(i)} \right\rangle - \mu_i \mu_i = \sigma_i^{\,2}$

- **Diverses funcions de diverses variables:** Si tenim *m* funcions $f_1, f_2, \ldots f_m$ de *n* variables aleatòries $x_{(1)}, x_{(2)}, \ldots x_{(n)}$, aquestes funcions tindran associades unes variàncies determinades per les variàncies de les variables $x_{(i)}$, i estaran correlacionades entre si per ser funcions de les mateixes variables:

$$Var(f_i) \equiv V(f_i) = \left\langle (f_i - \overline{f}_i)^2 \right\rangle$$
$$\mathrm{cov}(f_i, f_j) = \left\langle (f_i - \overline{f}_i)(f_j - \overline{f}_j) \right\rangle$$

Formulació matricial de la propagació

- **Propagació dels errors**
  - Desenvolupant les funcions f_i en sèrie de Taylor al voltat de les mitjanes de les x_j:

$$f_i \approx f_i(\mu_1, \mu_2, \ldots) + \left(\frac{\partial f_i}{\partial x_1}\right)(x_1 - \mu_1) + \left(\frac{\partial f_i}{\partial x_2}\right)(x_2 - \mu_2) + \ldots$$

  - Substituint en l'expressió de la variància $V(f_i) = \left\langle (f_i - \overline{f_i})^2 \right\rangle$ arribem a:

$$V(f_i) = \left\langle \left(\frac{\partial f_i}{\partial x_1}(x_1 - \mu_1) + \frac{\partial f_i}{\partial x_2}(x_2 - \mu_2) + \ldots \right)^2 \right\rangle = \left(\frac{\partial f_i}{\partial x_1}\right)^2 \left\langle (x_1 - \mu_1)^2 \right\rangle + \ldots + 2\left(\frac{\partial f_i}{\partial x_1}\right)\left(\frac{\partial f_i}{\partial x_2}\right)\left\langle (x_1 - \mu_1)(x_2 - \mu_2) \right\rangle \ldots =$$

$$= \sum_j \left(\frac{\partial f_i}{\partial x_j}\right)^2 V(x_j) + \sum_j \sum_{k \neq j} \left(\frac{\partial f_i}{\partial x_j}\right)\left(\frac{\partial f_i}{\partial x_k}\right) \operatorname{cov}(x_j, x_k)$$

Formulació matricial de la propagació

- **Propagació dels errors**
  - Substituint en l'expressió de les covariàncies $\mathrm{cov}(f_k, f_l) = \left\langle (f_k - \overline{f}_l)(f_k - \overline{f}_l) \right\rangle$ arribem a:

$$\mathrm{cov}(f_k, f_l) = \sum_i \sum_j \left(\frac{\partial f_k}{\partial x_i} \right) \left(\frac{\partial f_l}{\partial x_j} \right) \mathrm{cov}(x_i, x_j) \quad \text{amb} \quad \mathrm{cov}(f_i, f_i) \equiv V(f_i)$$

  - Si definim la matriu rectangular *m*x*n* $G_{ki} = \left(\frac{\partial f_k}{\partial x_i} \right)$ i denotem per $\mathbf{V_x}$ i $\mathbf{V_f}$ les matrius de covariància de *x* i *f* respectivament, podem escriure la relació matricial:

$$\mathbf{V}_f = \mathbf{G}\mathbf{V}_x\mathbf{G}^T$$

$$G = \begin{pmatrix} \frac{\partial f_1}{\partial x_{(1)}} & \cdots & \frac{\partial f_1}{\partial x_{(n)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_{(1)}} & \cdots & \frac{\partial f_m}{\partial x_{(n)}} \end{pmatrix}$$

on G^T és la matriu transposada de G. Amb aquesta notació:

$$\mathrm{cov}(f_k, f_l) = \left(V_f\right)_{kl} = G_{ki} V_{ij} G^T_{jl}$$

Formulació matricial de la propagació

$$\mathbf{V}_f = \mathbf{G}\mathbf{V}_x\mathbf{G}^T$$

$$\begin{pmatrix} \sigma_{f_1}^2 & \cdots & \mathrm{cov}(f_1, f_m) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(f_m, f_1) & \cdots & \sigma_{f_m}^2 \end{pmatrix}_{m\times m} =$$

$$= \begin{pmatrix} \frac{\partial f_1}{\partial x_{(1)}} & \cdots & \frac{\partial f_1}{\partial x_{(n)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_{(1)}} & \cdots & \frac{\partial f_m}{\partial x_{(n)}} \end{pmatrix}_{m\times n} \begin{pmatrix} \sigma_{x_1}^2 & \cdots & \mathrm{cov}(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(x_n, x_1) & \cdots & \sigma_{x_n}^2 \end{pmatrix}_{n\times n} \begin{pmatrix} \frac{\partial f_1}{\partial x_{(1)}} & \cdots & \frac{\partial f_m}{\partial x_{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_{(n)}} & \cdots & \frac{\partial f_m}{\partial x_{(n)}} \end{pmatrix}_{n\times m}$$

Estimadors de les distribucions

1. Mètode de màxima versemblança
    1. Estimador mitjana de la distribució de Gauss
    2. Estimador variància de la distribució de Gauss
    3. Ponderació
2. Màxima versemblança amb diverses variables

Bibliografia*:*

- *Statistics. A guide to the use of Statistical Methods in the Physical Sciences* (Barlow)
- *Data reduction and error analysis for the physical sciences* (Bevington, Robinson)

Mètode de màxima versemblança

- Ens plantegem el següent problema: obtenir a partir d'una mostra de dades els millors valors possibles de les propietats de la distribució d'origen de la mostra. Aquests valors es diuen **estimadors**.
- Un estimador ha de ser:
  - Consistent: si a' representa l'estimador del paràmetre a $\lim_{N\to\infty} a' = a$
  - No esbiaixat: $\langle a' \rangle = a$
  - Eficient: La seua variància ha de ser menuda.

Exemple: Volem trobar l'alçada mitjana de tots els estudiants de la UVEG a partir d'una mostra aleatòria de N estudiants. Suposem que el valor vertader és $H = \frac{1}{30000}\sum_1^{30000} h_i$.

Es proposen els següents estimadors:

$$h_1 = \frac{1}{N}\sum_{i=1}^{N} h_i \qquad h_2 = \frac{1}{10}\sum_{i=1}^{10} h_i \qquad h_3 = \frac{1}{N-1}\sum_{i=1}^{N} h_i \qquad h_4 = 1.8 \text{ m}$$

$$h_5 = \begin{cases} h_1 = \frac{1}{N/2}\sum_{i=1}^{N/2} h_{2i} & \text{si N parell} \\ h_1 = \frac{1}{(N+1)/2}\sum_{i=1}^{(N+1)/2} h_{2i-1} & \text{si N imparell} \end{cases}$$

Mètode de màxima versemblança

- **Paràmetres consistents**: h_1, h_3 i h_5
  - La *llei dels grans nombres* ens diu que les propietats d'una mostra, s'aproximen a les propietats de la població com més gran es fa la mostra, per tant, en el nostre cas,

$$\lim_{N\to 30000} h_1 = H; \quad \lim_{N\to 30000} h_2 = \frac{30000}{29999} H \qquad \lim_{N\to 30000} h_5 \approx H$$

- **Paràmetres sense biaix**: h_1, h_2 i h_5

$$\langle h_1 \rangle = \left\langle \tfrac{1}{N}\sum_{i=1}^{N} h_i \right\rangle = \tfrac{1}{N}\sum_{i=1}^{N}\langle h_i\rangle = \tfrac{1}{N}\sum_{i=1}^{N} H = \tfrac{1}{N} NH = \mathrm{H}$$

$$\langle h_2 \rangle = \left\langle \tfrac{1}{10}\sum_{i=1}^{10} h_i \right\rangle = \tfrac{1}{10}\sum_{i=1}^{10}\langle h_i\rangle = \tfrac{1}{10}\sum_{i=1}^{10} H = \mathrm{H}$$

$$\langle h_3 \rangle = \left\langle \tfrac{1}{N-1}\sum_{i=1}^{N} h_i \right\rangle = \tfrac{1}{N-1}\sum_{i=1}^{N}\langle h_i\rangle = \tfrac{1}{N-1}\sum_{i=1}^{N} H = \frac{N}{N-1}\mathrm{H} \neq \mathrm{H}$$

$$\langle h_4 \rangle = 1.8 \neq H$$

$$\langle h_5 \rangle = \begin{cases} \left\langle \tfrac{1}{N/2}\sum_{i=1}^{N/2} h_{2i} \right\rangle = \tfrac{1}{N/2}\sum_{i=1}^{N/2}\langle h_{2i}\rangle = \tfrac{1}{N/2}\tfrac{N}{2}\mathrm{H} = \mathrm{H} & \text{si N parell} \\ \left\langle \tfrac{1}{(N+1)/2}\sum_{i=1}^{(N+1)/2} h_{2i-1} \right\rangle = \tfrac{1}{(N+1)/2}\sum_{i=1}^{N+1/2}\langle h_{2i-1}\rangle = \tfrac{1}{(N+1)/2}\tfrac{N+1}{2}\mathrm{H} = \mathrm{H} & \text{si N imparell} \end{cases}$$

Mètode de màxima versemblança

– Eficiència: Pel *teorema del límit central*, com que h_5 utilitza la meitat dels valors, la seua variància és el doble que la de h_1:

$$V(h_5) = 2 \times V(h_1)$$

$\Rightarrow$ h_1 és més eficient que h_2

Conclusió: el millor estimador de l'alçada mitjana dels universitaris de la UVEG és $h_1 = \frac{1}{N}\sum_1^N h_i$.

Mètode de màxima versemblança

- **Mètode de màxima versemblança:**
  - El *principi de màxima versemblança* afirma que, a partir d'una mostra aleatòria de dades x_1, x_2,...x_N, el millor estimador de qualsevol paràmetre a de la distribució d'origen d'aquestes dades és el valor de a que maximitza la **funció versemblança**,

$$L(x_{1,}x_{2,}...,x_N;a) = P(x_1,a)\times...\times P(x_N,a) = \prod_{i=1}^{N} P(x_i,a)$$

essent $P(x_i,a)$ la probabilitat d'obtenir el valor x_i a partir de la distribució d'origen . Per tant, L representa la probabilitat d'observar la seqüencia de valors x_1, x_2,...x_N.

Mètode de màxima versemblança

- Treballarem primer amb la hipòtesi que les distribucions d'origen són **gaussianes**, la qual cosa és aplicable en la majoria de les situacions físiques.
  - Si la distribució d'origen és una gaussiana de mitjana μ i desviació típica σ, la funció de màxima versemblança **associada a la mitjana** és:

$$L(\mu') = \prod_{i=1}^{N} P(x_i, \mu') = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu'}{\sigma}\right)^2}$$

  - Multiplicant els N factors de probabilitats:

$$L(\mu') = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mu'}{\sigma}\right)^2\right]$$

Mètode de màxima versemblança

- Quan N→∞ qualsevol estimador consistent és no esbiaixat (límit asimptòtic).
  - En efecte,

$$\langle a' \rangle = \int a' L(a; x_1, x_{2,\ldots}) dx_1 dx_2 \ldots \xrightarrow{N\to\infty} \int a L(a; x_1, x_{2,\ldots}) dx_1 dx_2 \ldots = a \int L(a; x_1, x_{2,\ldots}) dx_1 dx_2 \ldots = a$$

- Provarem que, a més, $V^{-1}(a') = -\left[\frac{d^2 \ln L}{da^2}\right]_{a=a'}$
  - Suposarem que el valor vertader de a és a_0. L'estimador a' satisfà la relació $\left(\frac{d \ln L}{da}\right)_{a=a'} = 0$
  - Si l'estimador és consistent a'-a_0 és menut si N gran i podem escriure aquesta derivada en sèrie de Taylor al voltant de a_0 (a' difereix del valor vertader a_0 per ser la derivada de ln*L* en a_0 diferent de zero per causa de les fluctuacions).

$$\left(\frac{d \ln L}{da}\right)_{a=a'} = \left(\frac{d \ln L}{da}\right)_{a_0} + (a'-a_0)\left(\frac{d^2 \ln L}{da^2}\right)_{a_0} = 0 \;\rightarrow\; V(a'-a_0) = \frac{1}{\left(\frac{d^2 \ln L}{da^2}\right)^2_{a_0}} V\left(\frac{d \ln L}{da}\right) \quad (1)$$

  - La variància de la derivada de ln *L* és

$$V\left[\left(\frac{d \ln L}{da}\right)_{a_0}\right] = \left\langle \left(\frac{d \ln L}{da}\right)^2 \right\rangle - \left\langle \left(\frac{d \ln L}{da}\right) \right\rangle^2$$

Mètode de màxima versemblança

- Per la condició de normalització

$$\int L dx_1 \ldots = 1 \Rightarrow \int \frac{dL}{da} dx_1 \ldots = 0 \Rightarrow \int \frac{d \ln L}{da} L dx_1 \ldots = 0 \rightarrow \left\langle \left(\frac{d \ln L}{da} \right) \right\rangle = 0$$

$$\int \left(\frac{d^2 \ln L}{da^2} L + \int \frac{d \ln L}{da} \frac{dL}{da} \right) dx_1 \ldots = 0 \rightarrow \int \frac{d^2 \ln L}{da^2} L dx_1 \ldots + \int \left(\frac{d \ln L}{da} \right)^2 L dx_1 \ldots = 0 \rightarrow \left\langle \left(\frac{d^2 \ln L}{da^2} \right) \right\rangle = - \left\langle \left(\frac{d \ln L}{da} \right)^2 \right\rangle$$

I tindrem

$$V \left[\left(\frac{d \ln L}{da} \right)_{a_0} \right] = \left\langle \left(\frac{d \ln L}{da} \right)^2 \right\rangle - \left\langle \left(\frac{d \ln L}{da} \right) \right\rangle^2 = - \left\langle \left(\frac{d^2 \ln L}{da^2} \right) \right\rangle \qquad (2)$$

- De (1)

$$V(a' - a_0) = - \frac{1}{\left(\frac{d^2 \ln L}{da^2} \right)_{a_0}^2} \left\langle \left(\frac{d^2 \ln L}{da^2} \right) \right\rangle$$

- I si N és gran

$$\left\langle \left(\frac{d^2 \ln L}{da^2} \right) \right\rangle \xrightarrow{N \to \infty} \left(\frac{d^2 \ln L}{da^2} \right)_{a_0} \Rightarrow V(a') = V(a' - a_0) = - \left(\frac{d^2 \ln L}{da^2} \right)_{a_0}^{-1}$$ q.e.d.

Mètode de màxima versemblança

1. **<u>Estimador mitjana d'una gaussiana:</u>** Pel principi de màxima versemblança, hem de maximitzar L, que és equivalent a maximitzar el logaritme de L (utilitzem μ´ per ser la mitjana el paràmetre que fem variar; la σ està fixada).

$$\ln L = -\sum \frac{(x_i - \mu')^2}{2\sigma^2} - N \ln \sigma - \frac{N}{2} \ln 2\pi$$

- Per a trobar el valor de μ´ que fa màxim el valor del logaritme de L igualem la derivada a zero.

$$\frac{d}{d\mu'}\left[-\sum \frac{(x_i - \mu')^2}{2\sigma'^2} - N \ln \sigma - \frac{N}{2} \ln 2\pi \right] = 0 \Rightarrow$$

$$\Rightarrow \sum \frac{(x_i - \mu')}{\sigma^2} = 0$$

  - I com que σ és constant, obtenim com a estimador de la mitjana de la gaussiana:

$$\mu' = \frac{1}{N} \sum_i x_i = \bar{x}$$

Mètode de màxima versemblança

- Comprovem que aquest estimador és consistent. Per la *llei dels grans nombres*

$$\lim_{N\to\infty}\mu' = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}x_i = \lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{k}n_j x_j = \int_{-\infty}^{+\infty} xP(x)dx = \int_{-\infty}^{\infty} x\,G_{\mu,\sigma}(x)\,dx = \mu$$

ja que *per la llei dels grans nombres* $\frac{n_j}{N}\to P(x)$

- Comprovem que aquest estimador és no esbiaixat. Pel teorema del límit central:

$$\langle\mu'\rangle = \langle\overline{x}\rangle = \frac{\langle x\rangle + \ldots + \langle x\rangle}{N} = \frac{N\mu}{N} = \mu$$

- Del mateix teorema deduïm que la variància d'aquest estimador és,

$$V(\mu') = V\left[\frac{1}{N}(x_1 + \ldots + x_N)\right] = \frac{1}{N^2}N\sigma^2 = \frac{\sigma^2}{N}$$

Aquest últim resultat el podríem haver obtingut per propagació d'errors de l'expressió de μ´.

O també a partir de

$$\sigma_\mu{}^2 = \sum\left[\sigma_i{}^2\left(\frac{\partial\mu'}{\partial x_i}\right)^2\right] = \sum\left[\sigma_i{}^2\left(\frac{1}{N}\right)^2\right] = \frac{\sigma^2}{N}$$

$$V^{-1}(a') = -\left[\frac{d^2\ln L}{d\mu^2}\right]_{\mu=\mu'} = -\left[\frac{d}{d\mu}\left(\tfrac{1}{\sigma^2}\sum(x_i-\mu)\right)\right]_{\mu=\mu'} = \frac{N}{\sigma^2}$$

Mètode de màxima versemblança

2. **<u>Estimador variància d'una gaussiana:</u>** Suposem ara que tenim un conjunt de mesures amb mitjana μ coneguda i desviació σ desconeguda. La funció versemblança de paràmetre σ´ és:

$$L(\sigma') = \left(\frac{1}{\sigma'\sqrt{2\pi}}\right)^N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma'}\right)^2\right]$$

– Per a maximitzar derivem el logaritme de *L* igualat a zero: $\frac{d\ln L}{d\sigma'} = 0$

$$\frac{d}{d\sigma'}\left[-\sum\frac{(x_i-\mu)^2}{2\sigma'^2} - N\ln\sigma' - \frac{N}{2}\ln 2\pi\right] = 0$$

– I arribem a $$\sum\frac{(x_i-\mu)^2}{\sigma'^3} - \frac{N}{\sigma'} = 0$$

– I, per tant, l'estimador de la variància, coneguda la mitjana, és:

$$V'(x) = \sigma'^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i-\mu)^2$$

Mètode de màxima versemblança

- En la pràctica no coneixem la mitjana; podem reemplaçar-la pel seu estimador $\overline{x}$.

$$V'(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i^{\,2} - \overline{x}^2\right) \qquad (1)$$

- Però en aquesta aproximació estem esbiaixant l'estimador. En efecte,

$$\left\langle V'(x)\right\rangle = \left\langle \frac{1}{N}\sum_{i=1}^{n}(x_i^{\,2} - \overline{x}^2)\right\rangle = \left\langle x^2\right\rangle - \left\langle \overline{x}^2\right\rangle$$

- Pel TLC $\langle \overline{x}\rangle = \mu$ i $\mu = \langle x\rangle$ i podem escriure

$$\left\langle V'(x)\right\rangle = \left\langle x^2\right\rangle - \left\langle x\right\rangle^2 - \left\langle \overline{x}^2\right\rangle + \left\langle \overline{x}\right\rangle^2 = V(x) - V(\overline{x})$$

- A més a més pel TLC $V(\overline{x}) = \frac{V(x)}{N}$ i arribem a la desigualtat que mostra que (1) està esbiaixat:

$$\left\langle V'(x)\right\rangle = V(x) - V(\overline{x}) = \frac{N-1}{N}V(x) \neq V(x)$$

Mètode de màxima versemblança

– L'estimador variància en la forma correcta seria (*correcció de Bessel*):

$$V'(x) = \frac{N}{N-1} \times \frac{1}{N} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$V'(x) \equiv s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

(Variància mostral de la gaussiana)

– s^2 representa un estimador consistent i no esbiaixat de la variància i per tant el més adequat.

– La variància de σ és: $V^{-1}(\sigma') = -\left[\frac{d^2 \ln L}{d\sigma^2}\right]_{\sigma=\sigma'} = -\left[\frac{d}{d\sigma}\left(\frac{1}{\sigma^3}\sum (x_i - \mu)^2 - \frac{N^2}{\sigma}\right)\right]_{\sigma=\sigma'} = -\frac{2N}{\sigma'^2}$

i per tant $\sigma(\sigma') = \frac{\sigma'}{\sqrt{2N}}$ i anàlogament $\sigma(s) = \frac{s}{\sqrt{2(N-1)}}$

– I per propagació d'errors arribem a $V(\sigma'^2) = \frac{2\sigma'^4}{N}$ i $V(s^2) = \frac{2s^4}{N-1}$

Mètode de màxima versemblança

3. **<u>Ponderació:</u>** Imaginem ara que els diferents punts de la mostra obtinguda experimentalment han sigut mesurats amb diferents precisions, i per tant partim de distribucions d'origen amb la mateixa mitjana però diferents desviacions. La funció de màxima versemblança és:

$$L(\mu') = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu'}{\sigma_i}\right)^2}$$

– Maximitzant ln *L*, tindrem

$$\frac{d}{d\mu'}\left[-\sum \frac{(x_i - \mu')^2}{2\sigma_i^{\,2}} - N \ln \sigma_i - \frac{N}{2} \ln 2\pi \right] = 0 \Rightarrow \frac{d}{d\mu'} \sum \frac{(x_i - \mu')^2}{2\sigma_i^{\,2}} = \sum \frac{(x_i - \mu')}{\sigma_i^{\,2}} = 0$$

– I la **mitjana ponderada** serà:

$$\mu' = \frac{\sum \left(x_i / \sigma_i^{\,2} \right)}{\sum \left(1 / \sigma_i^{\,2} \right)}$$

Mètode de màxima versemblança

- **Error de la mitjana ponderada:** Per propagació dels errors en la fórmula de la mitjana ponderada, com que

$$\frac{\partial \mu'}{\partial x_i} = \frac{\partial}{\partial x_i} \frac{\sum\left(x_i / \sigma_i^{\,2}\right)}{\sum\left(1/\sigma_i^{\,2}\right)} = \frac{\left(1/\sigma_i^{\,2}\right)}{\sum\left(1/\sigma_i^{\,2}\right)}$$

$$\sigma_{\mu'}^{\,2} = \sum\left[\sigma_i^{\,2}\left(\frac{\partial \mu'}{\partial x_i}\right)^2\right] = \sum\left[\sigma_i^{\,2}\left(\frac{1/\sigma_i^{\,2}}{\sum(1/\sigma_i^{\,2})}\right)^2\right] = \frac{\sum(1/\sigma_i^{\,2})}{\left[\sum(1/\sigma_i^{\,2})\right]^2}$$

I finalment,

$$\sigma_{\mu'}^{\,2} = \frac{1}{\sum(1/\sigma_i^{\,2})}$$

Màxima versemblança amb diverses variables

- A partir d'una mostra aleatòria de *N* dades experimentals $x_1, x_2, ..., x_N$ volem estimar els millors *n* valors $a_1, a_2, ..., a_n$ que defineixen la distribució d'origen d'aquelles dades.
- Construïm la funció de màxima versemblança $L = L(x_1, x_2, ..., x_N; a_1, a_2, ..., a_n)$ i imposem les condicions de màxim, a partir de les quals obtindrem els estimadors $a_1, a_2, ..., a_n$

$$\frac{\partial \ln L(x_1, x_2, ..., x_N; a_1, a_2, ..., a_n)}{\partial a_j} = 0 \qquad j = 1, ..., n$$

- Es pot demostrar que la inversa de la matriu de covariància ve determinada per les equacions:

$$V_{ij}^{-1} = \mathrm{cov}^{-1}\left(a_i, a_j\right) = -\left[\frac{\partial^2 \ln L}{\partial a_i \partial a_j}\right]_{a=a'}$$

– En particular

$$V^{-1}(a_i) = -\left[\frac{\partial^2 \ln L}{\partial^2 a_i}\right]_{a=a'}$$

Mínims quadrats

1. Mètode de mínims quadrats
2. Ajust d'una recta
   I. Pendent i ordenada
   II. Errors i covariància
   III. Expressió del χ^2 en funció dels errors
   IV. Ajust ponderat de la recta
   V. Extrapolació
   VI. Linealització
3. La distribució χ^2
4. Formulació matricial dels mínims quadrats
5. Ajusts de funcions no lineals en els paràmetres

Bibliografia*:*

- *Statistics. A guide to the use of Statistical Methods in the Physical Sciences* (Barlow)
- *Data reduction and error analysis for the physical sciences* (Bevington, Robinson)
- *Análisis de errores* (C. Sánchez del Río)

Mètode de mínims quadrats

- El **mètode de mínims quadrats**, com el mètode de màxima versemblança, és un procediment que ens permet determinar un conjunt de paràmetres desconeguts a partir d'un conjunt de dades.

- En la seua forma bàsica s'utilitza quan tenim dues variables x i y on:
  - Els valors de x són coneguts amb molta precisió.
  - Els valors de y són coneguts amb precisions σ.
  - Tenim una relació *f(x;a)* que prediu el valor y per a qualsevol valor x, però que conté com a paràmetre desconegut *a*, el qual hem de determinar.

Mètode de mínims quadrats

- El *principi de mínims quadrats* pot derivar-se del *principi de màxima versemblança*, si les mesures es distribueixen de forma gaussiana. En efecte, la probabilitat de tenir un valor particular y_i per un valor donat x_i és

$$P(x_i;a)=\left(\frac{1}{\sigma_i\sqrt{2\pi}}\right)\exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i-f(x_i;a}{\sigma_i}\right)^2\right]$$

- El logaritme de la funció versemblança és, doncs,

$$\ln L=-\frac{1}{2}\sum\left[\frac{y_i-f(x_i;a)}{\sigma_i}\right]^2-N\ln\sigma_i-\frac{N}{2}\ln 2\pi$$

- I maximitzar aquesta funció equival a **minimitzar** la suma ponderada dels quadrats de les diferències entre les dades i els valors predits, és a dir,

$$\sum_i\left[\frac{y_i-f(x_i;a)}{\sigma_i}\right]^2$$

Mètode de mínims quadrats

- També pot plantejar-se com un principi obvi, que no necessita més justificants, que el mateix significat d'aquest estimador: es fa variar el paràmetre *a* per a obtenir els valors predits $f(x_i)$ tan pròxims com siga possible als valors experimentals y_i (els quadrats comporten eliminar les grans desviacions). L'estimador rep el nom de χ^2

$$\chi^2 = \sum_i \left[\frac{y_i - f(x_i;a)}{\sigma_i} \right]^2$$

- La minimització exigeix trobar la solució *a* de l'equació

$$\frac{d\chi^2}{da} = 0 \Rightarrow \sum_i \frac{1}{\sigma_i^2} \frac{df(x_i;a)}{da} \left[y_i - f(x_i;a) \right] = 0$$

- El valor així estimat —denotat per a´— serà pròxim al valor verdader. La variància d'aquest estimador calculat en funció dels valors y_i , podem calcular-la amb la fórmula de propagació d'errors,

$$V(a') = \sum_i \left(\frac{\partial a'}{\partial y_i} \right)^2 V(y_i)$$

Mètode de mínims quadrats

- **Exemple: proporcionalitat directa** $y = mx$
- La quantitat a minimitzar és $\chi^2 = \sum_i \left[\frac{y_i - mx_i}{\sigma_i} \right]^2$
- Derivem respecte a *m* $\frac{d\chi^2}{dm} = \sum_i -2x_i \frac{[y_i - mx_i]}{\sigma_i^2}$
- Si els errors σ_i **són tots iguals**, els podem traure factor comú i tenim

$$\frac{d\chi^2}{dm} = \frac{2}{\sigma^2} \sum_i \left(x_i y_i - m x_i^2 \right)$$

- Igualem a zero i queda

$$\sum_i \left(x_i y_i - m' x_i^2 \right) = 0 \Rightarrow \sum_i x_i y_i = m' \sum_i x_i^2$$

- I, per tant, la constant de proporcionalitat és

$$m' = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{\overline{xy}}{\overline{x^2}}$$

Mètode de mínims quadrats

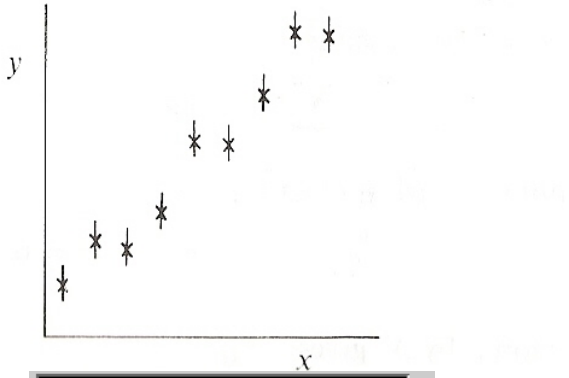
- Per a calcular la variància utilitzem $V(a') = \sum_i \left(\frac{\partial a'}{\partial y_i} \right)^2 V(y_i)$
- Partint de
$$m' = \sum_i \frac{x_i}{N\overline{x^2}} y_i$$
- Tindrem
$$V(m') = \sum_i \left(\frac{\partial m'}{\partial y_i} \right)^2 V(y_i) = \sum_i \left(\frac{x_i}{N\overline{x^2}} \right)^2 \sigma^2$$
- És a dir, la variància de la constant queda:

$$V(m') = \frac{\sigma^2}{N\overline{x^2}}$$

Ajust d'una recta

- Considerem ara un conjunt de mesures (x_i , y_i), totes les *Y* **amb el mateix error σ** i que presumiblement tenen una **dependència lineal** amb la variable *X*, és a dir, els punts es distribueixen al voltant d'una línia recta de pendent *m i* ordenada en l'origen n:

$$y = mx + n$$

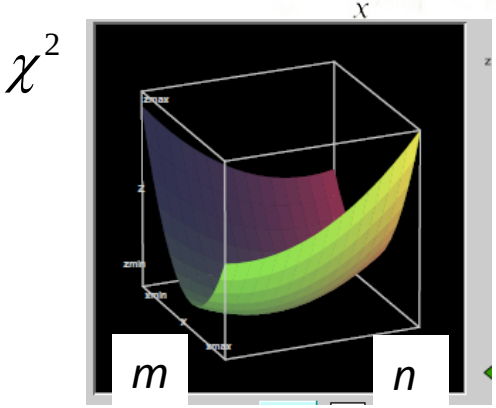


- La funció χ^2 és ara $\chi^2 = \sum_i \left[\frac{y_i - mx_i - n}{\sigma} \right]^2$

  i conté dos paràmetres a determinar, *m* i *n*,

  i per tant la minimització es fa mitjançant dues equacions, una per cada variable:

$$\frac{\partial \chi^2}{\partial m} = 0 \qquad \text{i} \qquad \frac{\partial \chi^2}{\partial n} = 0$$



- Derivant respecte de *n* tenim:

$$\sum_i 2\left(y_i - m' x_i - n'\right) = 0 \xrightarrow{\times \frac{1}{N}} \overline{y} - m'\overline{x} - n' = 0$$

Ajust d'una recta

- Derivant respecte de *m* tenim:

$$\sum_i -2x_i\left(y_i - m'x_i - n'\right) = 0 \xrightarrow{\times\frac{1}{N}} \overline{xy} - m'\overline{x^2} - n'\overline{x} = 0$$

- Tenim, doncs, un sistema de dues equacions amb dues incògnites *m´* i *n´*:

$$\begin{cases} \overline{x}m' + n' = \overline{y} \\ \overline{x^2}m' + \overline{x}n' = \overline{xy} \end{cases}$$

- Si aïllem *n´* en la primera i substituïm en la segona:

$$m' = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\mathrm{cov}(x,y)}{V(x)}$$

- I l'ordenada $n' = \overline{y} - m'\overline{x}$ (la recta passa pel punt $(\overline{x}, \overline{y})$, centre de gravetat dels valors x_i, y_i)

- O substituint el valor de m´

$$n' = \frac{\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2}$$

Ajust d'una recta

Error del pendent *m*

- Podem escriure el pendent de la següent manera:

$$m' = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \sum_i \frac{(x_i - \overline{x})}{N\left(\overline{x^2} - \overline{x}^2\right)} y_i$$

- La propagació d'errors ens dóna:

$$V(m') = \sum_i \left(\frac{\partial m'}{\partial y_i}\right)^2 V(y_i) = \sum_i \left[\frac{(x_i - \overline{x})}{N\left(\overline{x^2} - \overline{x}^2\right)}\right]^2 \sigma^2 = \frac{\sum_i \left(x_i^2 - 2\overline{x}x_i + \overline{x}^2\right)}{N^2\left(\overline{x^2} - \overline{x}^2\right)^2}\sigma^2 = \frac{N\left(\overline{x^2} - \overline{x}^2\right)}{N^2\left(\overline{x^2} - \overline{x}^2\right)^2}\sigma^2$$

- Que es pot expressar com a:

$$V(m') = \frac{\sigma^2}{N\left(\overline{x^2} - \overline{x}^2\right)}$$

Ajust d'una recta

Errors de l'ordenada *n*

- Podem escriure l'ordenada de la següent manera:

$$n' = \frac{\overline{x^2}\,\overline{y} - \overline{x}\;\overline{xy}}{\overline{x^2} - \overline{x}^2} = \sum_i \frac{\left(\overline{x^2} - \overline{x}x_i\right)}{N\left(\overline{x^2} - \overline{x}^2\right)} y_i$$

- La propagació d'errors ens dóna:

$$V(n') = \sum_i \left(\frac{\partial n'}{\partial y_i}\right)^2 V(y_i) = \sum_i \left[\frac{\overline{x^2} - \overline{x}x_i}{N\left(\overline{x^2} - \overline{x}^2\right)}\right]^2 \sigma^2 = \frac{\sigma^2}{N\left(\overline{x^2} - \overline{x}^2\right)^2} \frac{1}{N} \sum_i \left[\left(\overline{x^2}\right)^2 - 2\overline{x^2}\;\overline{x}x_i + \overline{x}^2 x_i^2\right] = \frac{\sigma^2}{N\left(\overline{x^2} - \overline{x}^2\right)^2}\left(\overline{x^2}\right)\left(\overline{x^2} - \overline{x}^2\right)$$

- I finalment:

$$V(n') = \frac{\sigma^2 \overline{x^2}}{N\left(\overline{x^2} - \overline{x}^2\right)}$$

Ajust d'una recta

Estimació dels errors a partir dels valors de y

– Es pot estimar la desviació de les mesures de y_i si considerem que estan distribuïdes de forma gaussiana respecte del valor verdader mx_i+n. La funció de versemblança serà:

$$L(\sigma') = \left(\frac{1}{\sigma'\sqrt{2\pi}}\right)^N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i - mx_i - n}{\sigma'}\right)^2\right]$$

– I la condició de màxima versemblança ens donaria:

$$\frac{d}{d\sigma'}\left[-\sum\frac{(y_i - mx_i - n)^2}{2\sigma'^2} - N\ln\sigma' - \frac{N}{2}\ln 2\pi\right] = 0 \Rightarrow \sigma'^2 = \frac{1}{N}\sum_{i=1}^{n}(y_i - mx_i - n)^2$$

– Si substituïm els valors verdaders m i n pels seus estimadors m' i n' hem de corregir l'estimació de σ, i es pot provar que finalment la millor estimació de l'error de y és:

$$\sigma'^2 = \frac{1}{N-2}\sum_{i=1}^{n}(y_i - mx_i - n)^2$$

Estimador de la desviació típica (*standard error of the estimate* $s_{y/x}$)

– La comparació d'aquest valor σ´ i l'estimació dels errors feta directament σ, és un test sobre la hipòtesi de linealitat entre *x* i *y.*

Ajust d'una recta

Covariància i coeficient de correlació de *m* i *n*

- Com que

$$\overline{y} = m'\overline{x} + n' \Rightarrow \sigma_{\overline{y}}^2 = \left(\frac{\partial \overline{y}}{\partial m'}\right)^2 \sigma_{m'}^2 + \left(\frac{\partial \overline{y}}{\partial n'}\right)^2 \sigma_{n'}^2 + 2\left(\frac{\partial \overline{y}}{\partial m'}\right)\left(\frac{\partial \overline{y}}{\partial n'}\right)\mathrm{cov}(m', n')$$

- I tenint en compte que

$$\overline{y} = \frac{1}{N}\sum_i y_i \rightarrow \sigma_{\overline{y}}{}^2 = \frac{\sigma^2}{N}; \quad \frac{\partial \overline{y}}{\partial m'} = \overline{x}; \quad \frac{\partial \overline{y}}{\partial n'} = 1; \quad \sigma^2(m') = \frac{\sigma^2}{N\left(\overline{x^2} - \overline{x}^2\right)}; \quad \sigma^2(n') = \frac{\sigma^2 \overline{x^2}}{N\left(\overline{x^2} - \overline{x}^2\right)}$$

- Resulta:

$$\mathrm{cov}(m', n') = \frac{1}{2\overline{x}}\left[\frac{\sigma^2}{N} - \overline{x}^2 \frac{\sigma^2}{N\left(\overline{x^2} - \overline{x}^2\right)} - \frac{\sigma^2 \overline{x^2}}{N\left(\overline{x^2} - \overline{x}^2\right)}\right]$$

- I finalment:

$$\mathrm{cov}(m', n') = -\frac{\sigma^2 \overline{x}}{N\left(\overline{x^2} - \overline{x}^2\right)}$$

$$\rho\left(m', n'\right) = \frac{\mathrm{cov}(m', n')}{\sigma_{m'}\sigma_{n'}} = -\frac{\overline{x}}{\sqrt{\overline{x^2}}}$$

Ajust d'una recta

Ajust ponderat de la recta

- **Si els errors** σ_i dels punts experimentals **no són iguals**, la funció a minimitzar és:

$$\chi^2 = \sum_i \left[\frac{y_i - m x_i - n}{\sigma_i} \right]^2$$

I tindrem

$$\begin{cases} \dfrac{\partial \chi^2}{\partial n} = 0 \Rightarrow -\sum_i 2\left(\dfrac{y_i - m' x_i - n'}{\sigma_i} \right) \dfrac{1}{\sigma_i} = 0 \rightarrow \sum_i \dfrac{y_i}{\sigma_i^2} - m' \sum_i \dfrac{x_i}{\sigma_i^2} - n' \sum_i \dfrac{1}{\sigma_i^2} = 0 \\ \dfrac{\partial \chi^2}{\partial m} = 0 \Rightarrow \sum_i -2\left(\dfrac{y_i - m' x_i - n'}{\sigma_i} \right) \dfrac{x_i}{\sigma_i} = 0 \rightarrow \sum_i \dfrac{x_i y_i}{\sigma_i^2} - m' \sum_i \dfrac{x_i x_i}{\sigma_i^2} - n' \sum_i \dfrac{x_i}{\sigma_i^2} = 0 \end{cases}$$

- Són les mateixes equacions del problema sense pesos. Podem trobar les solucions fent els canvis:

$$\begin{cases} \dfrac{1}{N}\sum_i x_i \rightarrow \dfrac{1}{\sum_i \frac{1}{\sigma_i^2}} \sum_i \dfrac{x_i}{\sigma_i^2}; \quad \dfrac{1}{N}\sum_i y_i \rightarrow \dfrac{1}{\sum_i \frac{1}{\sigma_i^2}} \sum_i \dfrac{y_i}{\sigma_i^2}; \quad \sum_i x_i y_i \rightarrow \sum_i \dfrac{x_i y_i}{\sigma_i^2} \\ \text{i per les variàncies:} \quad \sigma^2 \rightarrow \overline{\sigma^2} = \dfrac{\sum_i \frac{\sigma_i^2}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}} = \dfrac{N}{\sum_i \frac{1}{\sigma_i^2}} \end{cases}$$

Ajust d'una recta

Fent els canvis, tindríem:

$$\text{definint} \quad \Delta \equiv \left(\sum_i \frac{1}{\sigma_i^2}\right)\sum_i \frac{x_i^2}{\sigma_i^2} - \left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2$$

$$\text{Pendent} \rightarrow \quad m' = \frac{1}{\Delta}\left(\sum_i \frac{1}{\sigma_i^2}\sum_i \frac{x_i y_i}{\sigma_i^2} - \sum_i \frac{x_i}{\sigma_i^2}\sum_i \frac{y_i}{\sigma_i^2}\right)$$

$$\text{Ordenada} \rightarrow \quad n' = \frac{1}{\sum_i \frac{1}{\sigma_i^2}}\left(\sum_i \frac{y_i}{\sigma_i^2} - m'\sum_i \frac{x_i}{\sigma_i^2}\right)$$

substituint m' s'obté:

$$n' = \frac{1}{\Delta}\left(\sum_i \frac{x_i^2}{\sigma_i^2}\sum_i \frac{y_i}{\sigma_i^2} - \sum_i \frac{x_i}{\sigma_i^2}\sum_i \frac{x_i y_i}{\sigma_i^2}\right)$$

$$V(m') = \frac{1}{\Delta}\sum_i \frac{1}{\sigma_i^2}$$

$$V(n') = \frac{1}{\Delta}\sum_i \frac{x_i^2}{\sigma_i^2}$$

$$\text{cov}(m', n') = -\frac{1}{\Delta}\sum_i \frac{x_i}{\sigma_i^2}$$

Ajust d'una recta

Exemple 1: Un estudiant està investigant la llei $1/r^2$; amb un comptador Geiger compta el nombre de partícules detectades a diferents distàncies. Com que el sistema és automàtic el recompte es fa cada 15 segons. Per a aquesta anàlisi es registren 30 adquisicions de 15 segons en cada posició, i se sumen els resultats (*Bevington, ex. 6.2*)

Number of counts detected in $7\frac{1}{2}$-min intervals as a function of distance from the source†

i	Distance d_i (m)	$x_i = 1/d_i^2$ (m^{-2})	Counts C_i	σ_{C_i}	Weight $(1/C_i^2)$ w_i	$w_i x_i$	$w_i C_i$	$w_i x_i^2$	$w_i x_i C_i$	Fitted counts $a + bx_i$
1	0.20	25.00	901	30.0	0.00111	0.0278	1	0.694	25.0	887
2	0.25	16.00	652	25.5	0.00153	0.0254	1	0.393	16.0	610
3	0.30	11.11	443	21.0	0.00226	0.0251	1	0.279	11.1	461
4	0.35	8.16	339	18.4	0.00295	0.0241	1	0.197	8.2	370
5	0.40	6.25	283	16.8	0.00353	0.0221	1	0.138	6.3	311
6	0.45	4.94	281	16.8	0.00356	0.0176	1	0.087	4.9	271
7	0.50	4.00	240	15.5	0.00417	0.0167	1	0.067	4.0	242
8	0.60	2.78	220	14.8	0.00455	0.0126	1	0.035	2.8	205
9	0.75	1.78	180	13.4	0.00556	0.0099	1	0.018	1.8	174
10	1.00	1.00	154	12.4	0.00649	0.0065	1	0.007	1.0	150
Sums					0.03570	0.1868	10	1.912	81.0	

$\sigma_i = \sqrt{y_i} \qquad w_i = 1/\sigma_i^2 = 1/y_i$

$\Delta = \Sigma w_i \Sigma w_i x_i^2 - (\Sigma w_i x_i)^2 = 0.03570 \times 1.912 - (0.1868)^2 = 0.0334$

$a = [\Sigma w_i C_i \Sigma w_i x_i^2 - \Sigma w_i x_i \Sigma w_i x_i C_i]/\Delta = [10 \times 1.912 - 0.1868 \times 81.0]/\Delta = 119.5$

$b = [\Sigma w_i \Sigma w_i x_i C_i - \Sigma w_i x_i \Sigma w_i C_i]/\Delta = [0.03570 \times 81.0 - 0.1868 \times 10]/\Delta = 30.7$

$\sigma_a^2 \simeq \Sigma w_i x_i^2/\Delta = 1.912/0.0334 = 57.3 \qquad \sigma_a \simeq 7.6$

$\sigma_b^2 \simeq \Sigma w_i/\Delta = 0.03570/0.0334 = 1.07 \qquad \sigma_b \simeq 1.1$

†A linear fit to the data of the function $C = a + bx$ by the method of determinants gives $a = 119 \pm 8$ and $b = 31 \pm 1$, with $\chi^2 = 11.1$ for 8 degrees of freedom. The χ^2 probability for the fit is about 20%.

Els errors vénen determinats per l'estadística poissoniana i es calculen com l'arrel del nombre de partícules detectades.

En aquest exemple la recta ajustada és:

$$y = a + bx$$

$$\mathrm{cov}(b,a) = -\frac{\sum w_i x_i}{\Delta} = -5.6$$

Ajust d'una recta

Representació gràfica de l'exemple 1.

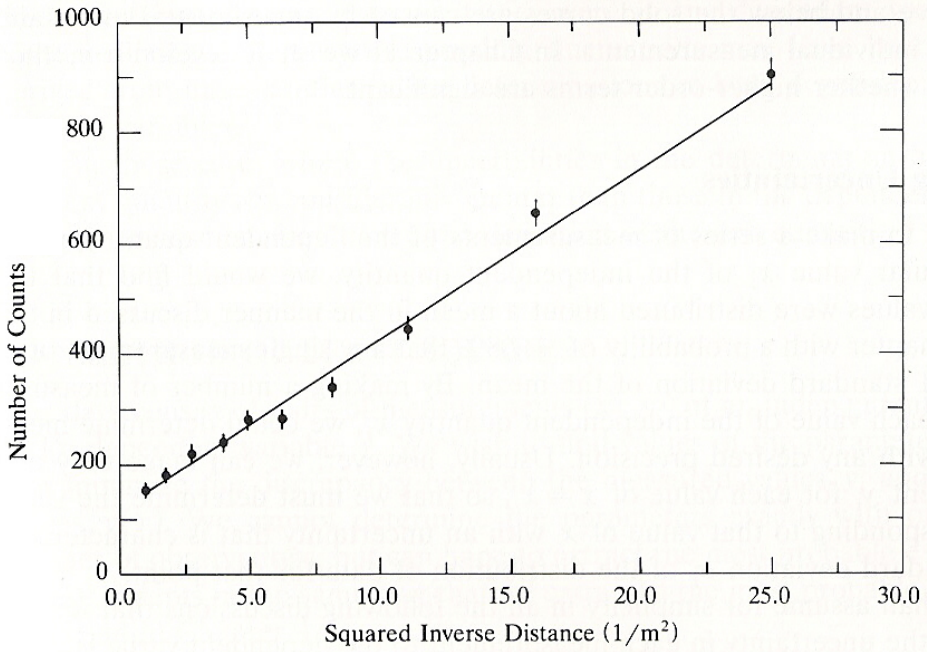
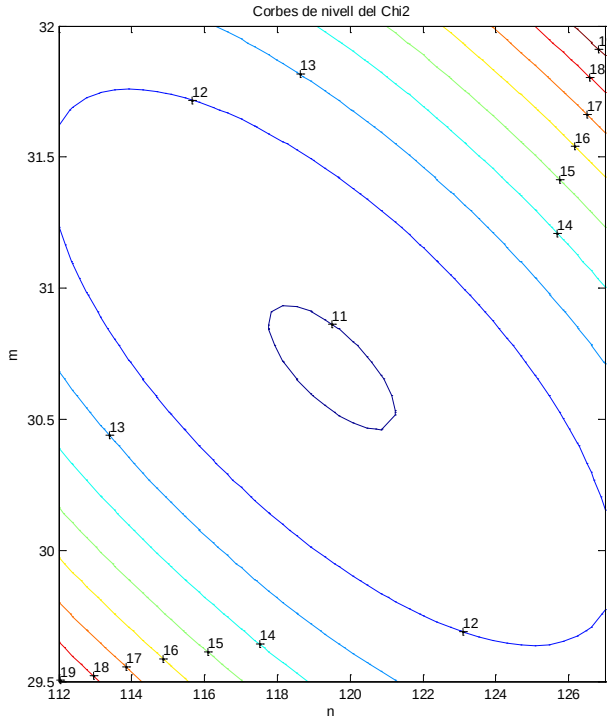
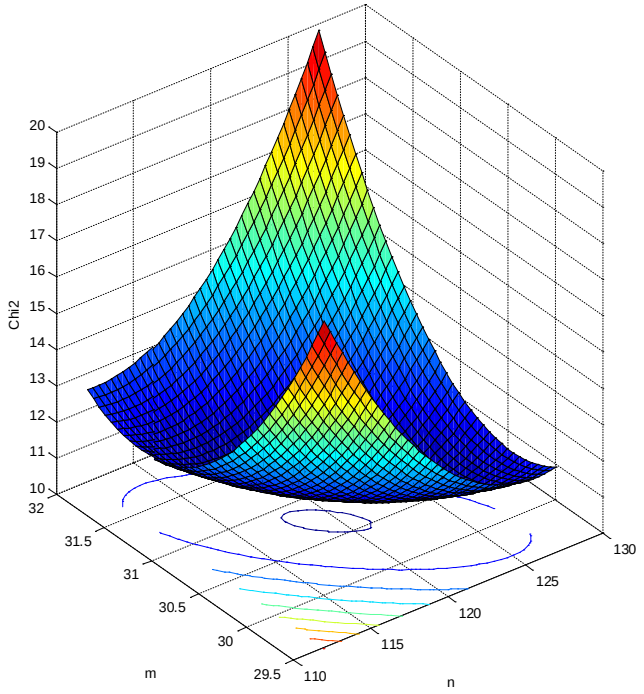


FIGURE 6.2
Number of counts in constant time intervals from a radioactive source as a function of the inverse distance from source to Geiger counter (Example 6.2). The vertical error bars indicate the statistical uncertainties in the counts. The straight line is the result of a least-squares fit to the data.

Ajust d'una recta

Representació gràfica del χ^2 de l'exemple 1 feta amb MATLAB.





Ajust d'una recta

Extrapolació:

- Una vegada tenim el pendent i l'ordenada, podem extrapolar o interpolar el valor de *Y* per una *X* donada, i obtenir el seu error.
- El valor extrapolat seria:

$$Y = m' X + n'$$

- I el seu error, seria l'arrel de la variància,

$$\sigma_Y^2 = \left(\frac{\partial Y}{\partial m'}\right)^2 \sigma_x^2 + \left(\frac{\partial Y}{\partial n'}\right)^2 \sigma_y^2 + 2\left(\frac{\partial Y}{\partial m'}\right)\left(\frac{\partial Y}{\partial n'}\right) \operatorname{cov}(m,n)$$

$$V(Y) = X^2 V(m) + V(n) + 2X \operatorname{cov}(m,n)$$

- El terme de la covariància pot ser molt important i hem de recordar que pot ser negatiu.
- En **l'exemple anterior**, l'extrapolació a una distància D= 2 m, ens donaria un recompte de

$$C = 30.7 \times (1/2\text{m})^2 + 119.5 = 127.2$$

$$V(C) = 0.25^2 \times 1.07 + 57.3 + 2 \times 0.25 \times (-5.6) = 54 \Rightarrow \sigma(C) = \sqrt{54} = 7.3$$

$$C = 127 \pm 7$$

Ajust d'una recta

Linealització:

- Funció exponencial: $y = ae^{-bx}$

- Prenent logaritmes, l'exponencial es transforma en una línia recta.

$$y' \equiv \ln y = \ln a - bx = a' - bx \qquad \text{amb } a' = \ln a$$

- El mètode de mínims quadrats minimitza el valor de χ^2

$$\chi^2 = \sum_i \left[\frac{\ln y_i - \ln a + bx_i}{\sigma'_i} \right]^2 \qquad \text{amb} \quad \sigma'_i = \frac{d \ln y}{dy}\sigma_i = \frac{1}{y_i}\sigma_i$$

- Els valors dels paràmetres a i b i els seus errors seran: fent $\Delta' \equiv \left(\sum_i \frac{1}{\sigma'^2_i}\right)\sum_i \frac{x_i^2}{\sigma'^2_i} - \left(\sum_i \frac{x_i}{\sigma'^2_i}\right)^2$

$$a' = \frac{1}{\Delta'}\left(\sum_i \frac{x_i^2}{\sigma'^2_i}\sum_i \frac{y'_i}{\sigma'^2_i} - \sum_i \frac{x_i}{\sigma'^2_i}\sum_i \frac{x_i y'_i}{\sigma'^2_i}\right) \qquad a = e^{a'} \qquad b = -\frac{1}{\Delta'}\left(\sum_i \frac{1}{\sigma'^2_i}\sum_i \frac{x_i y'_i}{\sigma'^2_i} - \sum_i \frac{x_i}{\sigma'^2_i}\sum_i \frac{y'_i}{\sigma'^2_i}\right)$$

$$V(a') = \frac{1}{\Delta'}\sum_i \frac{x_i^2}{\sigma'^2_i} \qquad V(a) = e^{2a'}V(a') \qquad V(b) = \frac{1}{\Delta'}\sum_i \frac{1}{\sigma'^2_i}$$

La distribució χ^2

- La quantitat χ^2 és la suma de les diferències al quadrat entre els valors observats i els valors teòrics predits, ponderats pels errors de les mesures.

$$\chi^2 = \sum_i \left[\frac{y_i - f(x_i; a_{1,}, ..., a_k)}{\sigma_i} \right]^2$$

- Si la funció s'ajusta bé amb els valors observats, el valor de χ^2 serà menut, per la qual cosa un valor gran de χ^2 després de la minimització, ens dirà que la funció teòrica no és la millor.
- Però un valor molt menut de χ^2 tampoc és bo i probablement ens està indicant una possible sobreestimació dels errors.
- El χ^2 és una variable aleatòria que es distribueix segon la funció de probabilitat, que depèn del paràmetre v.

$$P(\chi^2; \nu) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} \left(\chi^2\right)^{\frac{\nu-2}{2}} \exp(-\frac{\chi^2}{2})$$

Funció Gamma: $\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx$

$\Gamma(1) = 1 \quad \Gamma(\frac{1}{2}) = \sqrt{\pi} \quad \Gamma(n+1) = n\Gamma(n)$

$\Gamma(n+1) = n! \quad \text{si } n = 0, 1, ...$

$\Gamma(n+1) = n(n-1)(n-2)...\frac{3}{2} \times \left(\frac{1}{2}\sqrt{\pi}\right) \quad \text{si } n = \frac{1}{2}, \frac{3}{2}, ...$

La distribució χ^2

Demostració:

- Si cada punt y_i està distribuït al voltant del seu f_i d'acord amb una gaussiana, la variable normalitzada $u_i=[y_i-f_i]/\sigma_i$ estarà distribuïda com a una gaussiana tipificada (mitjana= 0 i sigma =1).
- Podem interpretar geomètricament els valors (u_1,...u_N) com un punt en un espai de probabilitat N-dimensional, amb probabilitat

$$P=\prod_{i=1}^{N}P(y_i,\,f_i)=\prod_{i=1}^{N}\frac{1}{\sigma_i\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y_i-f_i}{\sigma_i}\right)^2}\propto\exp(-\frac{1}{2}\sum_i u_i^{\,2})=\exp(-\chi^2/2)$$

- Els punts d'aquest espai amb igual probabilitat constitueixen una hipersuperfície de dimensió N-1 i radi $\sqrt{u_1^{\,2}+\ldots+u_N^{\,2}}=\sqrt{\chi^2}=\chi$
- La probabilitat d'obtenir un chi2 en l'interval χ i χ+dχ és proporcional a *P* multiplicat pel volum de la hipercorona que defineix aquest interval i, per tant,

$$P(\chi)\propto\exp(-\chi^2/2)\times\chi^{N-1}d\chi$$

- I finalment:

$$P(\chi^2)=\frac{dn}{d\chi^2}=\frac{dn}{d\chi}\left|\frac{d\chi}{d\chi^2}\right|=P(\chi)\frac{1}{2\chi}\propto\chi^{N-1}\exp(-\chi^2/2)\frac{1}{\chi}=\chi^{N-2}e^{-\chi^2/2}=\left(\chi^2\right)^{\frac{N-2}{2}}e^{-\chi^2/2}$$

La distribució χ^2

- Les restriccions imposades per les *k* equacions de minimització (una per paràmetre a ajustar) disminueixen la dimensió de l'espai de probabilitat inicial a una dimensió ν=*N-k.*
- La distribució depèn de *ν* (nombre de **graus de llibertat**) que és el nombre de punts en la suma N, menys el nombre de variables a minimitzar en l'ajust.
- La constant de proporcionalitat ve donada per la condició de normalització i resulta (vegeu p. e. *Análisis de errores* de C. Sánchez de Río).

$$P(\chi^2;\nu) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\left(\chi^2\right)^{\frac{\nu-2}{2}} \exp(-\frac{\chi^2}{2})$$

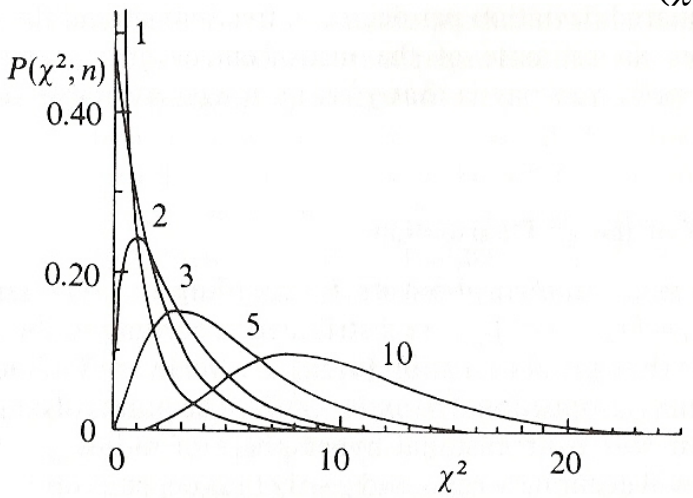


Fig. 6.4. Some χ^2 distributions.

El valor esperat i la variància d'aquesta distribució són:

$$\left\langle \chi^2 \right\rangle = \nu \qquad V\left(\chi^2\right) = 2\nu$$

Esperem un χ^2 reduït ($\tilde{\chi}^2 = \chi^2/\nu$) al voltant d'1 si la funció f ajustada és correcta. Podem calcular la probabilitat que les dades mesurades hagueren donat lloc a un $\tilde{\chi}^2$ superior al valor obtingut $\tilde{\chi}_0{}^2$

Límits: Si ajustem les dades a una funció amb $\tilde{\chi}^2 = \tilde{\chi}_0{}^2$ i

$$P(\tilde{\chi}^2 \geq \tilde{\chi}_0{}^2) < 5\% \quad (1\%)$$

rebutgem la distribució considerada al 5% (1%) de **nivell de significació.**

La distribució χ^2

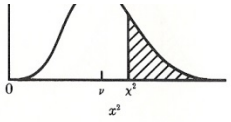


TABLE C.4
χ^2 distribution. Values of the reduced chi-square $\chi^2_\nu = \chi^2 / \nu$ corresponding to the probability $P_x(\chi^2; \nu)$ of exceeding χ^2 vs. the number of degrees of freedom ν

	P							
ν	0.99	0.98	0.95	0.90	0.80	0.70	0.60	0.50
1	0.00016	0.00063	0.00393	0.0158	0.0642	0.148	0.275	0.455
2	0.0100	0.0202	0.0515	0.105	0.223	0.357	0.511	0.693
3	0.0383	0.0617	0.117	0.195	0.335	0.475	0.623	0.789
4	0.0742	0.107	0.178	0.266	0.412	0.549	0.688	0.839
5	0.111	0.150	0.229	0.322	0.469	0.600	0.731	0.870
6	0.145	0.189	0.273	0.367	0.512	0.638	0.762	0.891
7	0.177	0.223	0.310	0.405	0.546	0.667	0.785	0.907
8	0.206	0.254	0.342	0.436	0.574	0.691	0.803	0.918
9	0.232	0.281	0.369	0.463	0.598	0.710	0.817	0.927
10	0.256	0.306	0.394	0.487	0.618	0.727	0.830	0.934
11	0.278	0.328	0.416	0.507	0.635	0.741	0.840	0.940
12	0.298	0.348	0.436	0.525	0.651	0.753	0.848	0.945
13	0.316	0.367	0.453	0.542	0.664	0.764	0.856	0.949
14	0.333	0.383	0.469	0.556	0.676	0.773	0.863	0.953
15	0.349	0.399	0.484	0.570	0.687	0.781	0.869	0.956
16	0.363	0.413	0.498	0.582	0.697	0.789	0.874	0.959
17	0.377	0.427	0.510	0.593	0.706	0.796	0.879	0.961
18	0.390	0.439	0.522	0.604	0.714	0.802	0.883	0.963
19	0.402	0.451	0.532	0.613	0.722	0.808	0.887	0.965
20	0.413	0.462	0.543	0.622	0.729	0.813	0.890	0.967
22	0.434	0.482	0.561	0.638	0.742	0.823	0.897	0.970
24	0.452	0.500	0.577	0.652	0.753	0.831	0.902	0.972
26	0.469	0.516	0.592	0.665	0.762	0.838	0.907	0.974
28	0.484	0.530	0.605	0.676	0.771	0.845	0.911	0.976
30	0.498	0.544	0.616	0.687	0.779	0.850	0.915	0.978
32	0.511	0.556	0.627	0.696	0.786	0.855	0.918	0.979
34	0.523	0.567	0.637	0.704	0.792	0.860	0.921	0.980
36	0.534	0.577	0.646	0.712	0.798	0.864	0.924	0.982
38	0.545	0.587	0.655	0.720	0.804	0.868	0.926	0.983
40	0.554	0.596	0.663	0.726	0.809	0.872	0.928	0.983
42	0.563	0.604	0.670	0.733	0.813	0.875	0.930	0.984
44	0.572	0.612	0.677	0.738	0.818	0.878	0.932	0.985
46	0.580	0.620	0.683	0.744	0.822	0.881	0.934	0.986
48	0.587	0.627	0.690	0.749	0.825	0.884	0.936	0.986
50	0.594	0.633	0.695	0.754	0.829	0.886	0.937	0.987
60	0.625	0.662	0.720	0.774	0.844	0.897	0.944	0.989
70	0.649	0.684	0.739	0.790	0.856	0.905	0.949	0.990
80	0.669	0.703	0.755	0.803	0.865	0.911	0.952	0.992
90	0.686	0.718	0.768	0.814	0.873	0.917	0.955	0.993
100	0.701	0.731	0.779	0.824	0.879	0.921	0.958	0.993
120	0.724	0.753	0.798	0.839	0.890	0.928	0.962	0.994
140	0.743	0.770	0.812	0.850	0.898	0.934	0.965	0.995
160	0.758	0.784	0.823	0.860	0.905	0.938	0.968	0.996
180	0.771	0.796	0.833	0.868	0.910	0.942	0.970	0.996
200	0.782	0.806	0.841	0.874	0.915	0.945	0.972	0.997

TABLE C.4
χ^2 distribution (*continued*)

	P							
ν	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.708	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.916	1.204	1.609	2.303	2.996	3.912	4.605	6.908
3	0.982	1.222	1.547	2.084	2.605	3.279	3.780	5.423
4	1.011	1.220	1.497	1.945	2.372	2.917	3.319	4.617
5	1.026	1.213	1.458	1.847	2.214	2.678	3.017	4.102
6	1.035	1.205	1.426	1.774	2.099	2.506	2.802	3.743
7	1.040	1.198	1.400	1.717	2.010	2.375	2.639	3.475
8	1.044	1.191	1.379	1.670	1.938	2.271	2.511	3.266
9	1.046	1.184	1.360	1.632	1.880	2.187	2.407	3.097
10	1.047	1.178	1.344	1.599	1.831	2.116	2.321	2.959
11	1.048	1.173	1.330	1.570	1.789	2.056	2.248	2.842
12	1.049	1.168	1.318	1.546	1.752	2.004	2.185	2.742
13	1.049	1.163	1.307	1.524	1.720	1.959	2.130	2.656
14	1.049	1.159	1.296	1.505	1.692	1.919	2.082	2.580
15	1.049	1.155	1.287	1.487	1.666	1.884	2.039	2.513
16	1.049	1.151	1.279	1.471	1.644	1.852	2.000	2.453
17	1.048	1.148	1.271	1.457	1.623	1.823	1.965	2.399
18	1.048	1.145	1.264	1.444	1.604	1.797	1.934	2.351
19	1.048	1.142	1.258	1.432	1.586	1.773	1.905	2.307
20	1.048	1.139	1.252	1.421	1.571	1.751	1.878	2.266
22	1.047	1.134	1.241	1.401	1.542	1.712	1.831	2.194
24	1.046	1.129	1.231	1.383	1.517	1.678	1.791	2.132
26	1.045	1.125	1.223	1.368	1.496	1.648	1.755	2.079
28	1.045	1.121	1.215	1.354	1.476	1.622	1.724	2.032
30	1.044	1.118	1.208	1.342	1.459	1.599	1.696	1.990
32	1.043	1.115	1.202	1.331	1.444	1.578	1.671	1.953
34	1.042	1.112	1.196	1.321	1.429	1.559	1.649	1.919
36	1.042	1.109	1.191	1.311	1.417	1.541	1.628	1.888
38	1.041	1.106	1.186	1.303	1.405	1.525	1.610	1.861
40	1.041	1.104	1.182	1.295	1.394	1.511	1.592	1.835
42	1.040	1.102	1.178	1.288	1.384	1.497	1.576	1.812
44	1.039	1.100	1.174	1.281	1.375	1.485	1.562	1.790
46	1.039	1.098	1.170	1.275	1.366	1.473	1.548	1.770
48	1.038	1.096	1.167	1.269	1.358	1.462	1.535	1.751
50	1.038	1.094	1.163	1.263	1.350	1.452	1.523	1.733
60	1.036	1.087	1.150	1.240	1.318	1.410	1.473	1.660
70	1.034	1.081	1.139	1.222	1.293	1.377	1.435	1.605
80	1.032	1.076	1.130	1.207	1.273	1.351	1.404	1.560
90	1.031	1.072	1.123	1.195	1.257	1.329	1.379	1.525
100	1.029	1.069	1.117	1.185	1.243	1.311	1.358	1.494
120	1.027	1.063	1.107	1.169	1.221	1.283	1.325	1.446
140	1.026	1.059	1.099	1.156	1.204	1.261	1.299	1.410
160	1.024	1.055	1.093	1.146	1.191	1.243	1.278	1.381
180	1.023	1.052	1.087	1.137	1.179	1.228	1.261	1.358
200	1.022	1.050	1.083	1.130	1.170	1.216	1.247	1.338

259

El test de χ^2 : aplicació a l'ajust d'una recta

- **Exemple:** La distribució χ^2de l'exemple de les 10 mesures del Geiger té 10-2=8 graus de llibertat i per tant:

$$P(\chi^2;8)=\frac{2^{-8/2}}{\Gamma(8/2)}\left(\chi^2\right)^{\frac{8-2}{2}}\exp(-\frac{\chi^2}{2})=\frac{1}{96}\left(\chi^2\right)^3\exp(-\frac{\chi^2}{2})$$

$$\left\langle\chi^2\right\rangle=8 \qquad \mathrm{Var}\left(\chi^2\right)=16$$

- Calculem:

$$\chi^2=\sum_i\left[\frac{y_i-mx_i-n}{\sigma_i}\right]^2=11.1\Rightarrow\chi'^2\equiv\chi^2/\nu=1.39$$

- De la taula:

$$\int_{\chi^2}^{\infty}P(\chi'^2;8)d\chi'^2\approx 0.20>5\%$$

```
>> chi2cdf(Chi2,nu,'upper')
```

- Interpretació: Si repetirem l'experiment moltes vegades, el 20% de vegades obtindríem valors de χ^2 en ajustar les diferents rectes obtingudes, iguals o superiors al nostre valor, la qual cosa indica un bon ajust (valors massa menuts o massa grans indicarien un model o unes dades incorrectes.)

Formulació matricial dels mínims quadrats

- Si la funció *f* a ajustar té *k* paràmetres $a_1, a_2, \ldots a_k$ escrivim el vector:

$$\vec{a} = (a_1, \ldots, a_k)$$

- Igualment construïm els vectors de *N* components:

$$\mathbf{y} = (y_1, \ldots, y_N) \qquad \mathbf{f} = \left(f(x_1; \vec{a}), \ldots, f(x_N; \vec{a}) \right)$$

- Es pot demostrar que el χ^2 ve donat per l'equació matricial

$$\chi^2 = (\mathbf{y} - \mathbf{f})^T \mathbf{V}_y^{-1} (\mathbf{y} - \mathbf{f})$$

on **V** és la matriu de covariància de les dades, que és diagonal si les **mesures** són **independents**, i en aquest cas:

$$V_{ij}^{-1} = \mathrm{cov}^{-1}(y_i, y_j) = \frac{1}{\sigma_i^2} \delta_{ij}$$

- Derivant χ^2 respecte de cada a_r i igualant a zero obtenim *k* equacions ***(equacions normals)***, que poden ser resoltes per a trobar l'estimació de mínims quadrats de **a**.

Formulació matricial dels mínims quadrats

1. Funcions lineals dels paràmetres:

$$f(x;\vec{a}) = a_o f_0(x) + a_1 f_1(x) + \ldots + a_k f_k(x) = \sum_{j=1}^{k} a_j f_j(x_i)$$

p. e. $a)$ $f(x) = a_o + a_1 x + a_2 x^2 + \ldots$ $b)$ $f(x) = asen(x) + b\cos(x)$

En aquest cas la minimització del χ^2 condueix a equacions normals lineals amb solució analítica.

Equacions normals

$$\chi^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left[y_i - \sum_{j=1}^{k} a_j f_j(x_i) \right]^2 \Longrightarrow \sum_{i=1}^{N} f_j(x_i)\frac{1}{\sigma_i^2}\left[y_i - \sum_{j=1}^{k} a_j f_j(x_i) \right] = 0 \quad \text{j=1,....k}$$

Formulació matricial dels mínims quadrats

En l'exemple b) $f(x) = asen(x) + b\cos(x)$

$$\chi^2 = \sum_i \left(\frac{y_i - asenx_i - bsenx_i}{\sigma_i} \right)^2$$

$$\begin{cases} \dfrac{\partial \chi^2}{\partial a} = 0 \Rightarrow \sum_i (y_i - asenx_i - b\cos x_i) senx_i = 0 \\ \dfrac{\partial \chi^2}{\partial b} = 0 \Rightarrow \sum_i (y_i - asenx_i - b\cos x_i)\cos x_i = 0 \end{cases} \Rightarrow \begin{cases} \overbrace{\left(\sum_i sen^2 x_i\right)}^{C_1} a + \overbrace{\left(\sum_i senx_i \cos x_i\right)}^{C_2} b = \overbrace{\sum_i y_i senx_i}^{C_0} \\ \underbrace{\left(\sum_i senx_i \cos x_i\right)}_{D_1} a + \underbrace{\left(\sum_i \cos^2 x_i\right)}_{D_2} b = \underbrace{\sum_i y_i \cos x_i}_{D_0} \end{cases}$$

$$\begin{cases} a = \dfrac{1}{\Delta} \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix} \\ b = \dfrac{1}{\Delta} \begin{vmatrix} C_1 & C_0 \\ D_1 & D_0 \end{vmatrix} \end{cases} \quad \text{amb} \quad \Delta = \begin{vmatrix} C_1 & C_2 \\ D_1 & D_2 \end{vmatrix} = \left(\sum_i sen^2 x_i\right)\left(\sum_i \cos^2 x_i\right) - \left(\sum_i senx_i \cos x_i\right)^2$$

Formulació matricial dels mínims quadrats

Quan la funció és lineal en els paràmetres podem escriure

$$\begin{cases} f(x_1) = a_o f_0(x_1) + a_1 f_1(x_1) + \ldots + a_k f_k(x_1) \\ f(x_2) = a_o f_0(x_2) + a_1 f_1(x_2) + \ldots + a_k f_k(x_2) \\ . \\ f(x_N) = a_o f_0(x_N) + a_1 f_1(x_N) + \ldots + a_k f_k(x_N) \end{cases} \qquad \begin{pmatrix} f(x_1) \\ f(x_2). \\ | \\ f(x_N) \end{pmatrix} = \underbrace{\begin{pmatrix} f_0(x_1) & f_1(x_1) \ldots\ldots f_k(x_1) \\ f_0(x_2) & f_1(x_2) \ldots\ldots f_k(x_2) \\ | & \\ f_0(x_N) & f_1(x_N) \ldots\ldots f_k(x_N) \end{pmatrix}}_{\mathrm{C}} \begin{pmatrix} a_0 \\ a_1 \\ | \\ a_k \end{pmatrix}$$

o bé $\mathbf{f} = \mathbf{C}\mathbf{a}$ amb $C_{ij} = f_j(x_i)$

a) $C = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ | & & \\ 1 & x_N & x_N^2 \end{pmatrix}$ *b)* $C = \begin{pmatrix} \sin(x_1) & \cos(x_1) \\ \sin(x_2) & \cos(x_2) \\ | & \\ \sin(x_N) & \cos(x_N) \end{pmatrix}$

Es pot provar que els paràmetres $a_1, a_2, \ldots a_k$ i la matriu de covariància dels paràmetres poden expressar-se en funció d'aquesta matriu C i de la matriu de covariància de les dades V_y com a:

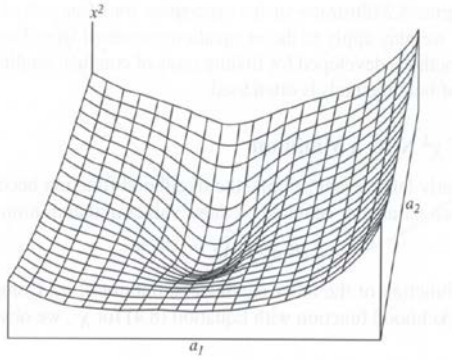
$$\mathbf{a} = \left[\mathbf{C}^T \mathbf{V}_y^{-1} \mathbf{C}\right]^{-1} \mathbf{C}^T \mathbf{V}_y^{-1} \mathbf{y}$$

$$\mathbf{V}(\vec{a}) = \left[\mathbf{C}^T \mathbf{V}_y^{-1} \mathbf{C}\right]^{-1}$$

Funcions no lineals en els paràmetres

2. Funcions no lineals dels paràmetres (p. e. $y = a_1 + sen(x + a_2)$ o $y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$ ...)

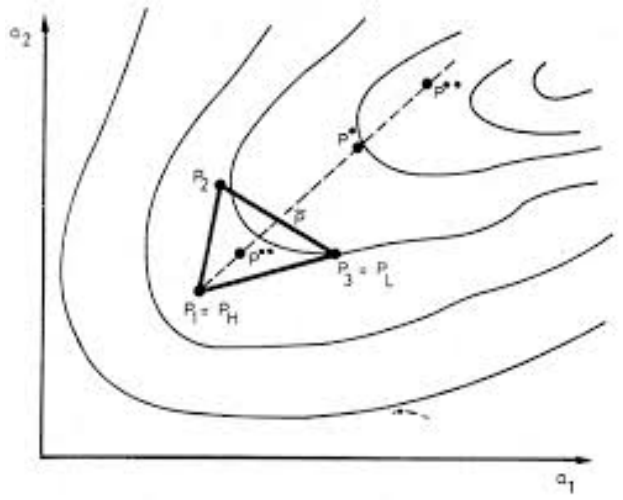
- Els ajusts no lineals generalment requereixen procediments numèrics (iteratius) per a minimitzar el χ^2.
- N'hi ha diferents mètodes, però cap és aplicable a totes les funcions.
- Els criteris per a seleccionar el mètode són:
  - la velocitat del càlcul,
  - l'estabilitat enfront de les divergències.
- Els mètodes poden classificar-se en dues categories:
  - **Mètodes de quadrícula** (*grid method*): es forma una quadrícula de punts igualment espaiats en les variables d'interès i s'avalua la funció en cada un dels punts. El punt amb el valor més menut és el mínim aproximat. Un d'aquests mètodes és el *mètode símplex.*
  - **Mètodes del gradient:** a partir de les derivades de la funció a minimitzar, es troba la direcció en la qual la funció decreix i es calcula per iteració el mínim de la funció. Un d'aquests mètodes és el *mètode de Newton.*



Funcions no lineals en els paràmetres

`MATLAB fminsearch` uses the **simplex search method** of Lagarias et al. [1]. This is a direct search method that does not use numerical or analytic gradients.

If n is the length of x, a simplex in n-dimensional space is characterized by the n+1 distinct vectors that are its vertices. In two-space, a simplex is a triangle; in three-space, it is a pyramid. At each step of the search, a new point in or near the current simplex is generated. The function value at the new point is compared with the function's values at the vertices of the simplex and, usually, one of the vertices is replaced by the new point, giving a new simplex. This step is repeated until the diameter of the simplex is less than the specified tolerance.



[1] Lagarias, J.C., J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," SIAM Journal of Optimization, Vol. 9 Number 1, pp. 112-147, 1998.

J. A. Nelder, R. Mead: Comput. J. 7,308 (1965)

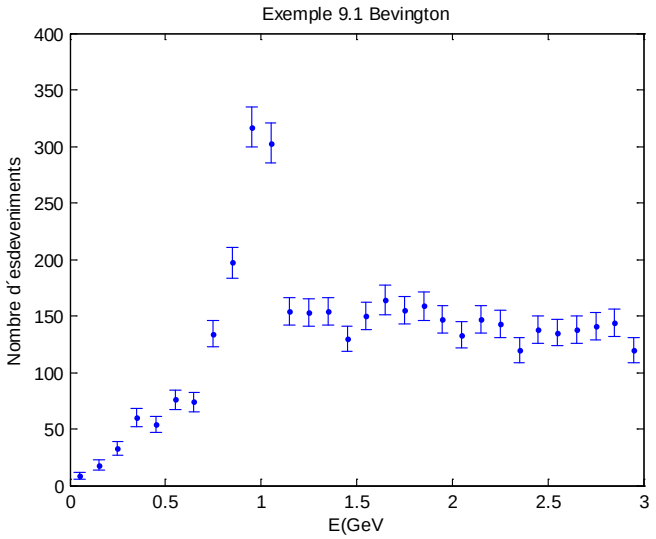
Funcions no lineals en els paràmetres

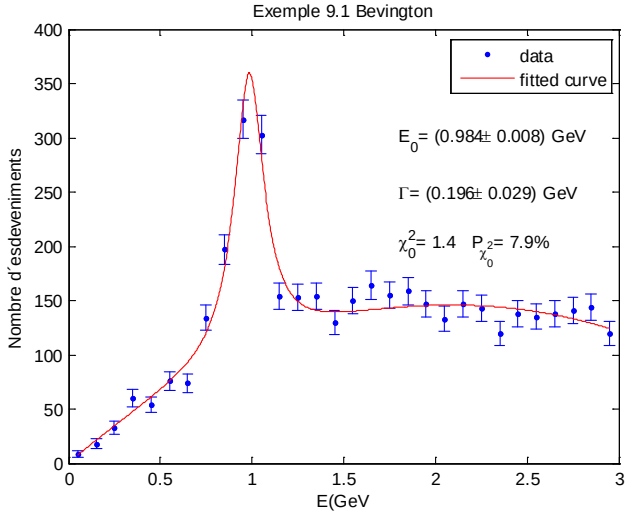
– ***Exemple:***

En un experiment de física de partícules s'ha mesurat l'energia d'un procés determinat i s'ha acumulat una estadística de 4000 esdeveniments. L'observació de l'histograma mostra un gran pic sobre un fons suau, que interpretem com una ressonància descrita per una distribució de Lorentz i un fons que depèn quadràticament amb l'energia. Representem l'histograma en passos de 0.10 GeV i ajustem les dades a la següent funció amb sis paràmetres lliures:

$$y(E) = a + bE + cE^2 + A_0 \frac{\Gamma/(2\pi)}{(E - E_0)^2 + (\Gamma/2)^2}$$

Ajust amb la instrucció `fit` de MATLAB





Tècniques de Montecarlo

1. Introducció
2. Nombres aleatoris
3. Nombres aleatoris segons distribucions de probabilitat
4. Distribucions aleatòries específiques
5. Exemples

El nom Montecarlo prové de la ciutat europea amb el casino més famós del món: la connexió està en els nombres aleatoris que apareixen en el mètode i també en els jocs d'atzar del casino.

Bibliografia*:*

- *Data reduction and error analysis for the physical sciences* (Bevington, Robinson)

Introducció

- El **mètode de Montecarlo** és una tècnica que, utilitzant la generació de nombres aleatoris, permet resoldre diversos tipus de problemes físics o matemàtics com ara l'extracció de paràmetres a partir de dades experimentals, sense haver de resoldre explícitament les integrals de les distribucions implicades.

- Bàsicament és un mètode per a calcular integrals múltiples a partir de mostres aleatòries.

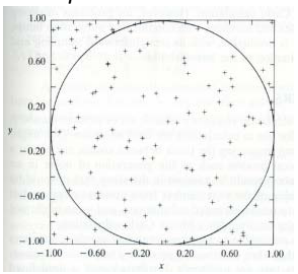
- En la pràctica, la tècnica MC proporciona un mètode per a simular experiments i generar dades experimentals a partir de models físics.

Introducció

Exemple 1: Càlcul de l'**àrea** d'un cercle, de radi conegut, sense utilitzar les integrals de superfície.

- Estratègia: inscrivim el cercle en un quadrat de longitud 2R i cobrim la superfície del quadrat amb marques de manera uniforme (per exemple, amb grans d'arròs).
- Comptem les marques: per una banda, les que cauen dins del quadrat N_q i, per altra banda, les que cauen dins del cercle N_c.
- L'àrea del cercle serà:

$$A_c = A_q \frac{N_c}{N_q}$$



- L'error en la determinació de l'àrea dependrà del nombre i la grandària dels grans d'arròs a més de la uniformitat en la seua distribució.
- Si generem aleatòriament la posició dels grans, reduïm el problema a un càlcul de probabilitats: La probabilitat de trobar el gra dins del cercle segueix una distribució binomial amb probabilitat $p = A_c/A_q$ i variància Npq, és a dir,

$$\sigma^2 = N_q p(1-p) = N_c(1-p)$$

Introducció

- Si el cercle té un radi R= 1 m, construïm un quadrat entre -1 m i 1 m (longitud 2 m) i generem N=100 parelles (x,y) de nombres aleatoris entre -1 i +1.

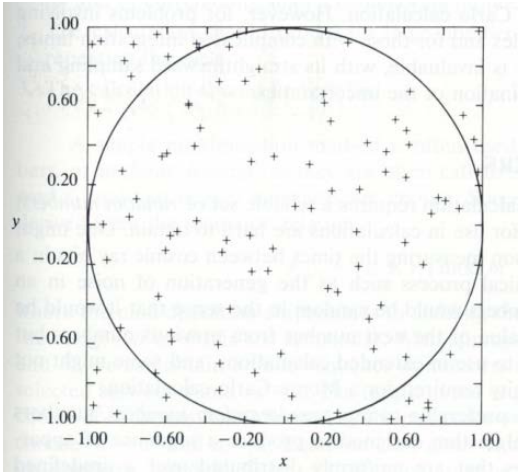
- Com que la probabilitat $p = \frac{\pi R^2}{2R \times 2R} = \frac{\pi}{4}$ **esperem** trobar Np=78.5 punts dins del cercle amb un error de

$$\sigma = \sqrt{N_c(1-p)} = \sqrt{78.5 \times \left(1 - \pi/4\right)} = 4.1$$

i per tant l'àrea:

$$A_c = A_q \frac{N_c}{N_q} = \left(4\text{m}^2\right) \times \frac{78.5}{100} = 3.14\text{m}^2$$

$$\sigma(A_c) = A_q \frac{\sigma(N_c)}{N_q} = \left(4\text{m}^2\right) \times \frac{4.1}{100} = 0.16\text{m}^2$$



Distribució de 100 punts generats uniformement. Es compten 73 punts dins del cercle, que correspon a una àrea A=2.92+-0.18 m²

$$A_c = A_q \frac{N_c}{N_q} = \left(4\text{m}^2\right) \times \frac{73}{100} = 2.92\text{m}^2$$

$$\sigma(A_c) = A_q \frac{\sigma(N_c)}{N_q} = \left(4\text{m}^2\right) \times \frac{\sqrt{73(1-0.73)}}{100} = 0.18\text{m}^2$$

Com que $\sigma(N_C)$ és proporcional a $\sqrt{N_q}$ l'error de l'àrea varia com a $1/\sqrt{N_q}$!!

Nombres aleatoris

- Els càlculs MC requereixen un conjunt fidedigne de nombres aleatoris.
- Un conjunt verdader de nombres aleatoris és difícil d'aconseguir (p. e. mesurant processos aleatoris naturals, com ara els intervals de temps de detecció dels raigs còsmics).
- Per als càlculs MC utilitzem ***nombres pseudoaleatoris,*** generats a partir d'un algoritme:
- Avantatges dels nombres pseudoaleatoris:
  - Generació dins del mateix programa MC d'un gran nombre de nombres aleatoris
  - Seqüència de nombres que es pot repetir, cosa que permet el control del programa (*debugging*).
  - Portabilitat entre programes i ordinadors.
- Requisits del generador de nombres aleatoris:
  - La seua distribució ha de ser uniforme en un rang específic, amb una seqüència no predictible i sense correlacions entre els nombres.
  - Que el període de la seqüència siga llarg abans de repetir el cicle.
  - Que els càlculs siguen ràpids.

Nombres aleatoris

- **Algoritme generador** de nombres aleatoris (*uniform deviates*): relació de recurrència a partir d'un valor inicial o *llavor* r_1 i les constants *a* i *m*

$$r_{i+1} = (a \times r_i) \bmod m \qquad i=1,2...$$

```
%scriptpseudorandomgenerator
a=5;m=37;r(1,1)=1;
for i=2:m-1
r(i,1)=mod(a*r(i-1,1),m);
end
```

- **Exemple:** *m=37; a=5 i* r_1*=1*

$$r_1 = 1$$
$$r_2 = (a \times r_1) \bmod m = (5 \times 1) \bmod 37 = \left[(0 \times 37) + 5\right] \bmod 37 = 5$$
$$r_3 = (a \times r_2) \bmod m = (5 \times 5) \bmod 37 = \left[(0 \times 37) + 25\right] \bmod 37 = 25$$
$$r_4 = (a \times r_3) \bmod m = (5 \times 25) \bmod 37 = \left[(3 \times 37) + 14\right] \bmod 37 = 14$$
$$r_5 = (a \times r_4) \bmod m = (5 \times 14) \bmod 37 = \left[(1 \times 37) + 33\right] \bmod 37 = 33$$
$$\ldots$$
$$r_{36} = (a \times r_{35}) \bmod m = (5 \times 3) \bmod 37 = \left[(0 \times 37) + 15\right] \bmod 37 = 15$$
$$r_{37} = (a \times r_{35}) \bmod m = (5 \times 15) \bmod 37 = \left[(2 \times 37) + 1\right] \bmod 37 = 1 = r_1$$

TABLE 5.1
Pseudorandom numbers†

i	r_i	i	r_i	i	r_i	i	r_i
1	1	10	6	19	36	28	31
2	5	11	30	20	32	29	7
3	25	12	2	21	12	30	35
4	14	13	10	22	23	31	27
5	33	14	13	23	4	32	24
6	17	15	28	24	20	33	9
7	11	16	29	25	26	34	8
8	18	17	34	26	19	35	3
9	16	18	22	27	21	36	15

†The generating equation is $r_{i+1} = (a \times r_i) \bmod m$, with $a = 5$ and $m = 37$. The cycle repeats $a_{37} = a_1$, $a_{38} = a_2$, and so forth.

- La longitud de la seqüència ve determinada per l'elecció de les constants i està limitat per la precisió del càlcul (quantitat de bits associats a cada nombre).

Nombres aleatoris

Generació de nombres aleatoris uniformes amb MATLAB: la instrucció `rand`

>> help rand

rand - Uniformly distributed pseudorandom numbers

This MATLAB function returns a pseudorandom scalar drawn from the standard uniform distribution on the open interval (0,1).

```
>> r=rand
r =
   0.4459
```

```
>> r=rand(5)
r =
   0.8142   0.3777   0.8019   0.3161   0.3867
   0.3950   0.7858   0.2944   0.5216   0.2134
   0.4940   0.7111   0.9610   0.7523   0.0447
   0.2553   0.5237   0.5889   0.5577   0.3256
   0.8240   0.6896   0.8805   0.1751   0.8705
```

```
>> r=rand(10,1)
r =
   0.5248
   0.1718
   0.0114
   0.1878
   0.3950
   0.8379
   0.8655
   0.2973
   0.7597
   0.5877
```

```
>> r=rand(1,10)
r =
   0.4289   0.9772   0.9703   0.9826   0.8593   0.5732   0.1782   0.7990   0.2976   0.4607
```

Nombres aleatoris segons distribucions de probabilitat

Com generar nombres aleatoris extrets de distribucions de probabilitat específiques?

1. Mètode de transformació

– Definim primer la **distribució uniforme** p(r) entre 0 i 1

$$p(r) = \begin{cases} 1 & \text{si } 0 \le r \le 1 \\ 0 & \text{per la resta de valors r} \end{cases}$$

que està normalitzada: $\int_{-\infty}^{+\infty} p(r)dr = 1$

- Per a obtenir nombres aleatoris a partir d'una distribució diferent, com per exemple,

$$P(x) = \begin{cases} \frac{1}{2} & \text{si } -1 \le x \le 1 \\ 0 & \text{per la resta de valors x} \end{cases}$$

Podem fer la transformació $x = 2r - 1$ que ens permet obtenir un valor x seguint P(x) per cada valor *r* extret de la distribució p(r).

Nombres aleatoris segons distribucions de probabilitat

- En general per a obtenir nombres aleatoris x seguint d'una funció de probabilitat P(x) a partir dels nombres aleatoris r distribuïts uniformement amb p(r), partim de la condició de conservació de la probabilitat:

$$p(r)dr = P(x)dx$$

- I per tant:

$$\int_{r=-\infty}^{r} p(r)dr = \int_{x=-\infty}^{x} P(x)dx \Rightarrow \int_{r=0}^{r} dr = r = \int_{x=-\infty}^{x} P(x)dx$$

- Per tant, per a trobar nombres aleatoris *x* seguint P(x), generem *r* uniformement i resolem l'equació anterior (que és equivalent a calcular la funció acumulada inversa).

Nombres aleatoris segons distribucions de probabilitat

– **Exemple:** Considerem la distribució descrita per l'equació:

$$p(x) = \begin{cases} A(1+ax^2) & \text{si } -1 \le x < 1 \\ 0 & \text{si } x < 1 \text{ ó } x \ge 1 \end{cases}$$

amb p(x) positiva o zero i la constant A tal que $\int_{-\infty}^{+\infty} p(x)dx = 1$

$$r = \int_{x=-\infty}^{x} p(x)dx = \int_{x=-\infty}^{x} A(1+x^2)dx = A(x + \tfrac{1}{3}ax^3 + 1 + \tfrac{1}{3}a)$$

Per a trobar els nombres aleatoris *x* seguint p(x), generem *r* uniformement i resolem l'equació de tercer grau anterior, que relaciona *x* amb *r.*

En general, ni l'equació integral, ni l'equació resultant poden resoldre's de manera analítica i hem de recórrer al càlcul numèric.

Nombres aleatoris segons distribucions de probabilitat

2. Mètode de rebuig:

Com el que hem utilitzat per a calcular l'àrea del cercle generant nombres aleatoris uniformement sobre la superfície del quadrat i rebutjant tots excepte els que cauen dins de la circumferència.

Exemple: Volem obtenir una distribució aleatòria entre [-1,1] seguint la distribució:

$$p(x) = 1 + ax^2$$

Generem una distribució x' uniformement distribuïda en l'interval [-1,1] que és l'abast de x, i una segona distribució y' uniformement distribuïda en l'interval [0,1+a], que és l'abast de p(x). Les equacions de transformació a partir dels valors aleatoris uniformes r_i i r_{i+1}:

$$x' = 2r_i - 1$$

$$y' = (1+a)r_{i+1}$$

Acceptem els valors x´ si el punt definit per (x´,y´) cau entre la corba P(x) i l'eix X, és a dir,

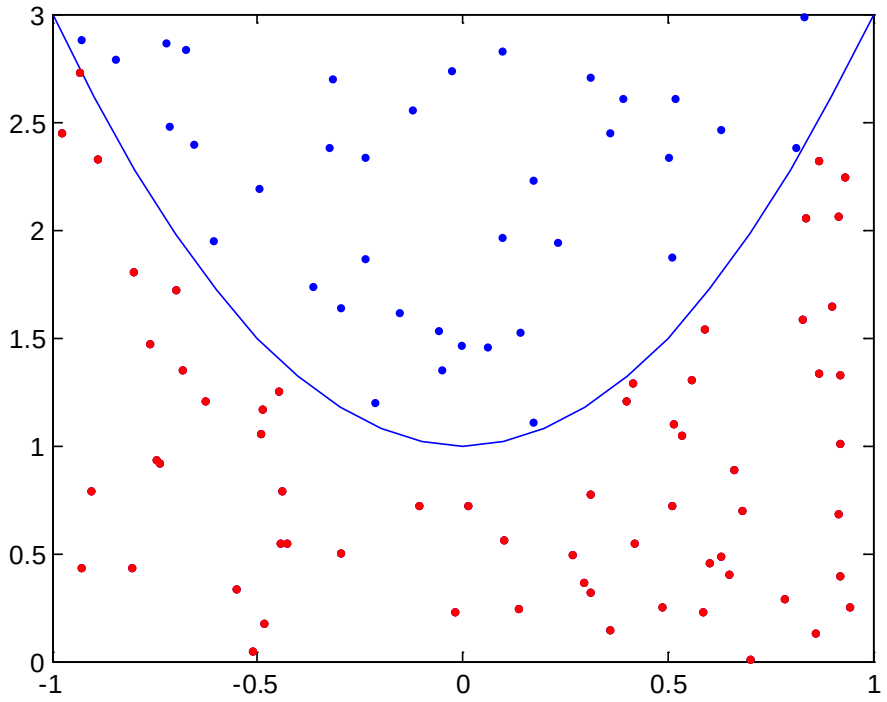
$$y' < p(x')$$

Nombres aleatoris segons distribucions de probabilitat

```
>>a=2;N=100
>> x=-1:0.1:1
>> y=1+a*x^2
>>plot(x,y)
>> hold on

>> r=rand(N,2);
>> xp=2*r(:,1)-1;
>> yp=(1+a)*r(:,2);
>> plot(xp,yp,'.')

>> X=xp(find(yp<1+a*xp.^2));
>> Y=yp(find(yp<1+a*xp.^2));
>> plot(X,Y,'r.')
>> hold off
```



Distribucions aleatòries específiques

1. Distribució gaussiana aleatòria

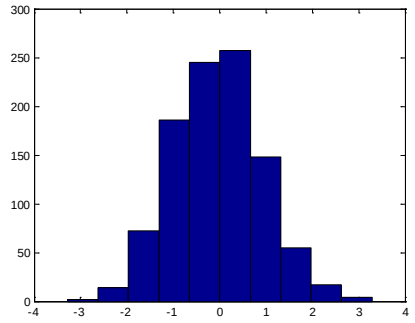
– Generarem nombres aleatoris seguint una gaussiana tipificada: $G_{0,1}(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

Després es calcula la variable x de la gaussiana $G_{\mu,\sigma}(x)$ amb el canvi:

$$x = \sigma z + \mu$$

– El mètode de transformació condueix a la integral $r = \int_{z=-\infty}^{z} G_{0,1}(z)dz$ que no es pot resoldre analíticament.

1.1. Una solució és la **integració numèrica** (resultats tabulats o calculables):



Histograma de 1000 nombres aleatoris fent ús del mètode de transformació i per integració numèrica amb MATLAB.

Generació de nombres aleatoris amb pdf Gauss a partir de nombres aleatoris uniformes amb la instrucció icdf de MATLAB

```
>>  mu=0;sigma=1;
>> r=rand(1,10)
r =
   0.6787   0.7577   0.7431   0.3922   0.6555   0.1712   0.7060   0.0318   0.2769   0.0462
>> z=icdf('normal',r,mu,sigma)
z =
   0.4642   0.6991   0.6530  -0.2735   0.4002  -0.9495   0.5419  -1.8545  -0.5920  -1.6832
```

Distribucions aleatòries específiques

1.2. Mètode de transformació de **Box i Müller** (1958)

Està basat en el fet que es pot trobar solució analítica a la generació d'una distribució gaussiana bidimensional (binormal):

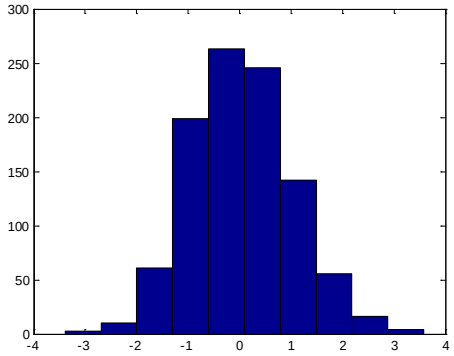
$$f(z_1, z_2) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(z_1^2 + z_2^2\right)} = \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \times \frac{1}{\sqrt{2\pi}} e^{-z_2^2/2}$$

A partir d'aquesta equació s'obtenen les expressions que generen aleatòriament dues gaussianes a partir dels valors aleatoris uniformes r_1 i r_2:

$$z_1 = \sqrt{-2\ln r_1}\cos 2\pi r_2$$
$$z_2 = \sqrt{-2\ln r_1}\sin 2\pi r_2$$

Generació de nombres aleatoris amb pdf Gauss (Box&Müller)

```
>> r=rand(1000,2);
>> z1=sqrt(-2*log(r(:,1))).*cos(2*pi*r(:,2));
>> hist(z1)
```



Distribucions aleatòries específiques

2. Distribució poissoniana aleatòria

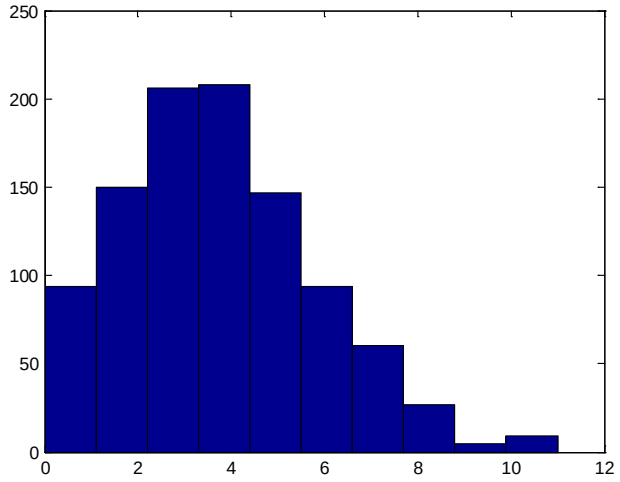
Apliquem el mètode de transformació, hem de resoldre ν de l'equació:

$$r = \sum_{\nu=0}^{\nu} \frac{\mu^{\nu}}{\nu!} e^{-\mu}$$

És a dir, que volem la funció acumulada inversa de la Poisson de mitjana μ

Generació de nombres aleatoris amb pdf Poisson a partir de nombres aleatoris uniformes amb la instrucció icdf de MATLAB

```
>>  mu=4;
>> r=rand(1,1000)
>> nu=icdf('Poisson',r,mu)
>>hist(nu)
```



Distribucions aleatòries específiques

3. Distribució exponencial aleatòria

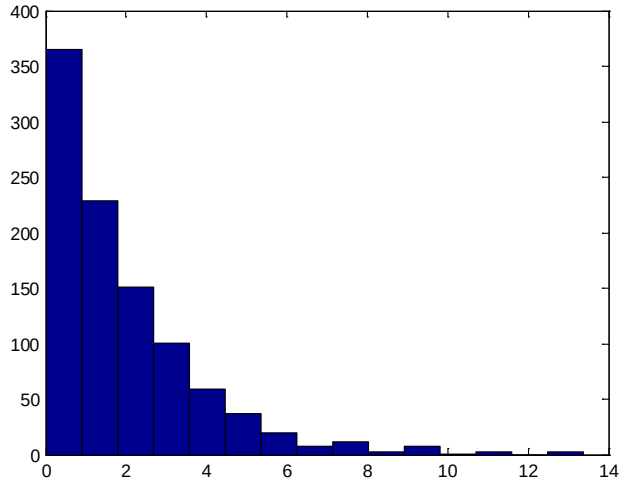
$$P(t;\tau)=\begin{cases}0 & \text{si } t<0\\ \frac{1}{\tau}e^{-t/\tau} & \text{si } t\geq 0\end{cases}$$

Apliquem el mètode de transformació, hem de resoldre t de l'equació $r=\int_{t=0}^{t} P(t)dt$

$$r=\int_{t=0}^{t} P(t)dt=\int_{t=0}^{t}\frac{1}{\tau}e^{-t/\tau}dt=1-e^{-t/\tau}\Rightarrow t=-\tau\ln(1-r)$$

Generació de nombres aleatoris amb pdf exponencial a partir de nombres aleatoris uniformes

```
>>  tau=2;
>> r=rand(1,1000)
>> t=-tau*log(1-r)
>>hist(t)
```



La desintegració radioactiva segueix aquesta distribució. En efecte, el nombre de desintegracions dN en un interval [t,t+ dt] és proporcional al nombre de nuclis N presents en l'instant t i a l'interval de temps, essent la constant de proporcionalitat la constant de desintegració λ de l'isòtop en qüestió:

$$dN=-\lambda N(t)dt\Rightarrow N/N_0=e^{-\lambda t}$$

La vida mitjana de l'isòtop és $\tau=\int_0^{\infty} te^{-\lambda t}=\lambda^{-1}$

Distribucions aleatòries específiques

Instruccions de generació aleatòria amb MATLAB: la instrucció `random`

>> help random

random - Random numbers

This MATLAB function where name is the name of a distribution that takes a single parameter, returns random numbers Y from the one-parameter family of distributions specified by name.

Y = random(name,A)
Y = random(name,A,B)
Y = random(name,A,B,C)
Y = random(name,A,m,n,...)
Y = random(name,A,[m,n,...])
Y = random(name,A,B,m,n,...)
Y = random(name,A,B,[m,n,...])
Y = random(name,A,B,C,m,n,...)
Y = random(name,A,B,C,[m,n,...])

Y = random(name,A) where name is the name of a distribution that takes a single parameter, returns random numbers Y from the one-parameter family of distributions specified by name. Parameter values for the distribution are given in A.

Y = random(name,A,B) returns random numbers Y from a two-parameter family of distributions. Parameter values for the distribution are given in A and B.

Y = random(name,A,m,n,...) or Y = random(name,A,[m,n,...]) returns an m-by-n-by... matrix of random numbers.

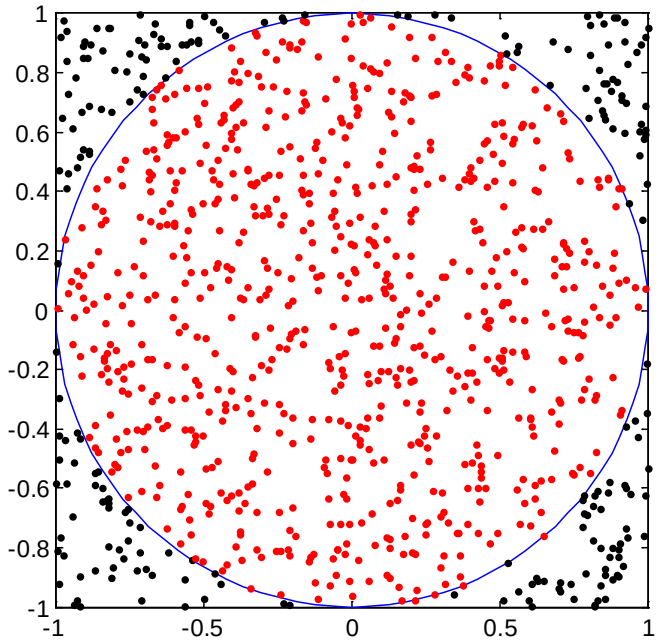
Generació de m columnes de N nombres aleatoris amb la instrucció random per diferents pdfs

```
>> G=random('normal',mu,sigma,N,m);
>>B=random('Binomial',n,p,N,m)
>> P=random('Poisson',mu,N,m);
>> U=random('Uniform',a,b,N,m);
>> Ch2=random('chi2',nu,N,m)
>> Exp=random('Exponential',tau;N,m)
```

Exemples

Exemple 1: Càlcul de l'**àrea** d'un cercle de radi 1 m per Montecarlo.

```
R=1;
N=1000;
P=2*rand(N,2)-1;
x=P(:,1);
y=P(:,2);
Nc=0;
t=0:pi/50:2*pi;
plot(sin(t),cos(t))
axis square
hold on
for i=1:N
if (x(i)^2+y(i)^2)<R^2
   Nc=Nc+1;
   plot(x(i),y(i),'r.')
else
   plot(x(i),y(i),'k.')
end
end
hold off
A=(2*R)^2*Nc/N
Sigma=(2*R)^2*(sqrt(Nc*(1-Nc/N)))/N
fprintf('Area= %5.3f m2  Sigma= %5.3f  m2 \n',A,Sigma)
```



Area= 3.112 m2 Sigma= 0.053 m2

Exemples

Exemple 2: En un cert instant de temps produïm una mostra pura d'un isòtop radioactiu. Aquest isòtop, que anomenarem X, es desintegra a un isòtop Y amb una vida mitjana d'una hora, i així mateix l'isòtop Y es desintegra a un altre isòtop Z amb una vida mitjana de 30 minuts. Finalment, l'isòtop Z es desintegra amb una vida mitjana de 20 minuts, a un isòtop estable W. Desitgem saber la proporció dels distints isòtops després d'un temps t de la seua producció, en particulars als valors, t=10 min, t=60 min i t=2h.

$$X \xrightarrow[1h]{} Y \xrightarrow[30\,\mathrm{min}]{} Z \xrightarrow[20\,\mathrm{min}]{} W$$

El problema es pot resoldre a partir d'un sistema d'equacions diferencials acoblades per a X, Y i Z en funció del temps, que té solució analítica (*equacions de Bateman*). Des del punt de vista de la simulació, però, la solució és força simple:

Planifiquem N proves, que és equivalent a començar en el temps inicial amb N nuclis del pare X.

1. Generem un nombre aleatori t1 de la pdf exponencial amb constant TX que correspon al moment en què es desintegrarà el nucli X elegit. Si t1<t, aquest nucli s'haurà desintegrat després d'un temps t, i disminuirem X en una unitat i augmentarem Y en una unitat. En cas contrari tornem al principi.
2. En el cas de la desintegració de X en t1 generaríem un nombre aleatori t2 amb una pdf exponencial amb constant TY per veure si Y s'ha desintegrat en t. La qual cosa és certa si t1+t2 <t; en aquest cas disminuïm Y en una unitat i augmentem Z en una unitat. En cas contrari tornem a 1.
3. En el cas de la desintegració de Y en t2, generem t3 amb pdf exponencial amb constant TZ. Si t1+t2+t3>t, Z s'haurà desintegrat en t i disminuïm Z en una unitat i augmentem W en una unitat. Com que W és estable tornem a 1 per a una nova prova.

Exemples

Montecarlo amb MATLAB per a simular la desintegració encadenada:

```
%Cadena de desintegracions
t=input('Entreu valor de t en minuts:  ');
N=input('Nombre de proves:   ');
%Vidas mitjanes en minuts
Tau=[60,30,20];
%Valors inicials del nombre d'isòtops de cada generació
N1=N;N2=0;N3=0;N4=0;
for i=1:N
   t1=random('exponential',Tau(1));
if t1<t
   N1=N1-1;N2=N2+1;
   t2=random('exponential',Tau(2));
   if t1+t2<t
      N2=N2-1;N3=N3+1;
      t3=random('exponential',Tau(3));
       if t1+t2+t3<t
         N3=N3-1;N4=N4+1;
       end
   end
end
end
fprintf('Isòtop\t Proporció \t  Error  \n')
fprintf(' Pare\t %6.3f\t   %8.3f\n ',N1/N,sqrt(N1)/N)
fprintf('Fill\t %6.3f\t   %8.3f\n ',N2/N,sqrt(N2)/N)
fprintf('Net\t %6.3f\t   %8.3f\n ',N3/N,sqrt(N3)/N)
fprintf('Besnét\t %6.3f\t   %8.3f\n ',N4/N,sqrt(N4)/N)
```

$$X \xrightarrow[1h]{} Y \xrightarrow[30\,\mathrm{min}]{} Z \xrightarrow[20\,\mathrm{min}]{} W$$

L'error de cada recompte es calcula sabent que la distribució de desintegracions és poissoniana amb $\sigma=\sqrt{N}$, sent N el nombre observat de desintegracions.

Entreu valor de t en minuts: 60
Nombre de proves: 10000

Isòtop	Proporció	Error
Pare	0.369	0.006
Fill	0.235	0.005
Nét	0.145	0.004
Besnét	0.252	0.005