

4. Test de hipòtesis i teoria de la decisió

Lluís Garrido

garrido@ecm.ub.es

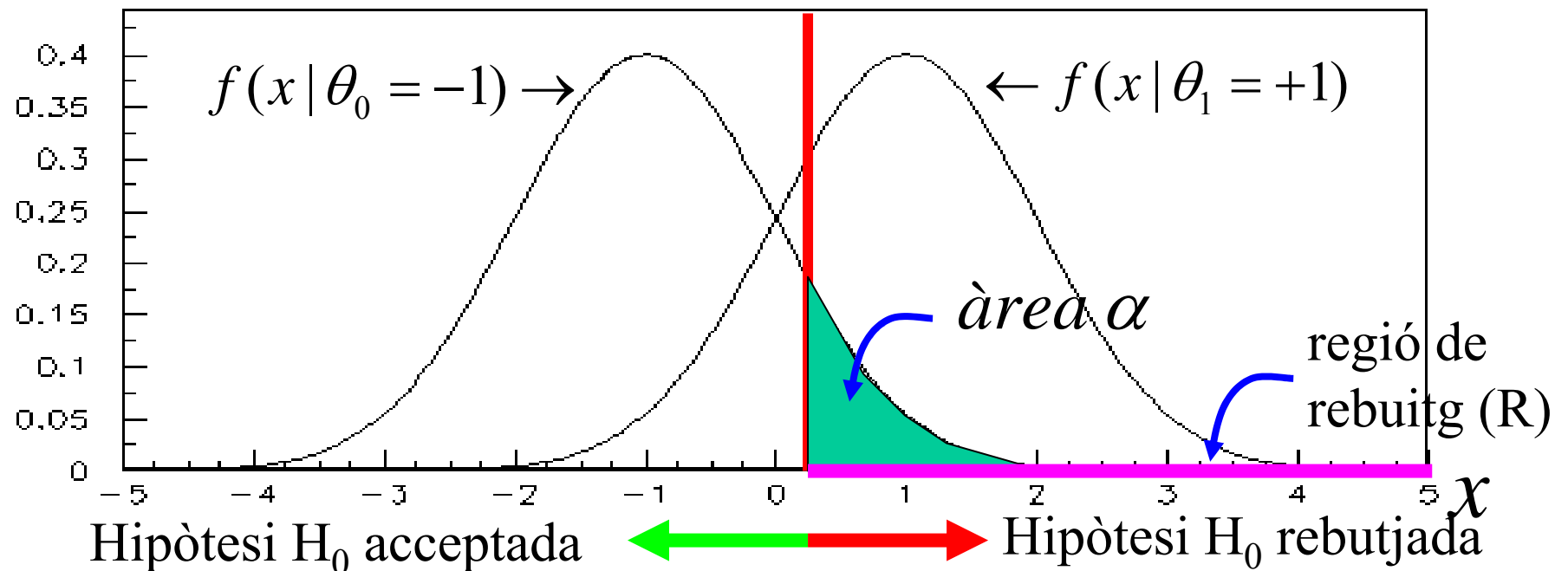
<http://www.ecm.ub.es/~garrido>

índex

- test de hipòtesi
 - mesura de la bondat d' un ajust
- teoria de la decisió
 - Separació senyal/background
- Informació de Kullback i de Shannon
 - Definició i propietats.
 - exemples d'aplicació

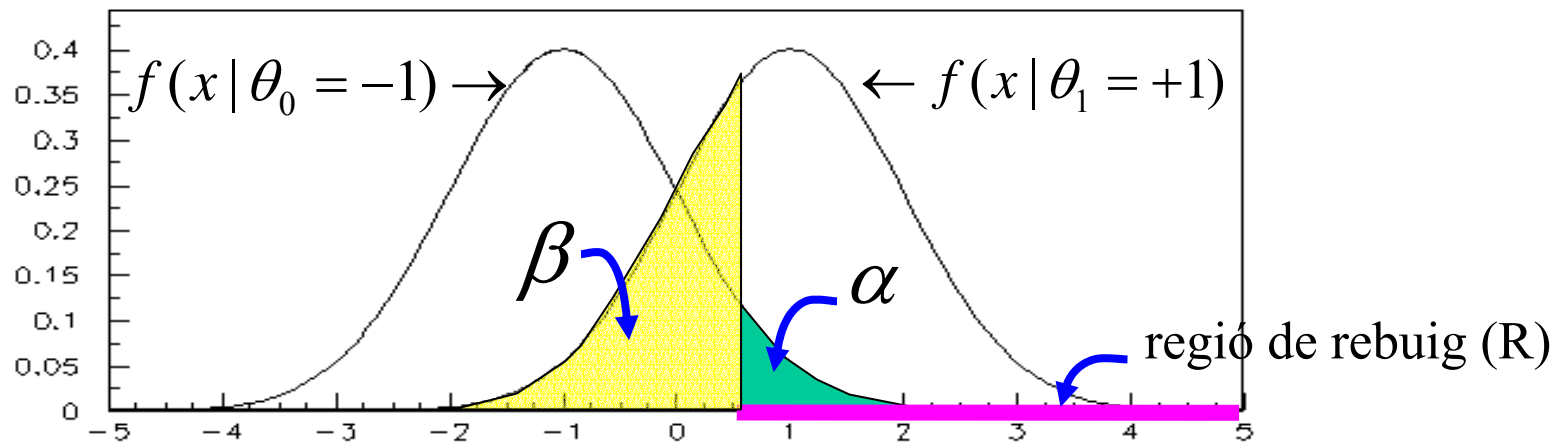
Test de hipòtesi

- Volem acceptar o rebutjar una hipòtesi H_0 (que la denominarem hipòtesi nul) contra una altra hipòtesi alternativa H_1 , basant-nos en les dades experimentals, i a un determinat “nivell de significació” α (probabilitat de que rebutgem H_0 siguin veritat. Voldrem α petit)
- exemple: sigui x un estadístic de les nostres mesures. Volem acceptar o rebutjar que la p.d.f de x es $f(x | \theta_0)$ en front de $f(x | \theta_1)$, una vegada hem obtingut el resultat x i a un nivell de significació donat α



Test d'hipòtesi (2)

- Existeix també la possibilitat que H_1 sigui veritat, però que acceptem H_0 . Això passarà amb una probabilitat β



$$P(x \in R | H_0) = \alpha \quad P(x \in R | H_1) = 1 - \beta$$

- ens interessa que β sigui el més petit possible. Per això definim:
Potència del test $\equiv 1 - \beta$
- Si tenim dos mètodes diferents per discriminar θ_0 de θ_1 evidentment agafarem aquell que ens doni una potència més gran

exemple

- Tenim un experiment que ens dóna esdeveniments que segueixen una distribució $N(\mu_0, \sigma^2)$. Fem un canvi en l'aparell i volem veure si el valor central no ha canviat mesurant n nous esdeveniments (considerem el cas en que σ no canvia i és coneguda).

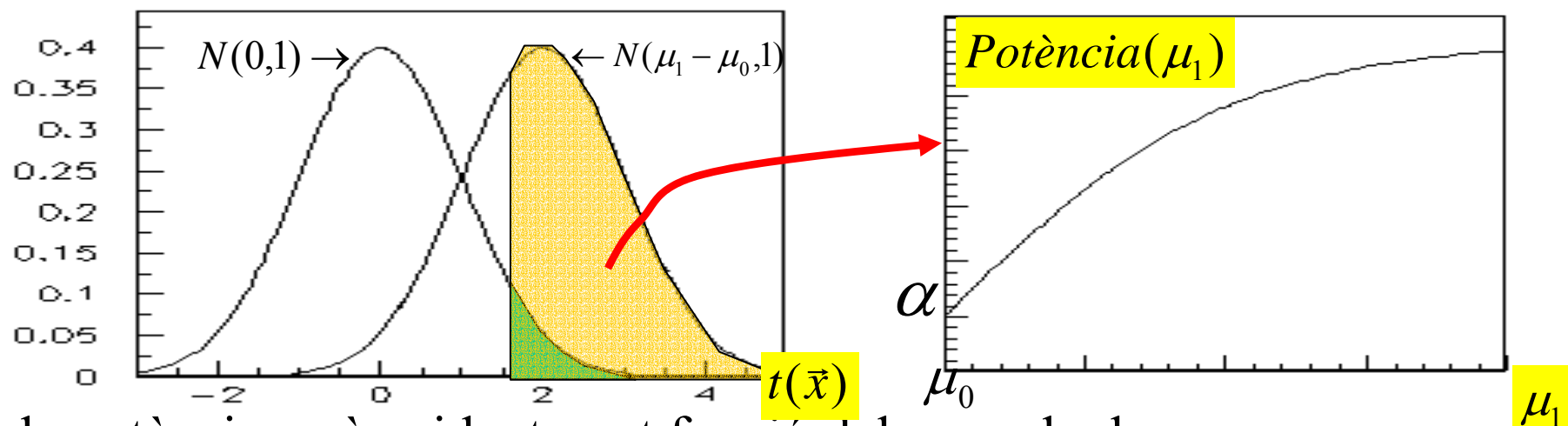
- $H_0: \mu = \mu_0$

- $H_1: \mu = \mu_1, (\mu_1 > \mu_0)$

- agafarem l'estadístic:

$$t(\bar{x}) = \left(\frac{1}{n} \sum_{i=1}^n x_i - \mu_0 \right) / (\sigma / \sqrt{n}) \Rightarrow p.d.f(t) = \begin{cases} N(0,1) & \text{si } \mu = \mu_0 \\ N(\mu_1 - \mu_0, 1) & \text{si } \mu = \mu_1 \end{cases}$$

- avaluant $t(x)$ podrem acceptar o rebutjar la hipòtesi H_0 contra H_1 al nivell de significació α que volem



- la potència serà evidentment funció del μ_1 amb el que comparem

Exemple (2)

- Però també podem utilitzar un altre mètode: el del signe
 - $H_0: \mu = \mu_0$ les n observacions estaran distribuïdes simètricament al voltant de μ_0
 - $H_1: \mu = \mu_1$ i no hi haurà simetria al voltant de μ_0
- sigui n_- el número de observacions en que $(x_i - \mu_0)$ sigui negatiu. (si H_0 no és veritat aquest número tendirà a ser petit comparat amb $n/2$)

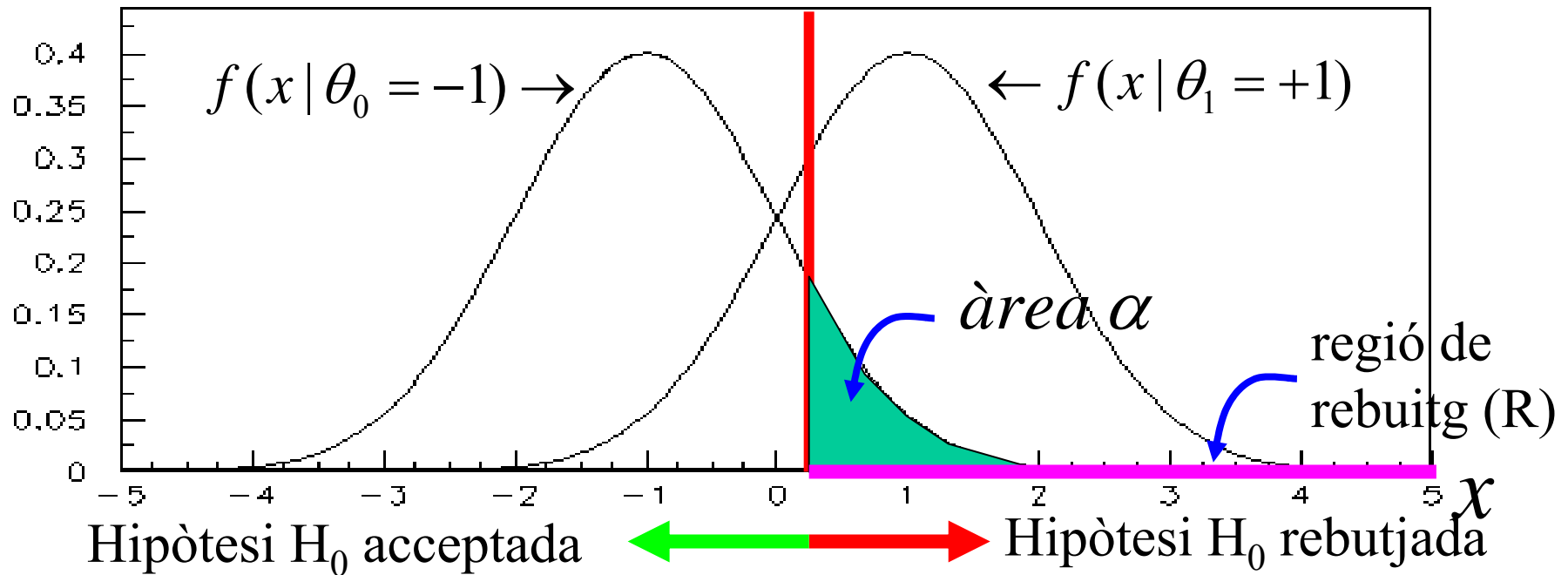
$$P(n_- = r | H_0) = \binom{n}{r} \left(\frac{1}{2}\right)^n \quad P(N_- = r | H_1) = \binom{n}{r} q^r (1 - q)^{n-r}$$

- com la pd.f és discreta en lloc de donar α donarem un n_{\min} . Si n_- és més petit que n_{\min} rebutjarem H_0 .

$$q = \int_{-\infty}^{\mu_0} N(\mu_1, \sigma) dx$$
- El nivell de significació determinat per aquesta n_{\min} serà: $\alpha = \sum_{i=0}^{n_{\min}} P(i | H_0)$
- i la potència del test serà: $1 - \beta = \sum_{i=0}^{n_{\min}} P(i | H_1)$
- Aquesta potència, com en el cas anterior depèn de μ_1
- Comparant els dos mètodes, per la mateixa α , resulta que el primer mètode té una potència més gran per qualsevol valor de μ_1

Test de hipòtesis a la pràctica

Correcta: donar primer α , fer la mesura d' x i acceptar o rebutjar H_0



Habitual: fer la mesura d' x i donar $\alpha = \int_x^{+\infty} f(y | \theta_0) dy$

Aquesta α dona el “nivell de confiança” (C.L.) de que H_0 sigui certa.

Exemple

Diferència en les mitjanes (σ iguals però desconegudes): test de student

Suposem que tenim dues mostres:

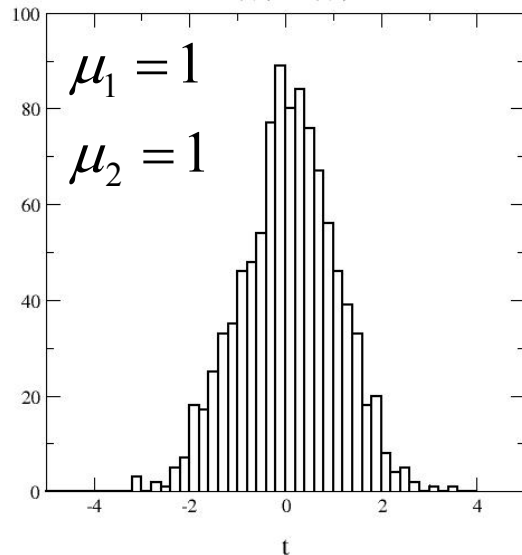
- $\{x_1, \dots, x_N\}$
- $\{y_1, \dots, y_M\}$

les quals sabem han estat **generades a partir de distribucions Normals amb la mateixa variància** però que poden tenir mitjanes diferents.

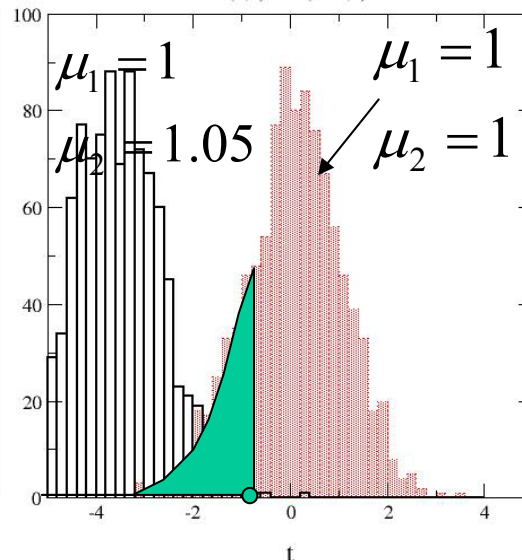
$$t = \frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{N} + \frac{1}{M}}} \quad S^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2 + \sum_{j=1}^M (y_j - \bar{y})^2}{N + M - 2}$$

student amb $N+M-2$ graus de llibertat

1000 parells de mostres de 100 valors
 $N(1,1)$ i $N(1,1)$



1000 parells de mostres de 100 valors
 $N(1,1)$ i $N(1.05,1)$

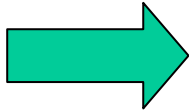
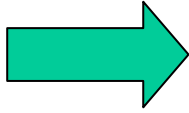


Mètode:

1) mesurar t_{obs}

2) $C.L. = \int_{-\infty}^{t_{obs}} t_{stud}(x; N + M - 2) dx$

Diferents test involucrant mitjanes i variàncies de distribucions Normals sota les condicions especificades



H_0	Conditions	Test statistic	Test distribution
$\mu = \mu_0$	σ^2 known	$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$N(0,1)$
	σ^2 unknown	$\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t(n-1)$
$\sigma^2 = \sigma_0^2$	μ known	$(n-1)s^2/\sigma_0^2 = \sum_{i=1}^n (x_i - \mu)^2/\sigma_0^2$	$\chi^2(n)$
	μ unknown	$(n-1)s^2/\sigma_0^2 = \sum_{i=1}^n (x_i - \bar{x})^2/\sigma_0^2$	$\chi^2(n-1)$
$\mu_1 - \mu_2 = 0$	$\sigma_1^2 = \sigma_2^2 = \sigma^2$ known	$\frac{\bar{x} - \bar{y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$	$N(0,1)$
	σ_1^2 and σ_2^2 known	$\frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$	
	$\sigma_1^2 = \sigma_2^2 = \sigma^2$ unknown	$\frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$ $S^2 \equiv \frac{1}{n+m-2} \left((n-1)s_1^2 + (m-1)s_2^2 \right)$	$t(n+m-2)$
	$\sigma_1^2 + \sigma_2^2$ unknown	$\frac{\bar{x} - \bar{y}}{\sqrt{s_1^2/n + s_2^2/m}}$	not exactly known, $\approx N(0,1)$
$\frac{\sigma_1^2}{\sigma_2^2} = 1$	μ_1 and μ_2 known	$\frac{s_1^2}{s_2^2} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_1)^2 / \frac{1}{m-1} \sum_{i=1}^m (y_i - \mu_2)^2$	$F(n,m)$
	μ_1 and μ_2 unknown	$\frac{s_1^2}{s_2^2} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 / \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2$	$F(n-1, m-1)$

Test de Neyman-Pearson

- En p.d.f. unidimensionals el problema de trobar la regió de rebuig “R” és fàcil: són els extrems. Però i en més dimensions?
- Mètode general per trobar la millor regió de rebuig (o màxima potencia del test) a un determinat nivell de significació donat:

$$\alpha = \int_R f(\bar{x} | \theta_0) d\bar{x} \quad 1 - \beta = \int_R f(\bar{x} | \theta_1) d\bar{x}$$

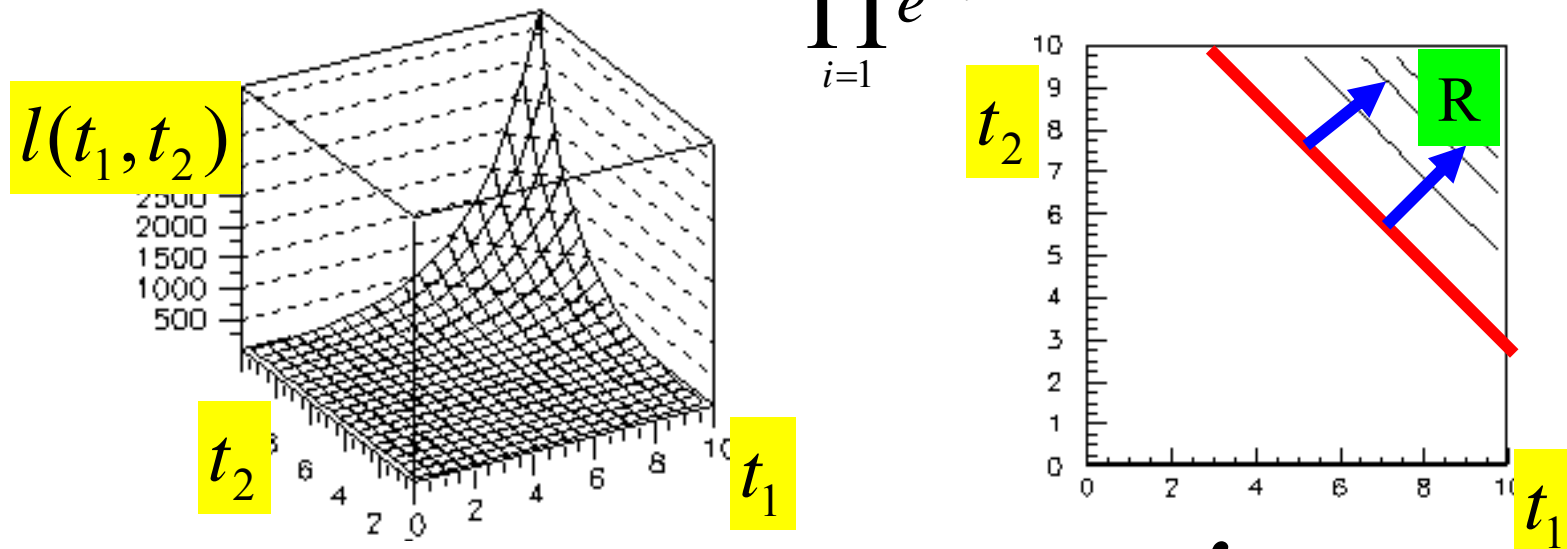
$$1 - \beta = \int_{R(\alpha)} \frac{f(\bar{x} | \theta_1)}{f(\bar{x} | \theta_0)} f(\bar{x} | \theta_0) d\bar{x} = E_R \left(\frac{f(\bar{x} | \theta_1)}{f(\bar{x} | \theta_0)} \middle| \theta = \theta_0 \right)$$

- serà màxim si i només si R és la fracció de l'espai que compta els valors majors de $f(\bar{x} | \theta_1) / f(\bar{x} | \theta_0)$
- la regió crítica vindrà donada per $l_n(\bar{x}; \theta_0, \theta_1) \equiv \frac{f(\bar{x} | \theta_1)}{f(\bar{x} | \theta_0)} \geq C_\alpha$
 C_α a determinar segons la α donada
- el criteri serà
 - acceptar H0 si $l_n < C_\alpha$
 - rebutjar H0 si $l_n \geq C_\alpha$

exemple

- Volem discriminar entre dos possibles valors de la vida mitjana d'una partícula a un cert nivell de significació α . $H_0: \tau = 1$, $H_1: \tau = 2$

- farem dos mesures t_1 i t_2
- $$l(t_1, t_2) = \frac{L(t_1, t_2 | \tau = 2)}{L(t_1, t_2 | \tau = 1)} = \frac{\prod_{i=1}^2 e^{-t_i/2} / 2}{\prod_{i=1}^2 e^{-t_i}} = \frac{1}{4} e^{(t_1+t_2)/2} > C_\alpha$$



- per trobar C_α hem de buscar la regió R on $\alpha = \int_R L(\vec{t}, 1) d\vec{t}$

$$1 - \alpha = \iint_{t_1+t_2 < 2 \ln(4C_\alpha)} e^{-t_1} e^{-t_2} dt_1 dt_2 \Rightarrow \begin{cases} \text{acceptar } H_0 & \text{si } (t_1 + t_2) < 2 \ln(4C_\alpha) \\ \text{rebutjar } H_0 & \text{si } (t_1 + t_2) > 2 \ln(4C_\alpha) \end{cases}$$

Maximitzant la potència local

- Hem estimat θ_0 màximitzant $L(\bar{x} | \theta)$, i volem discriminar entre
 - H0: $\theta = \theta_0$
 - H1: $\theta = \theta_0 + \Delta$
- si apliquem el test de Neyman-Pearson:

$$l_n(\bar{x}; \theta_0, \theta_0 + \Delta) = \frac{L(\bar{x} | \theta_0 + \Delta)}{L(\bar{x} | \theta_0)} \geq C_\alpha \Rightarrow \ln L(\bar{x} | \theta_0 + \Delta) - \ln L(\bar{x} | \theta_0) \approx \left. \frac{\partial \ln L}{\partial \theta} \right|_{\theta_0} \Delta \geq \ln C_\alpha$$

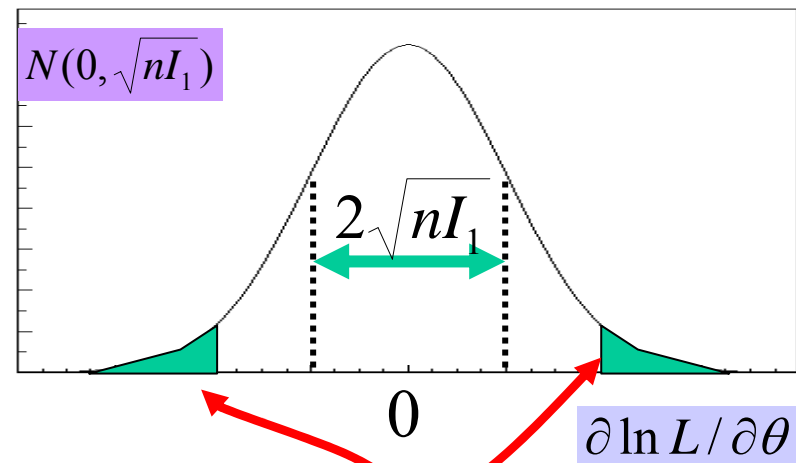
$$\left. \frac{\partial \ln L}{\partial \theta} \right|_{\theta_0} \geq \frac{\ln C_\alpha}{\Delta} \Rightarrow \left. \frac{\partial \ln L}{\partial \theta} \right|_{\theta_0} > k_\alpha \quad k_\alpha \equiv \left| \frac{\ln C_\alpha}{\Delta} \right|$$

- i sabem que $\partial \ln L / \partial \theta$ és una distribució normal amb:

$$E\left(\left. \frac{\partial \ln L}{\partial \theta} \right|_{\theta=\theta_0}\right) = 0 \quad E\left(\left(\left. \frac{\partial \ln L}{\partial \theta} \right|_{\theta=\theta_0}\right)^2 - 0\right) = nI^{(1)}$$

- es test serà:

$$\left| \left. \frac{\partial \ln L}{\partial \theta} \right|_{\theta_0} \right| > \lambda_\alpha \sqrt{nI_1}$$



- exemple: si $\lambda_\alpha = 2 \Rightarrow 5\%$ de nivell de significació

El test de Pearson

- Tenim n successos que poden caure en k llocs amb probabilitats p_i (típic histograma de k intervals)

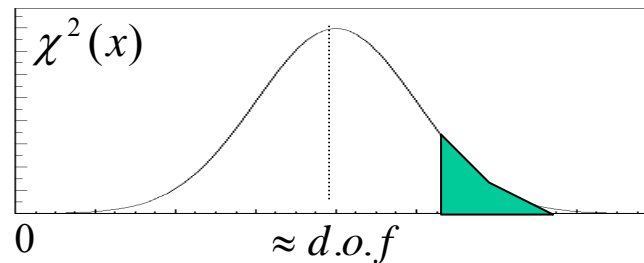
- volem acceptar o rebutjar $H_0: p_1 = \tilde{p}_1 \cdots p_k = \tilde{p}_k$

- construïm
$$\chi^2 \equiv \sum_{i=1}^k \frac{(n_i - n\tilde{p}_i)^2}{n\tilde{p}_i}$$

- si n_i és gran i H_0 és veritat tindrem que $(n_i - n\tilde{p}_i)^2 / n\tilde{p}_i \approx N(0,1)$ i per tant estem davant d'una χ^2 de “ $k-1$ ” d.o.f (“ $k-1$ ” i no “ k ” degut a la condició de normalització: $\sum_{i=1}^k n_i = n$). Si les p_i han estat estimades a partir de “ d ” paràmetres, la χ^2 serà de “ $k-1-d$ ” d.o.f

- test: per a una α de nivell de significació, la regió de rebuig vindrà donada per

$$\alpha = \int_{\chi_\alpha^2}^{\infty} \chi^2(x; d.o.f) dx$$

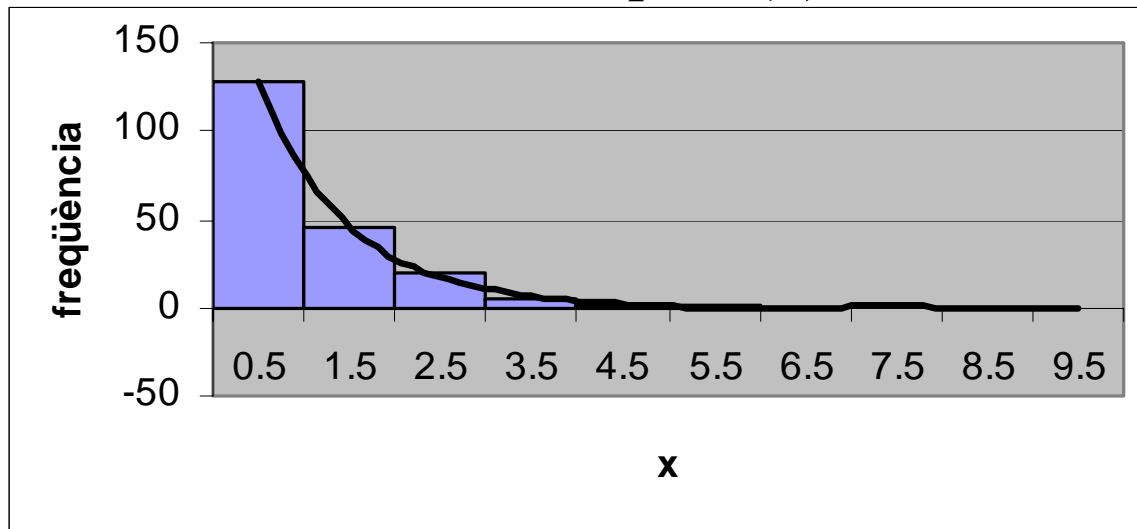


- pràctica usual en física es donar un C.L. Com

$$\int_{\chi_{obs}^2}^{\infty} \chi^2(x; d.o.f) dx$$

El test de Pearson: cas 1

Comparar dades (n successos que poden caure en k intervals) amb un model donat de pdf $f(x)$



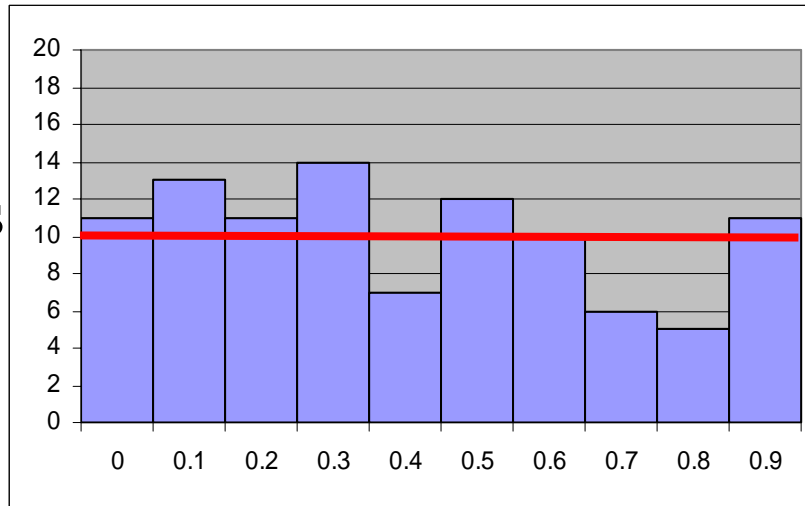
$$\tilde{p}_i = \int_{a_i}^{a_{i+1}} f(x) dx$$

$$\chi_{obs}^2 = \sum_{i=1}^k \frac{(n_i - n\tilde{p}_i)^2}{n\tilde{p}_i} \Rightarrow C.L. = \int_{\chi_{obs}^2}^{\infty} \chi^2(x; d.o.f) dx$$

$$dof = \begin{cases} k & \text{si } \mathbf{n} \text{ no esta fixat} \\ k-1 & \text{si } \mathbf{n} \text{ esta fixat} \\ k-1-d & \text{si } f(x) \text{ té } \mathbf{d} \text{ paràmetres que hem ajustat a les dades} \end{cases}$$

Exemple (dof=k-1)

Generem
n=100
nombres plans
i els agrupem
en k=10
interval·s.



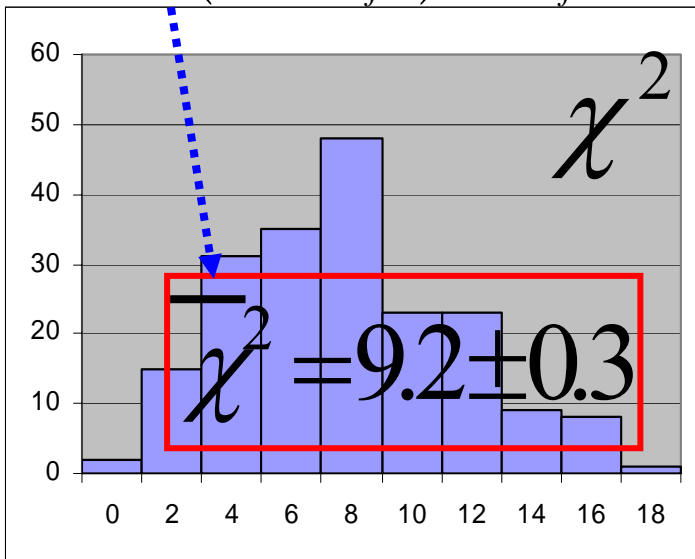
$$\tilde{p}_i = \int_{a_i}^{a_{i+1}} dx = 0.1$$

$$\chi_{obs}^2 = \sum_{i=1}^{10} \frac{(n_i - 10)^2}{10} = 11.35$$

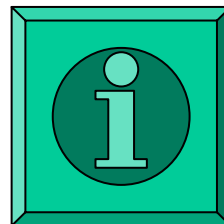
$$C.L. = \int_{11.35}^{\infty} \chi^2(x; 9) dx = 0.25$$

Podem repetir l'experiment 200 vegades i veure les distribucions de χ^2 i CI

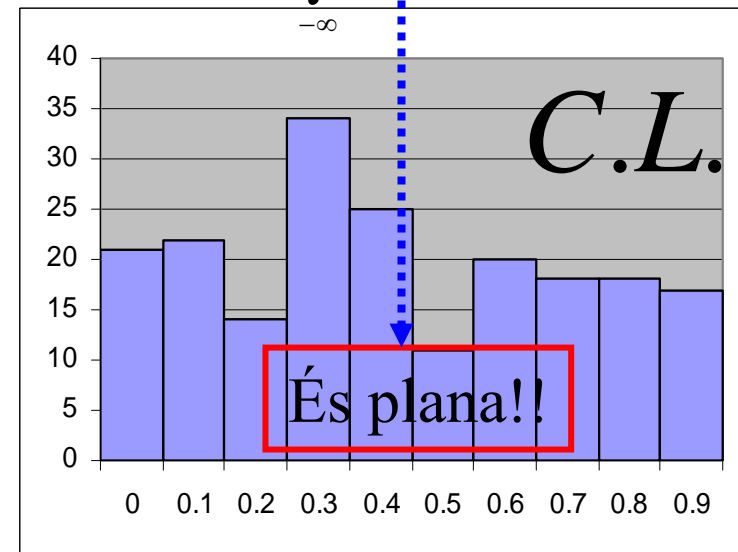
$$E(\chi^2(n_{dof})) = n_{dof}$$



Demo

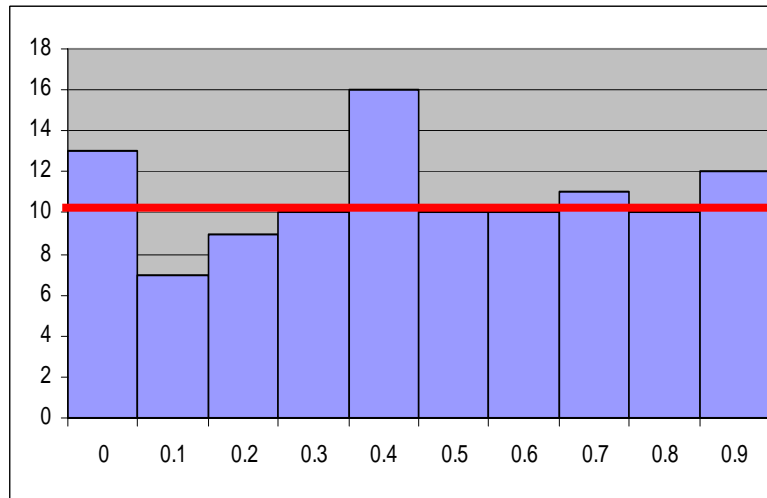


$$CL = 1 - \int_{-\infty}^{\chi_{obs}^2} \chi^2(x; d.o.f) dx$$



Exemple (dof=k)

Generem n
nombres plans
i els agrupem
en k=10
intervalos.
(n poisson
amb $\mu=100$)

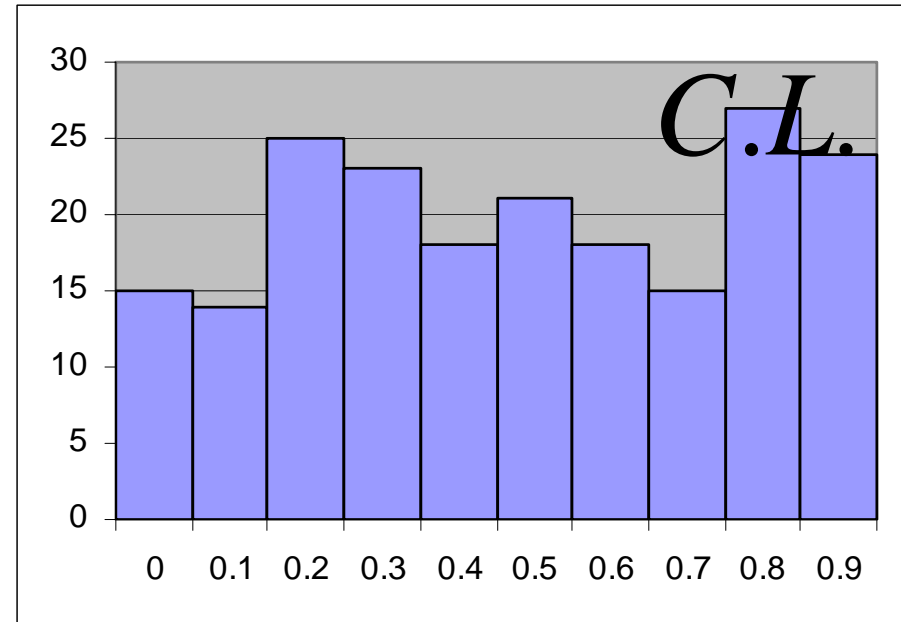
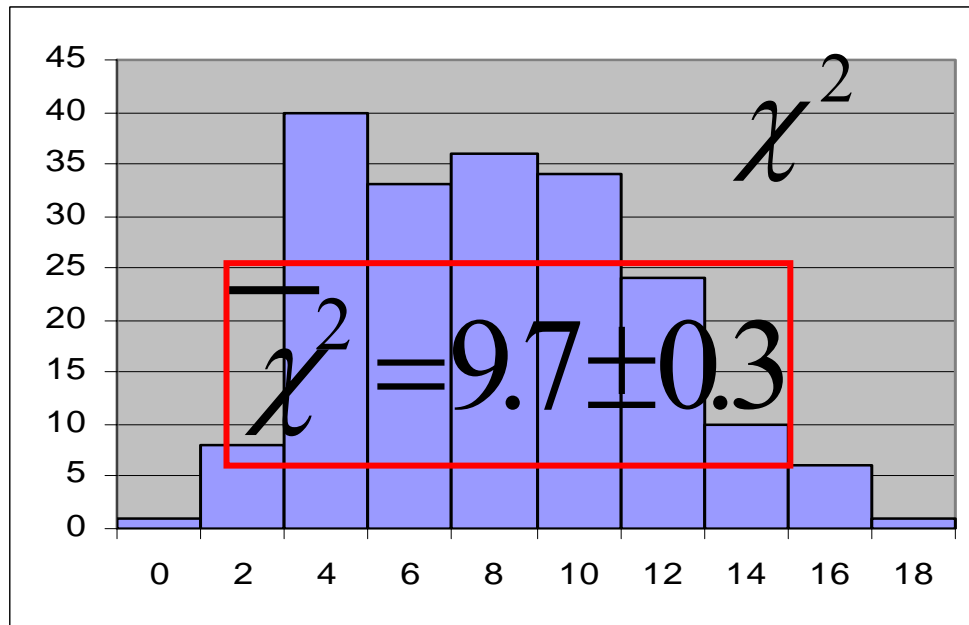


$$\tilde{p}_i = \int_{a_i}^{a_{i+1}} dx = 0.1$$

$$\chi^2_{obs} = \sum_{i=1}^{10} \frac{(n_i - 10)^2}{10} = 2.9$$

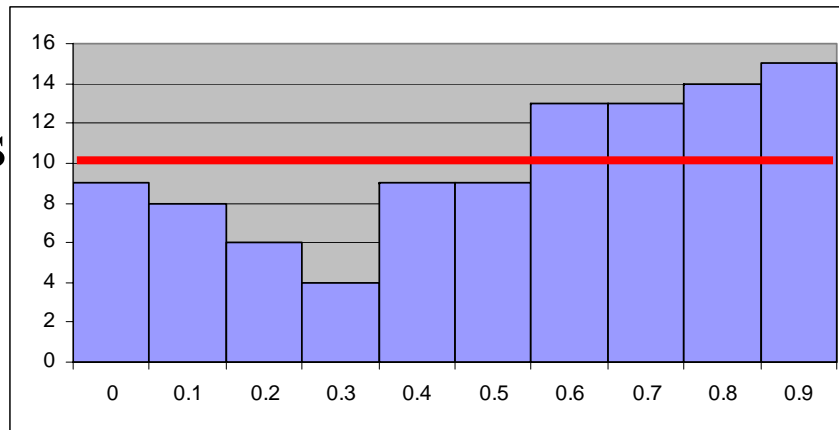
$$C.L. = \int_{11.35}^{\infty} \chi^2(x; 10) dx = 0.96$$

Podem repetir l'experiment 200 vegades i veure les distribucions de χ^2 i CI



Exemple (dof=k-1-d)

Generem 100 nombres plans i els agrupem en k=10 intervals.



$$\chi^2_{obs} = \sum_{i=1}^{10} \frac{(n_i - 10)^2}{10} = 11.8$$

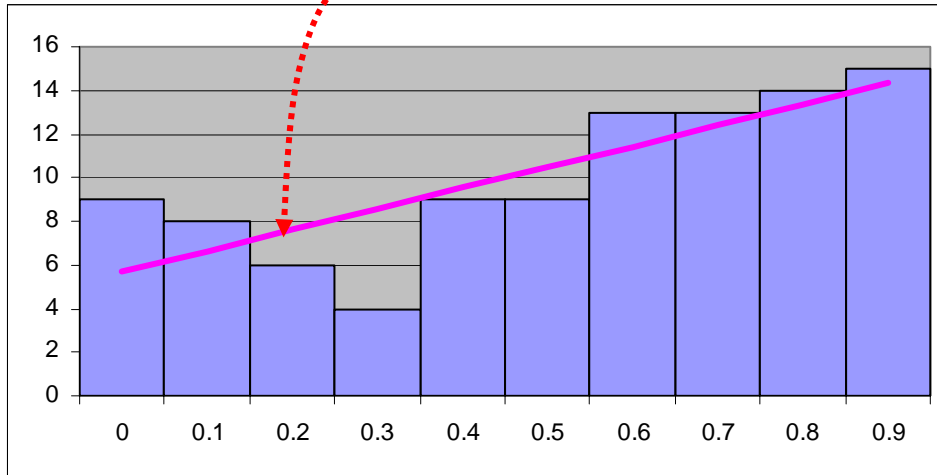
$$C.L. = \int_{11.35}^{\infty} \chi^2(x; 9) dx = 0.22$$

Però ara creiem que aquestes dades, en lloc de provenir de la pdf $f(x)=1$, venen de la pdf $f'(x; \alpha) = \alpha(x - 0.5) + 1$ $0 \leq x \leq 1$ d'on desconeixem α i per tant l'hem d'estimar.

$$\chi^2(\alpha) = \sum_{i=1}^{10} \frac{(n_i - n\tilde{p}_i(\alpha))^2}{n_i} \xrightarrow{\text{minimitzar}} \hat{\alpha}$$

$$\tilde{p}_i(\alpha) = \int_{a_i}^{a_{i+1}} (\alpha(x - 0.5) + 1) dx$$

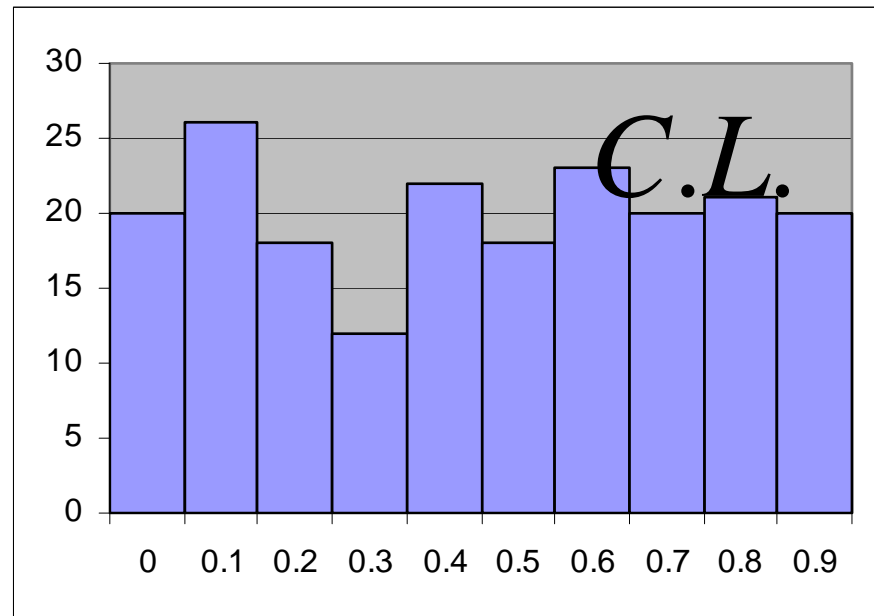
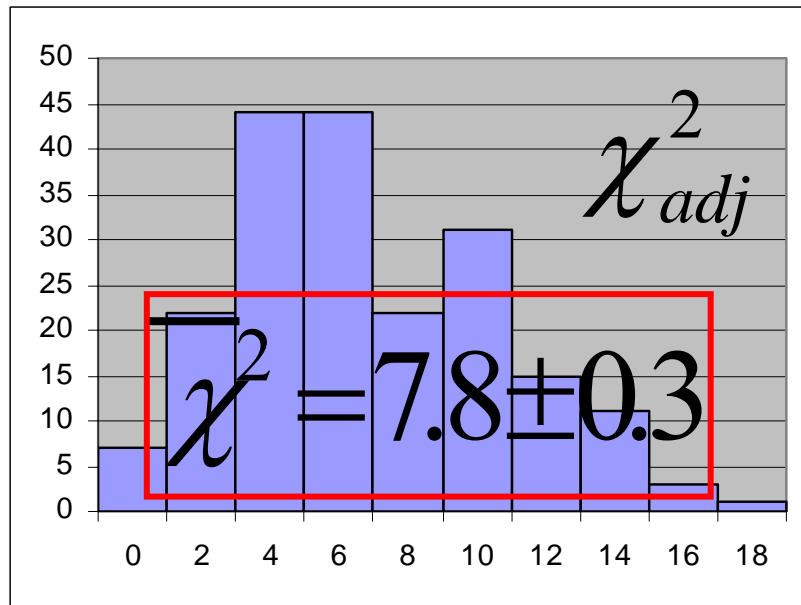
millor ajust



$$\chi^2_{adj} = \sum_{i=1}^{10} \frac{(n_i - n\tilde{p}_i(\hat{\alpha}))^2}{n_i} = 4.2$$

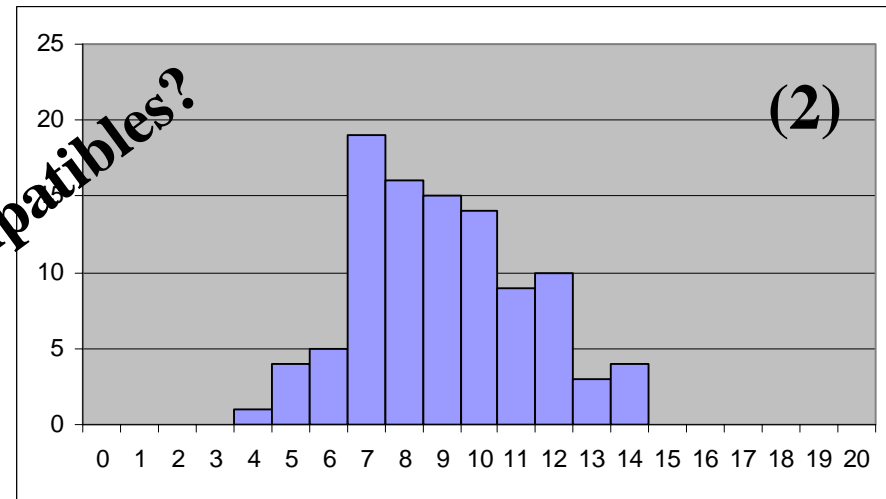
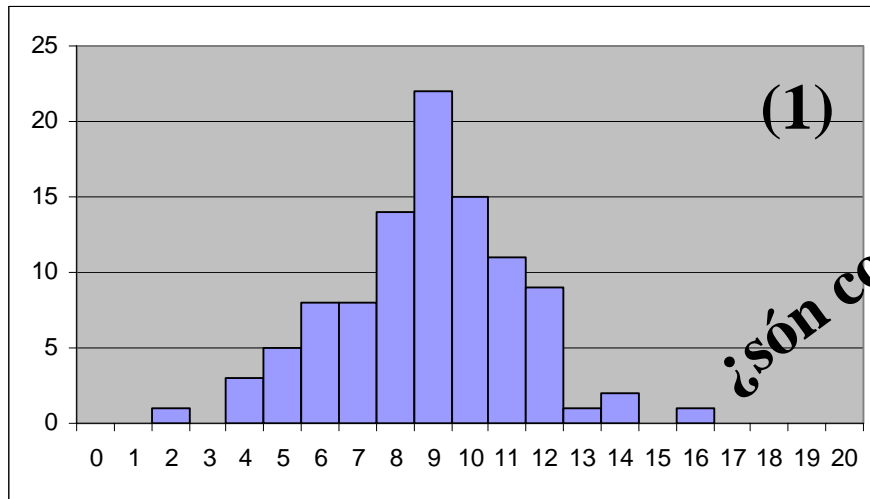
$$C.L. = \int_{11.35}^{\infty} \chi^2(x; 8) dx = 0.83$$

Podem repetir l'experiment 200 cops i veure les distribucions de χ^2_{adj} i CL



El test de Pearson: cas 2

Volem saber si el resultat dels dos mostrejos és compatible amb el que seria esperable si ambdós segueixen la mateixa distribució de probabilitat $f(x)$.



(de fet totes dues han estat generades segons una binomial $n=20, p=0.5$)

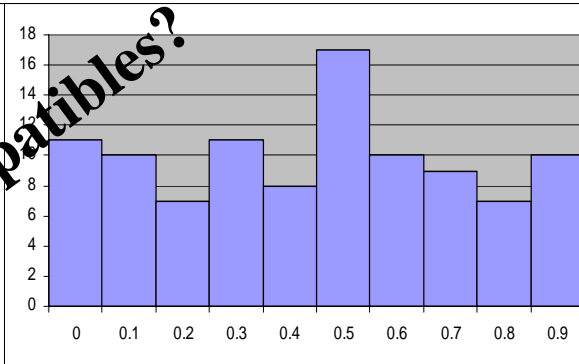
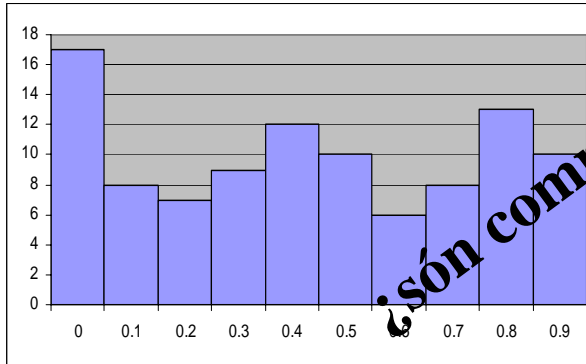
$$\chi_{obs}^2 = \sum_{i=1}^k \frac{(n_i^{(1)} - n_i^{(2)})^2}{n_i^{(1)} + n_i^{(2)}} \Rightarrow C.L. = \int_{\chi_{obs}^2}^{\infty} \chi^2(x; d.o.f) dx$$

cal ometre els termes amb $n_i^{(1)} = n_i^{(2)} = 0$

$$dof = \begin{cases} k & \text{si les dades s'han recollit **sense imposar** que } n_{TOT}^{(1)} = n_{TOT}^{(2)} \\ k-1 & \text{si les dades s'han recollit **imposant** que } n_{TOT}^{(1)} = n_{TOT}^{(2)} \end{cases}$$

Exemple

Generem dos conjunts de 100 nombres plans agrupats en k=10 intervals.

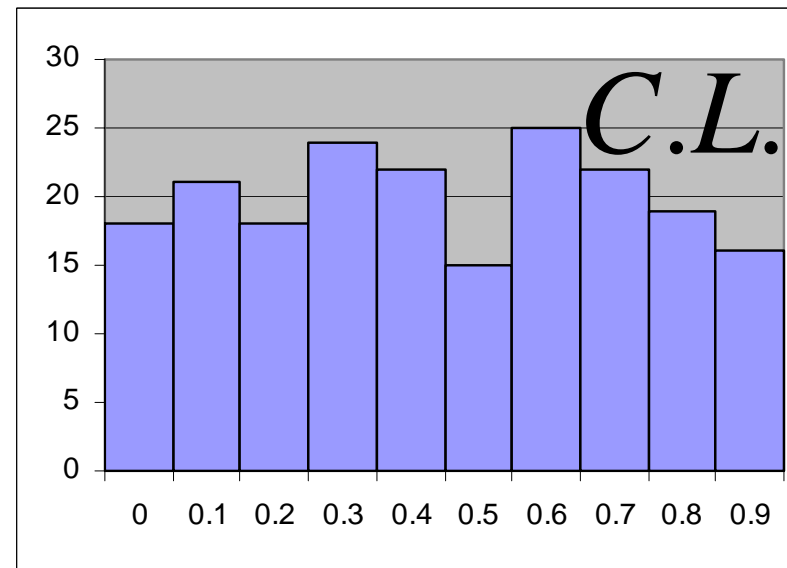
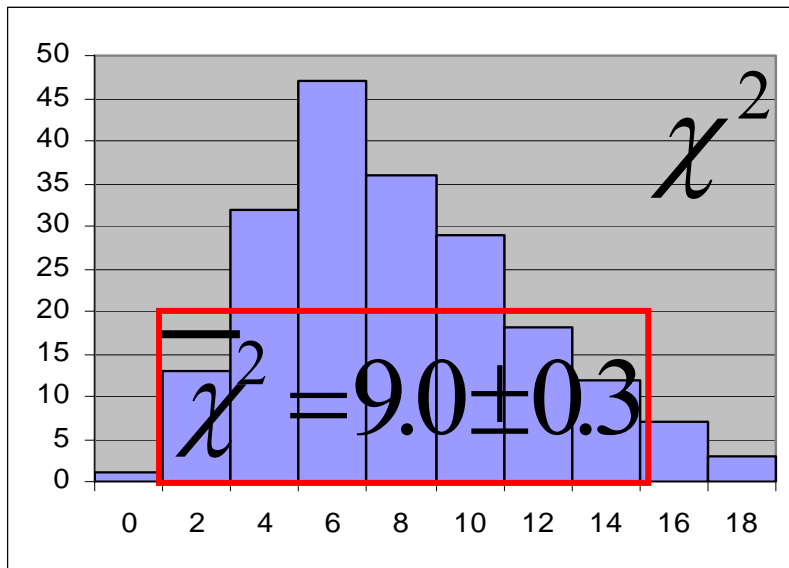


¿són compatibles?

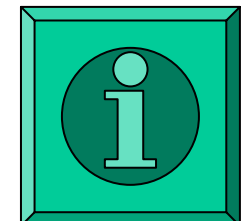
$$\chi^2_{obs} = \sum_{i=1}^k \frac{(n_i^{(1)} - n_i^{(2)})^2}{n_i^{(1)} + n_i^{(2)}} = 8.4$$

$$C.L. = \int_{\chi^2_{obs}}^{\infty} \chi^2(x; 9) dx = 0.5$$

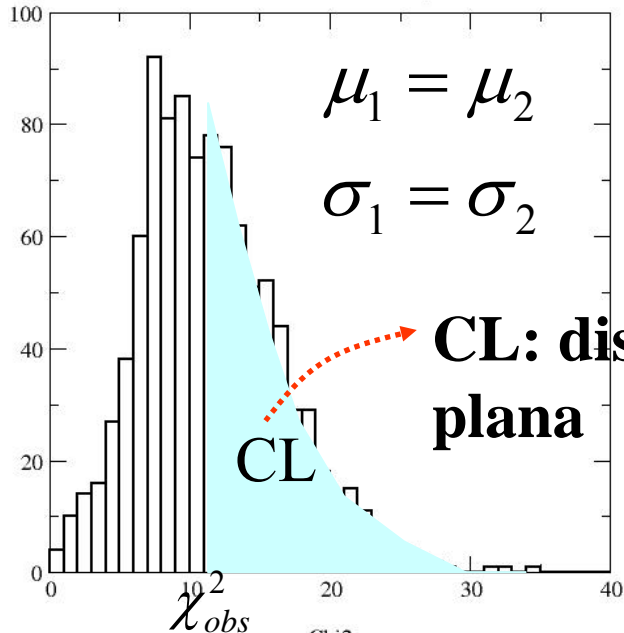
Podem repetir l'experiment 200 vegades i veure les distribucions de χ^2 i CI



Demo



1000 parells de mostres de 1000 elements $N(1,1) - N(1,1)$
20 intervals de mostreig

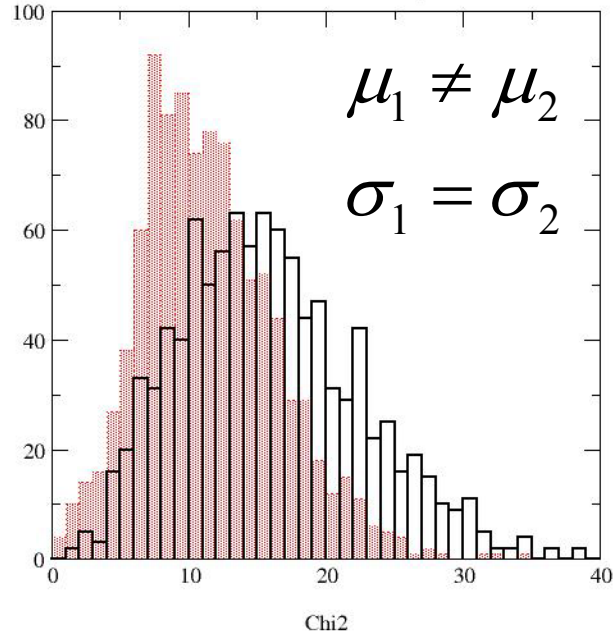


Compatibilitat entre Normals

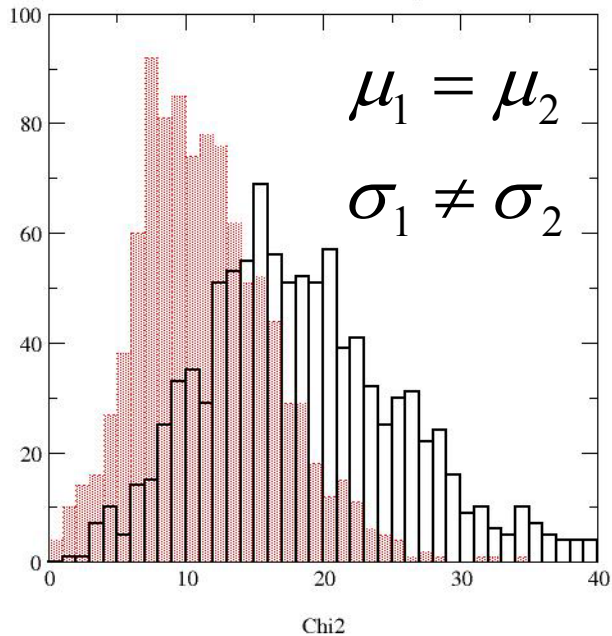
$$\chi_{obs}^2 = \sum_{i=1}^k \frac{(n_i^{(1)} - n_i^{(2)})^2}{n_i^{(1)} + n_i^{(2)}}$$

$$C.L. = \int_{\chi_{obs}^2}^{\infty} \chi^2(x; d.o.f) dx$$

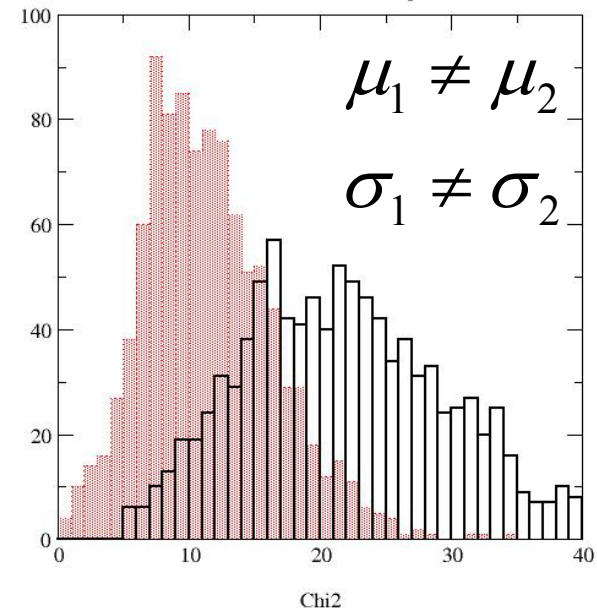
1000 parells de mostres de 1000 elements $N(1,1) - N(1.1,1)$
20 intervals de mostreig



1000 parells de mostres de 1000 elements $N(1,1) - N(1,1.1)$
20 intervals de mostreig



1000 parells de mostres de 1000 elements $N(1,1) - N(1.1,1.1)$
20 intervals de mostreig



En els altres 3 casos CL no tindrà distribució plana (estarà picat cap a 0)

Test free of binning(I)

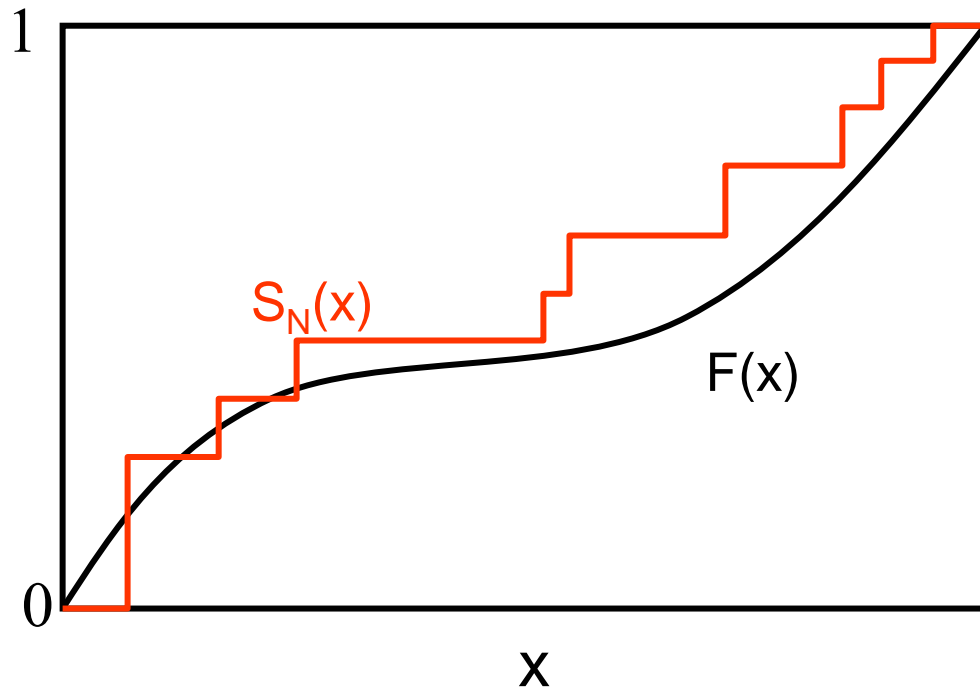
Test de Smirnov-Cramer-Von Mises

Suposem que tenim una mostra $\{x_1, \dots, x_N\}$ i volem saber si és compatible amb una certa distribució de densitat de probabilitat $f(x)$.

Definim:

$$S_N(x) \equiv \begin{cases} 0 & x < x_1 \\ i/N & x_i \leq x < x_{i+1} \\ 1 & x_N \leq x \end{cases}$$

$$F(x) = \int_{-\infty}^x f(x) dx$$



$$W^2(\bar{x}) = \int_{-\infty}^{+\infty} (S_N(x) - F(x))^2 f(x) dx = \int_{-\infty}^{x_1} F^2(x) dF(x) + \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} \left(\frac{i}{N} - F(x) \right)^2 dF(x) +$$

$$\int_{x_N}^{+\infty} (1 - F(x))^2 dF(x) = \frac{1}{N} \left\{ \frac{1}{12N} + \sum_{i=1}^N \left[F(x_i) - \frac{2i-1}{2N} \right]^2 \right\}$$

Per un valor fixat de x , $S_N(x)$ és una binomial amb

$$E[S_N(x)] = F(x) \quad (np)$$

$$E[(S_N(x) - F(x))^2] = \frac{1}{N} F(x)(1 - F(x)) \quad (npq)$$

i per tant

$$E(W^2) = E_{\bar{x}} \left[\int_{-\infty}^{+\infty} (S_N(x) - F(x))^2 f(x) dx \right] = \int_{-\infty}^{+\infty} E_{\bar{x}} [S_N(x) - F(x)]^2 f(x) dx =$$

$$= \int_{-\infty}^{+\infty} \frac{1}{N} F(x)(1 - F(x)) f(x) dx = \int_0^1 \frac{1}{N} F(1 - F) dF = \frac{1}{6N}$$

$$V(W^2) = E(W^4) - E(W^2)^2 = \frac{4N - 3}{180N^3}$$

Com ja coneixem $E(W^2)$ i $V(W^2)$ podem determinar la relació entre el nivell de significància α i el valor de W^2 de tall (de fet donarem $N \cdot W^2$)

Nivell α	Valor crític de NW^2
0.1	0.347
0.05	0.461
0.01	0.743
0.001	1.168

Donada α , si el NW^2 mesurat és major que el valor crític associat, hem de rebutjar la hipòtesi de que les dades venen descrites per $f(x)$ a aquest nivell de significància α escollit.

(a la pràctica és mesura NW^2 i es dona el nivell de confiança α associat)

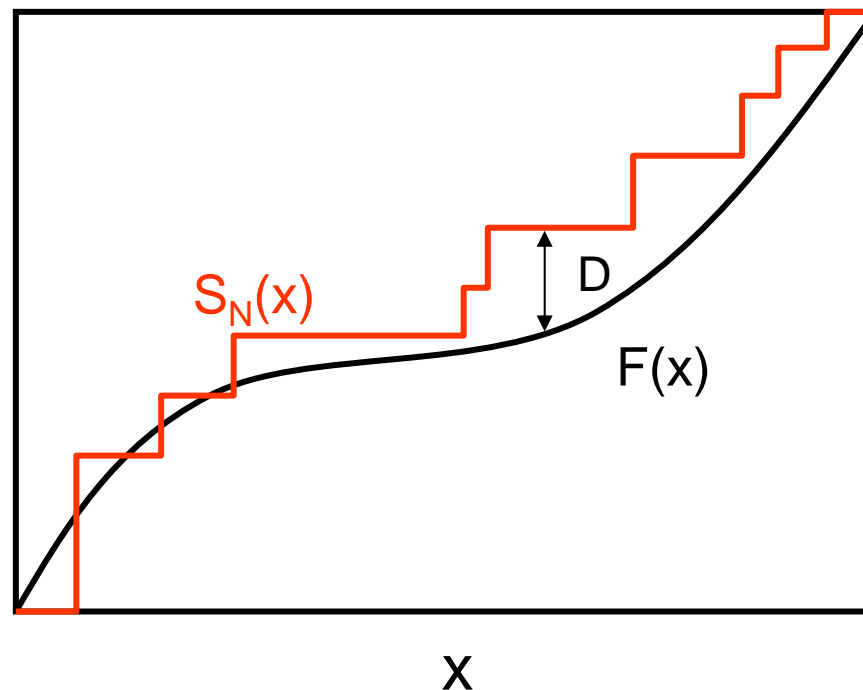
Test free of binning(II)

Test de Kolomogorov-Smirnov: cas 1

Suposem que tenim una mostra $\{x_1, \dots, x_N\}$ i volem saber si és compatible amb una certa distribució de densitat de probabilitat $f(x)$.

Definim l'estadístic D de Kolomogorov-Smirnof com

$$D = \max | S_N(x) - F(x) |$$



Quant més gran sigui D menys probable serà que la mostra sigui compatible amb la distribució $f(x)$. Podem usar aquest fet per avaluar la significància del valor D obtingut.

Si la mostra segueix realment la distribució $f(x)$ aleshores

$$P(x > D) = Q_{KS} \left(D \left[\sqrt{N} + 0.12 + \frac{0.11}{\sqrt{N}} \right] \right)$$

amb

$$Q_{KS}(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \lambda^2}$$

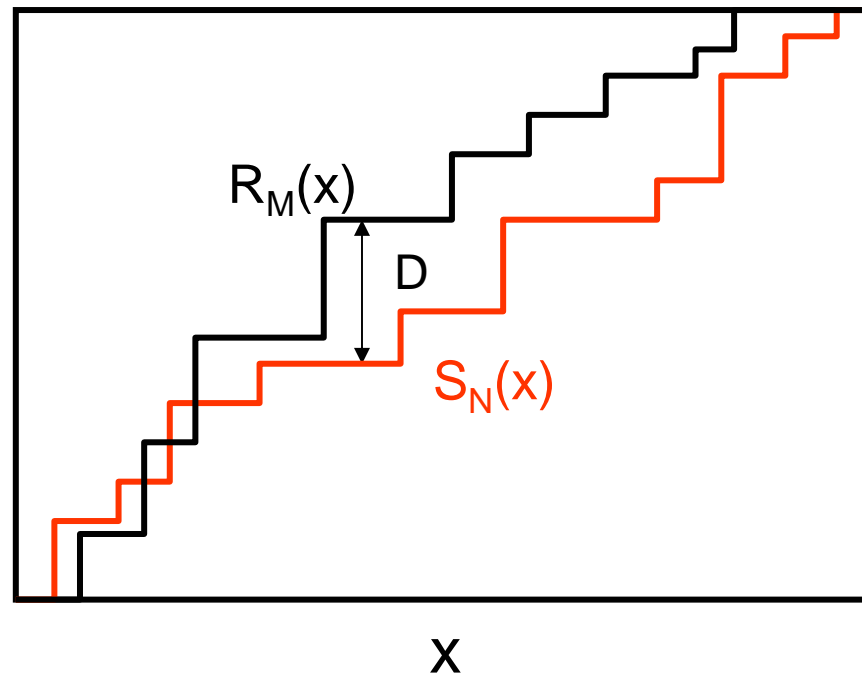
Quant més petita sigui $P(x \geq D)$ menys compatible serà la mostra amb $f(x)$

Test de Kolmogorov-Smirnov: cas 2

Suposem que tenim dues mostres $\{x_1, \dots, x_N\}$ i $\{y_1, \dots, y_M\}$ i volem saber si és possible que les dues provinguin de la mateixa distribució de densitat de probabilitat.

Definim l'estadístic D de Kolmogorov-Smirnov com

$$D = \max | S_N(x) - R_M(x) |$$



Quant més gran sigui D menys probable serà que les dues mostres provinguin de la mateixa distribució $f(x)$. Podem usar aquest fet per avaluar la significància del valor D obtingut.

Si les dues mostres provenen de la mateixa distribució

$$P(x > D) = Q_{KS} \left(D \left[\sqrt{N_e} + 0.12 + \frac{0.11}{\sqrt{N_e}} \right] \right)$$

amb

$$Q_{KS}(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \lambda^2} \quad N_e = \frac{NM}{N+M}$$

Quant més petita sigui $P(x \geq D)$ menys compatible serà que les dues mostres provinguin de la mateixa distribució de probabilitat.

Test de Wilks

En problemes reals podem trobar-nos en situacions en que no tenim cap coneixement a priori de la forma de la funció de densitat. En aquests casos normalment s'intenta modelar la mostra a partir de la combinació de funcions de densitat senzilles.

En aquests casos una pregunta habitual és *quantes funcions cal usar en la combinació?*

Exemple: si modelem una mostra com una combinació de distribucions normals $f(x) = a_1 N(\mu_1, \sigma_1) + a_2 N(\mu_2, \sigma_2) + \dots$, quantes cal usar per descriure correctament la mostra?

Augmentar el número de paràmetres de la distribució augmentarà la versemblança, però no suposarà necessàriament obtenir una descripció més realista

Suposem que definim dues funcions de densitat com a candidates per descriure una mostra:

- $f_1(x; \vec{\alpha}) \quad \vec{\alpha} \in \omega$
- $f_2(x; \vec{\beta}) \quad \vec{\beta} \in \Omega$

On ω i Ω són espais de paràmetres, de dimensions diferents.

Podem calcular la versemblança de la nostra mostra en els dos casos:

- $L(\omega)$ hipòtesi H_0 : la mostra està ben descrita per f_1
- $L(\Omega)$ hipòtesi H_1 : la mostra està ben descrita per f_2

Definim aleshores la Raó de versemblança i paràmetre ν

$$\lambda = \frac{L(\omega)}{L(\Omega)} \quad 0 < \lambda < 1 \quad \nu = -2 \ln(\lambda)$$

Wilks va demostrar que si la hipòtesi H_0 és certa aleshores ν segueix una **distribució χ^2 amb $t = \dim(\Omega) - \dim(\omega)$ graus de llibertat.**

Per tant, usant un test χ^2_t podem estimar la validesa (nivell de confiança) de la hipòtesi H_0 davant de la hipòtesi H_1 .

Exemple: tres hipòtesis

- Hipòtesi H_0 : $\ln(L_0)=300$ $\dim(\omega_0)=10$
- Hipòtesi H_1 : $\ln(L_1)=310$ $\dim(\omega_1)=20$
- Hipòtesi H_2 : $\ln(L_2)=314$ $\dim(\omega_2)=30$

Aleshores tindrem dues raons de versemblança:

- $v_{0-1} = 20$
- $v_{1-2} = 8$

I aleshores un test χ^2_{10} ens dona:

- H_0 respecte H_1 : C.L.=2.9%
- H_1 respecte H_2 : C.L.=62.9%

Hipòtesi escollida H_1

Teoria de la decisió

- Mesurem un estadístic x i volem decidir entre dos úniques possibilitats H_0 i H_1 (exemple: discriminar entre senyal i background)
- Agafarem aquella que doni la probabilitat més gran per la observació:
 - decidim H_0 si $p(H_0 | x) > p(H_1 | x)$ ↪ $= 1 - p(H_0)$
- **$p(x | H_0)p(H_0) > p(x | H_1)p(H_1)$**
- aquesta és l'aproximació Bayesiana. Pels anti-bayesians com desconeixem les probabilitats a priori $p(H_i)$ només podem parlar de test d'hipòtesi a un cert nivell de significació
- En física adoptarem un o l'altre punt de vista en funció del problema i si coneixem o no les probabilitats a priori.

- En altres problemes (exemple decisions militars) haurem d'associar també un grau de pèrdues “ l_i ” per cada una de les decisions H_i .

Decisió Realitat	$\theta=0$	$\theta=1$
H_0	0	l_0
H_1	l_1	0

Taula de “grau de pèrdua”

Escollirem H_0 si el grau de pèrdua a posteriori d'escollir H_0 quant H_1 és veritat

$$l_1 p(H_1 | x) = l_1 p(H_1) p(x | H_1) / p(x)$$

és menor que el grau de pèrdua a posteriori si escollim H_1 quant H_0 és veritat:

$$l_0 p(H_0 | x) = l_0 p(H_0) p(x | H_0) / p(x)$$

$$l_0 p(x | H_0) p(H_0) > l_1 p(x | H_1) p(H_1)$$

exemple(I): hem de pitjar "reset"?

Tenim un detector que cada dia pren dades contínuament amb una freqüència màxima θ_{\max} , però que pot anar perdent eficiència. Cada matí hem de decidir si hem de fer un "reset" del detector a partir de les "t" contes observades durant el darrer dia.

Decisió \ Realitat	"NO reset"	"reset"
$\theta = \theta_{\max}$	0	$p\theta_{\max}$
$\theta < \theta_{\max}$	$\theta_{\max} - \theta$	0

"reset" pren el seu temps

escollir H_0 si:
 $l_0 p(x | H_0) p(H_0) >$
 $l_1 p(x | H_1) p(H_1)$

Com l_1 depèn de θ , utilitzarem el seu valor esperat donats el "t" observats:

Escollir H_0 si: \rightarrow

$$p\theta_{\max} > E(\theta_{\max} - \theta) =$$

$$\int_0^{\theta_{\max}} (\theta_{\max} - \theta) P(\theta | t) d\theta$$

Exemple:

$$p(\theta | t) = p(t | \theta) \Pi(\theta)$$

$$\left\{ \begin{array}{l} p(t | \theta) = \frac{\theta^t}{t!} e^{-\theta} \\ \Pi(\theta) = \frac{1}{\theta_{\max}} \quad \text{si } \theta < \theta_{\max} \end{array} \right.$$

exemple(II): Seleccionant esdeveniments

Suposem que tenim dades $\{\vec{x}^{(i)} \in R^k ; i = 1, n\}$ amb dos tipus d'esdeveniments (**senyal/background**), que corresponen a dues diferents hipòtesis H_0 y H_1 , y que volem seleccionar els de tipus H_0 (senyal).

Cada esdeveniment és un punt $\vec{x}^{(i)} \in R^k$. ¿Quina és la regió que determinada pels esdeveniments considerats del tipus H_0 ?

Direm que és del tipus H_0 si:
$$\frac{p(\vec{x} | H_1)}{p(\vec{x} | H_0)} < \frac{p(H_0)}{1 - p(H_0)}$$

(agafem $l_0=l_1=1$)

És com trobar la regió d'acceptació de la hipòtesi H_0 en el test de Neyman-Pearson:

$$1 - \alpha = \int_A p(\vec{x} | H_0) d\vec{x} \quad \beta = \int_A p(\vec{x} | H_1) d\vec{x}$$

Un esdeveniment d'aquesta regió té probabilitat $(1-\alpha)*p(H_0)$ de ser “senyal” i $\beta*p(H_1)$ de ser “background” (α nivell de significació, $1-\beta$ potència)

El problema és que molt sovint desconeixem les expressions analítiques per les $p(x|H_i)$ i només tenim dades generades MC tan de senyal com de background.

En aquests casos la solució habitual per definir la regió d'acceptació és: **buscar els millors talls lineals (variant c_j)**

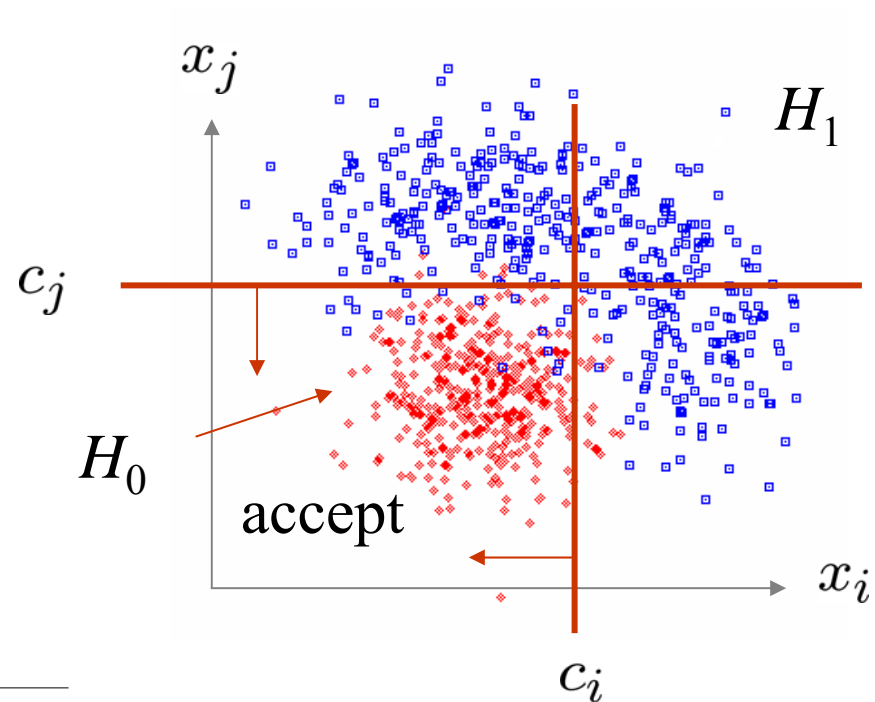
$$x_i < c_i$$

$$x_j < c_j$$

tractant de trobar regions on el background es petit i la senyal gran.

Per exemple maximitzant la puresa:

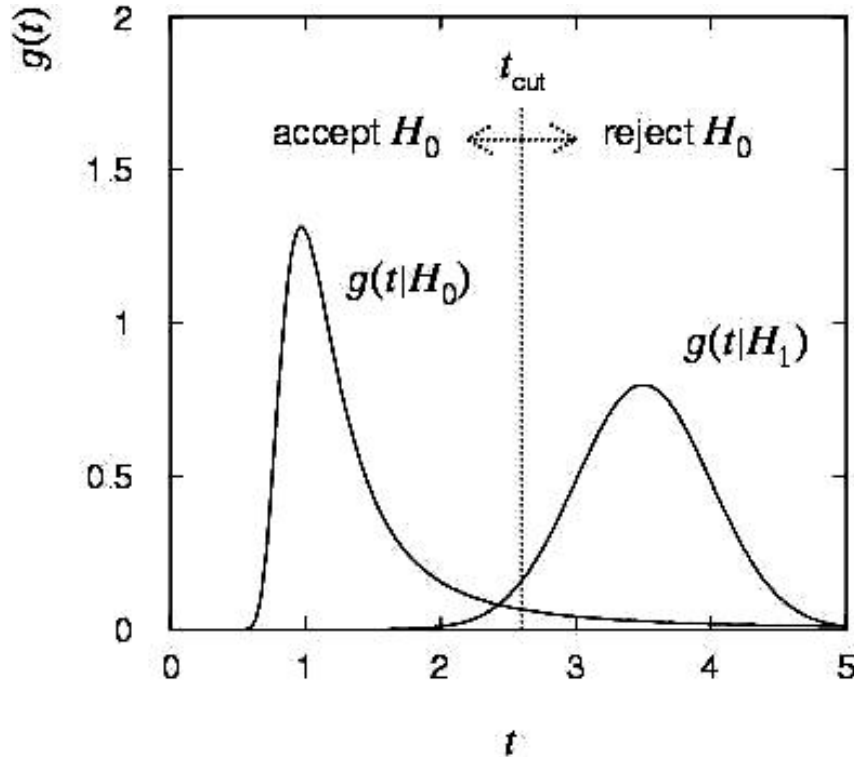
$$Pur(c_i) = \frac{sig}{sig + back} = \frac{(1 - \alpha)P(H_0)}{(1 - \alpha)P(H_0) + \beta P(H_1)}$$



Evidentment no és òptima!!!!

altres alternatives

Construir un estadístic $t(\vec{x})$ per comprimir les dades. En aquest cas la regió d'acceptació quedarà definida per un tall



Com ja hem vist l'òptim estadístic és:

$$t(\vec{x}) = \frac{p(\vec{x} | H_1)}{p(\vec{x} | H_0)}$$

amb el tall:

$$t_{cut} = \frac{p(H_0)}{1 - p(H_0)}$$

La transformació més fàcil a provar és una combinació lineal:

$$t(\vec{x}) = \sum_{j=1}^k a_j x_j$$

Com determinar les a_j ?

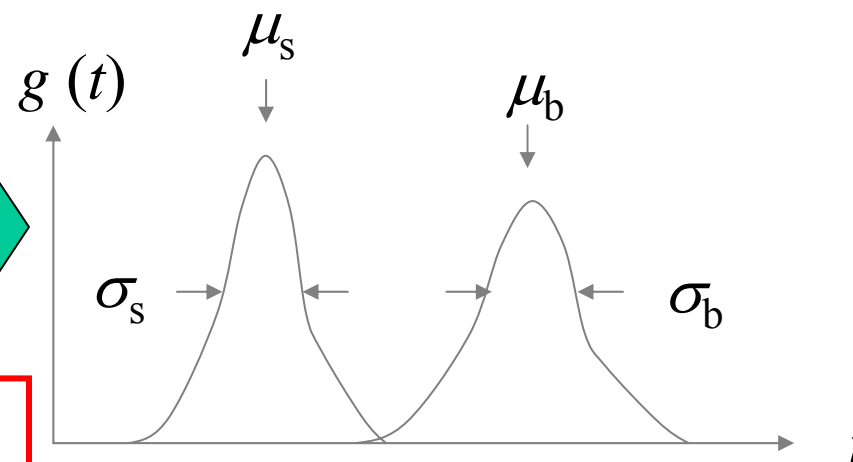
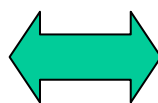
Discriminant de Fisher

$$t(\vec{x}) = \sum_{j=1}^k a_j x_j$$

Escollir les a_1, \dots, a_n de tal manera que les pdfs $g(t|s), g(t|b)$ tinguin la màxima 'separació'.

Volem:

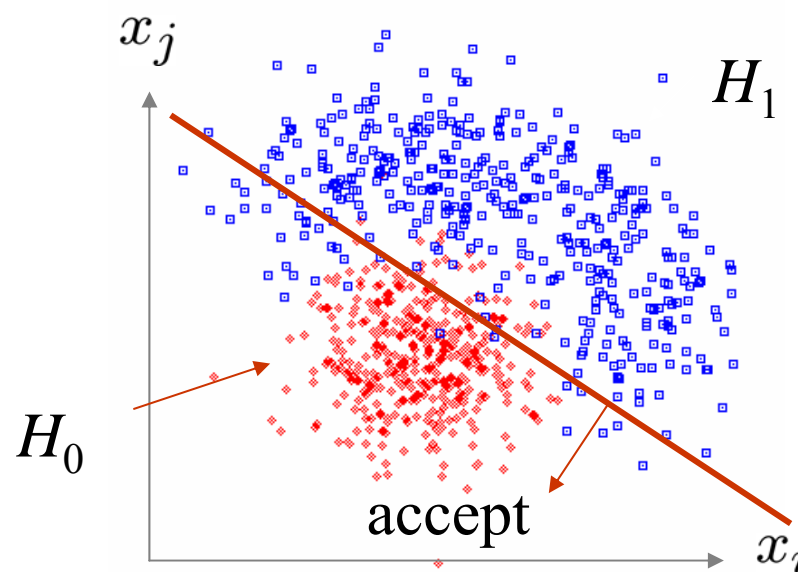
gran distancia entre les mitjanes i amplades petites



Maximitzar: $J(\vec{a}) = \frac{(\mu_s - \mu_b)^2}{\sigma_s^2 + \sigma_b^2}$

És Equivalent a Neyman-Pearson si la senyal y el background pdfs són Gaussianes nD amb iguals matrius V;

Si no es el cas NO és òptim

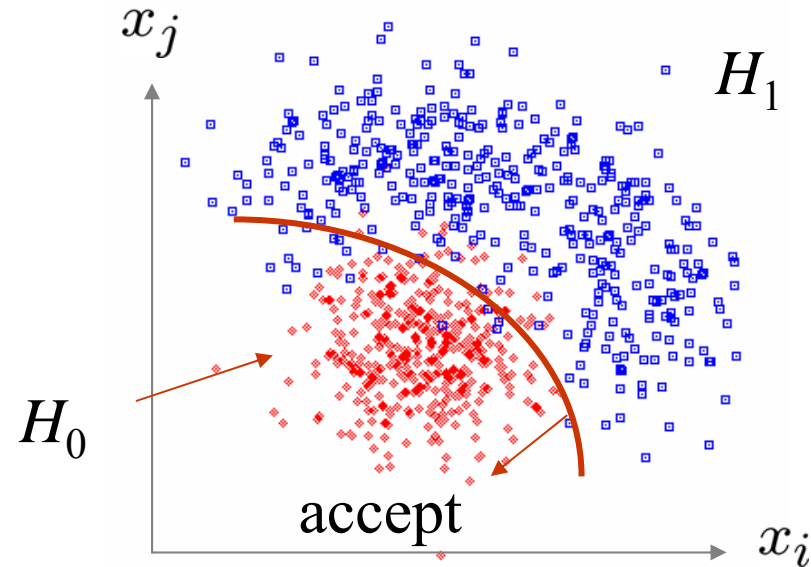


Estatístiques no lineals

La regió frontera no té per que ser un hyperpla , → estadístiques no lineals
(Neyman-Pearson és un d'ells)

Altres multivariate statistical methods:

Neural Networks,
Support Vector Machines,
Kernel density methods,
...



Exemple: imaginem que tenim dades MC tan de senyal(1) com de background(2). Minimitzem E variant els paràmetres w d'una família de funcions representada per $t(\vec{x} | \vec{w})$ (pot ser una xarxa neuronal)

$$E(\vec{w}) = \frac{1}{2} \sum_p \left(t(\vec{x}^{(p)}) - d^{(p)} \right)^2$$

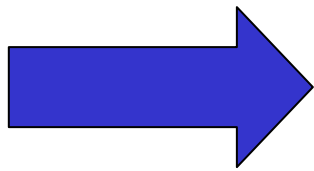
$$d^{(p)} = \begin{cases} 1 & \text{si } \vec{x}^{(p)} \in \text{clase 1} \\ 0 & \text{si } \vec{x}^{(p)} \in \text{clase 2} \end{cases}$$

- Aquesta funció error es pot expressar en funció de les p.d.f de cada una de les categories ($P_i(\vec{x}), i = 1, 2; \int P_i(\vec{x})d\vec{x} = 1$)

$$E[t(\vec{x})] = \int d\vec{x} (\alpha_1 P_1(\vec{x})(t(\vec{x}) - 1)^2 + \alpha_2 P_2(\vec{x})(t(\vec{x}))^2)$$

proporcions
 $\alpha_1 + \alpha_2 = 1$

$$\frac{\delta E[t(\vec{x})]}{\delta t(\vec{x}')} = 2 \int d\vec{x} (\alpha_1 P_1(\vec{x})(t(\vec{x}) - 1) + \alpha_2 P_2(\vec{x})t(\vec{x})) \delta(\vec{x} - \vec{x}') = 0$$



$$t(\vec{x}) = \frac{\alpha_1 P_1(\vec{x})}{\alpha_1 P_1(\vec{x}) + \alpha_2 P_2(\vec{x})}$$

- després de minimitzar, $t(\vec{x})$ ens dona la probabilitat per un \vec{x} donat de que sigui de la categoria 1.

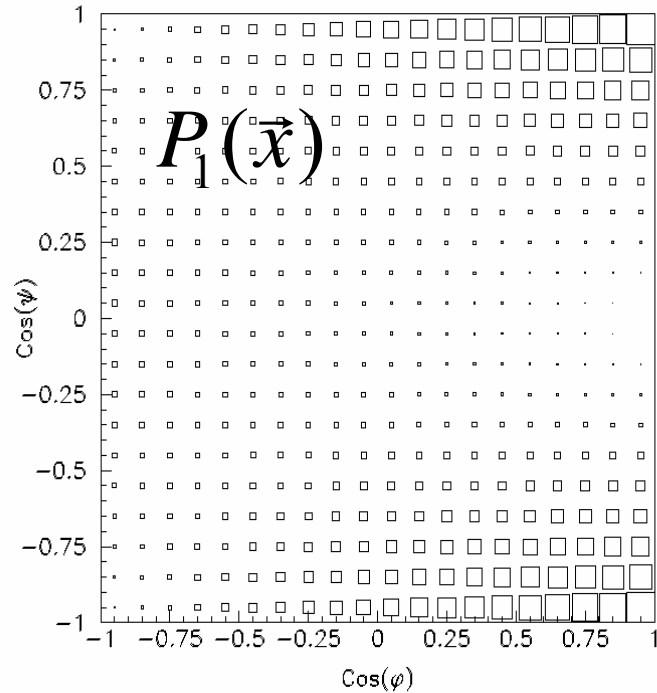
És equivalent al
 Neyman-Pearson!!!

$$t(\vec{x}) \approx P(1 | \vec{x})$$

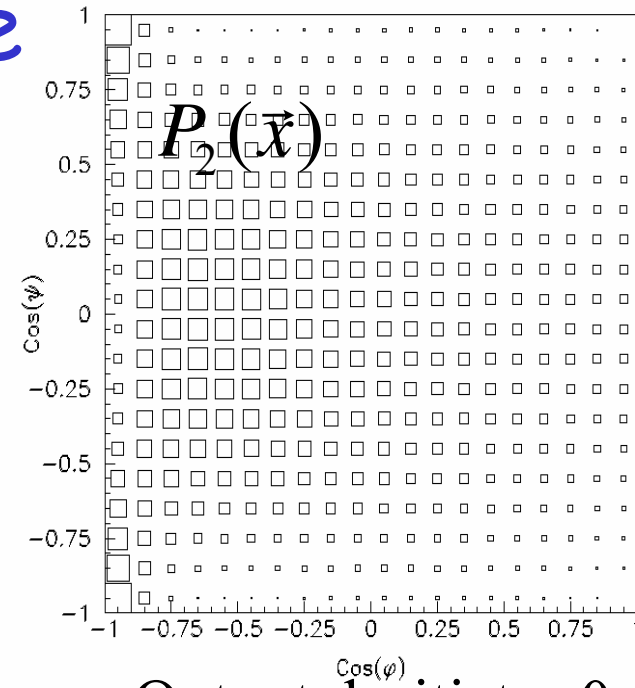
- Ho podem utilitzar per decidir a quina categoria pertany

$$si \quad t(\vec{x}) > 0.5 \quad \Rightarrow \quad \vec{x} \in \text{clase 1}$$

Exemple



Output desitjat = 1

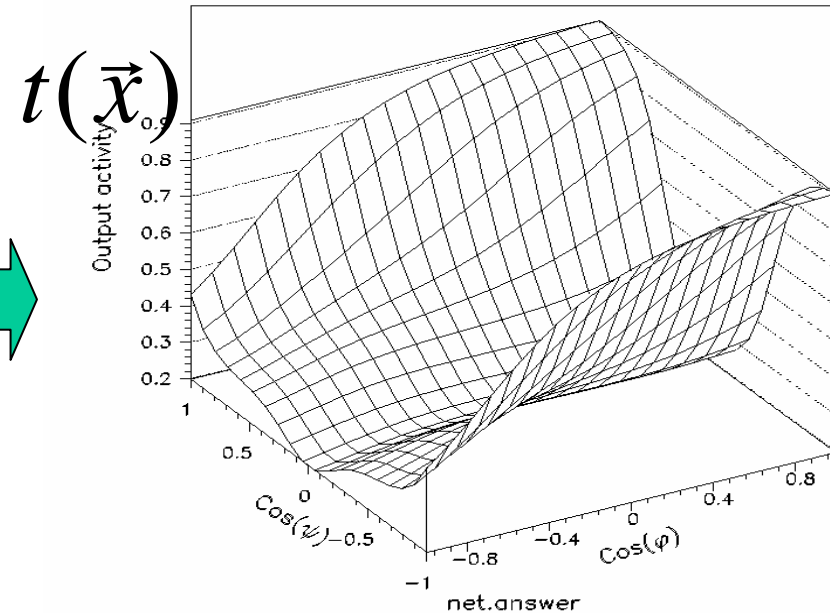
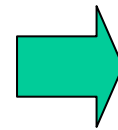


Output desitjat = 0

Minimitzem:

$$E(\vec{w}) = \frac{1}{2} \sum_p (t(\vec{x}^p) - d^p)^2$$

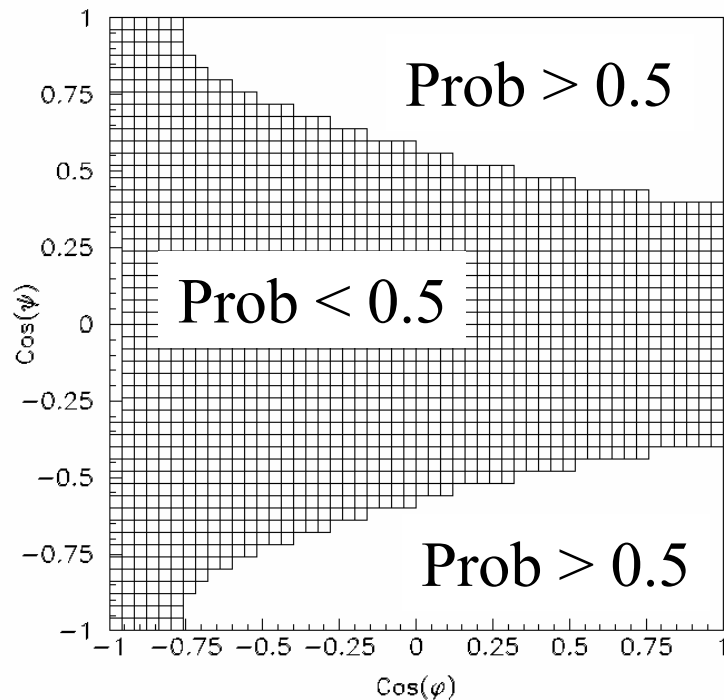
En aquest cas s'ha utilitzat una xarxa neuronal per $t(x)$.



Solució per talls no lineals

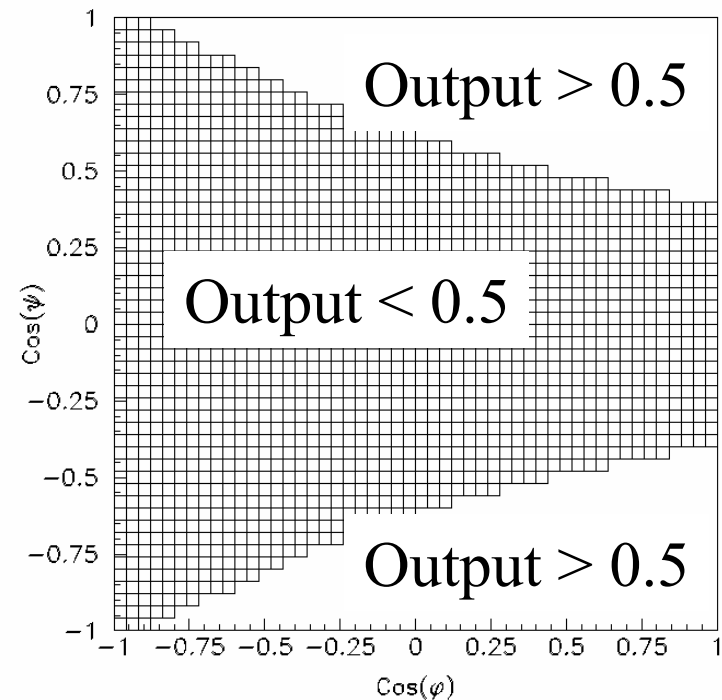
Solució òptima

encerts = 70.2%



Solució xarxa neural

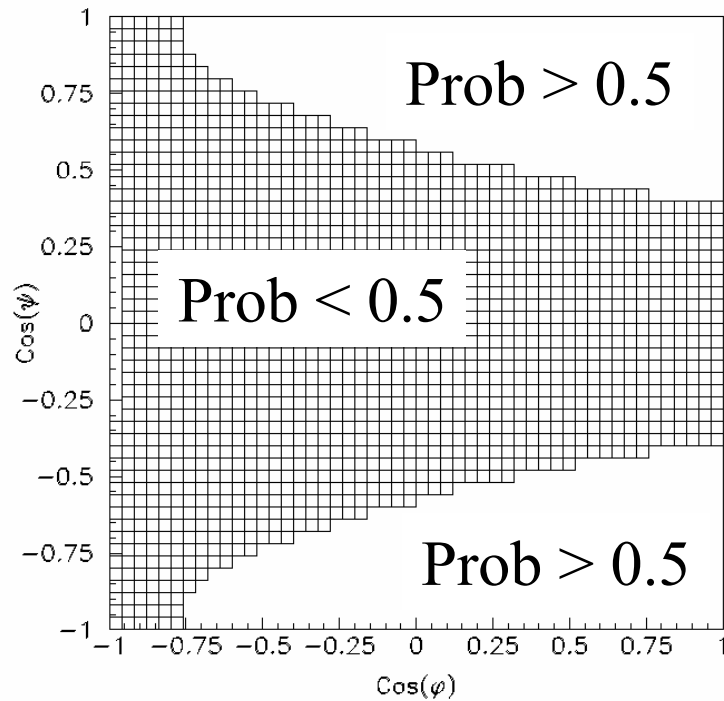
encerts = 70.0%



Solució per talls linials

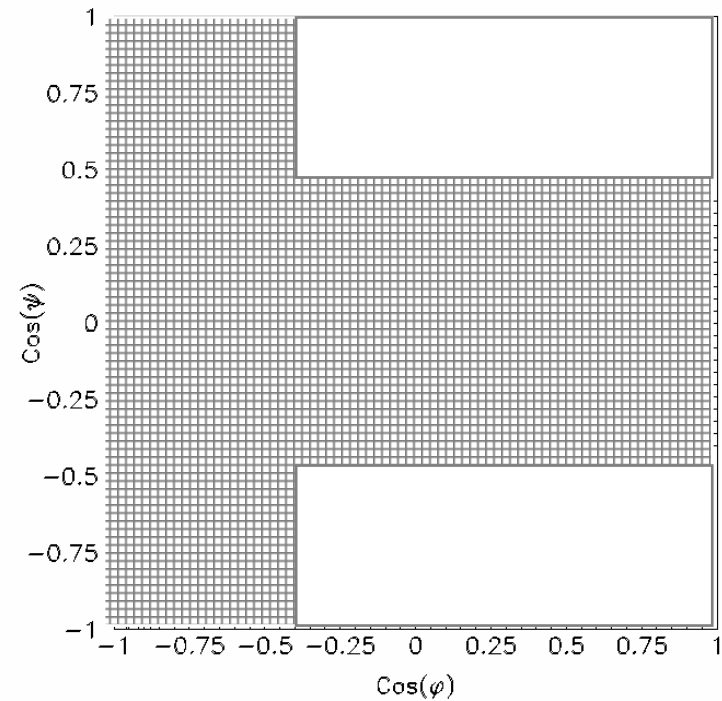
Solució òptima

encerts = 70.2%



Solució talls linials

encerts = 59%

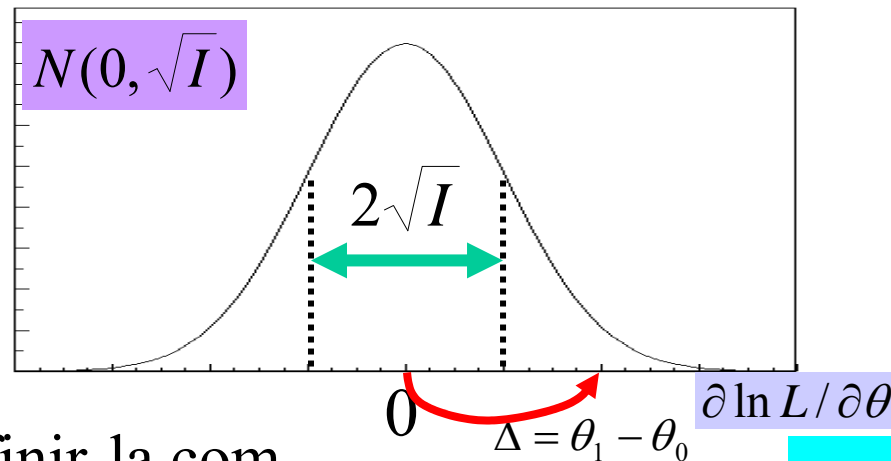


Informació distancia

- Sabem que la informació de Fisher està relacionada amb la resolució (màxima) que tenim per determinar el paràmetre θ de $f(x|\theta)$

$$I_F(\theta) = 1/\sigma_\theta^2$$

- suposem que tenim un “missatge” (experiment, observació,...) que ens permet canviar del coneixement actual del paràmetre $\theta_0 \rightarrow \theta_1$
- volem associar un nou concepte d’informació per aquest missatge. Aquesta informació ha d’estar lligada al “guany” de coneixement que hem tingut.



- És “natural” definir-la com

$$d(\theta_0, \theta_1) = \frac{\theta_1 - \theta_0}{\sigma_\theta} \rightarrow$$

Per qué no al quadrat?

$$I_d(\theta_0, \theta_1) \equiv \int_{\theta_0}^{\theta_1} \sqrt{I} d\theta$$

Informació de Kullback

- Tenim successos que poden caure en k llocs amb probabilitats p_i^0 .
Tenim un “missatge” que passa d’aquestes probabilitats $p_i^0 \rightarrow p_i^1$
- La informació de Kullback que porta aquest “missatge” és:

$$I_K \equiv \sum_{i=1}^k p_i^1 \ln(p_i^1 / p_i^0)$$

- és pot demostrar que $I_K \approx \frac{1}{2} (\sqrt{I_F} d\theta)^2$
- exemple:

tenim que les probabilitats p_i són funció d’un paràmetre θ . Per un succés:

$$I_F = E \left[\left(\frac{\partial \ln L(i | \theta)}{\partial \theta} \right)^2 \right] = E \left[\left(\frac{\partial \ln p_i(\theta)}{\partial \theta} \right)^2 \right] = E \left[\left(\frac{1}{p_i} \frac{\partial p_i}{\partial \theta} \right)^2 \right] = \sum_{i=1}^k \frac{1}{p_i} \left(\frac{\partial p_i}{\partial \theta} \right)^2$$

$$I_K \equiv \sum_{i=1}^k p_i(\theta_1) \ln \frac{p_i(\theta_1)}{p_i(\theta_0)} \approx \frac{1}{2} \sum_{i=1}^k \frac{1}{p_i} \left(\frac{\partial p_i}{\partial \theta} \right)^2 (\Delta\theta)^2 = \frac{1}{2} (\sqrt{I_F} \Delta\theta)^2$$

$$\Delta = \theta_1 - \theta_0, \sum_{i=1}^k p_i(\theta) = 1$$

Informació de Kullback(2)

- Com hem vist esta relacionada amb la informació que ens aporta un missatge que altera les nostres probabilitats de $p_i^0 \rightarrow p_i$
- de fet és l'única expressió (exceptuant constant multiplicativa) per aquesta informació que compleix les següents propietats:

- és una funció contínua de $p_i^0, p_i \Rightarrow I = I(\vec{p}; \vec{p}^0)$
- $I=0$ si $p_i^0 = p_i, \forall i$
- no depèn de l'ordre de les etiquetes

$$I(p_1, \dots, p_j, \dots, p_k, \dots, p_n; p_1^0, \dots, p_j^0, \dots, p_k^0, \dots, p_n^0) =$$

$$I(p_1, \dots, p_k, \dots, p_j, \dots, p_n; p_1^0, \dots, p_k^0, \dots, p_j^0, \dots, p_n^0)$$

- $I(1/n, \dots, 1/n, 0, \dots, 0; 1/n_0, \dots, 1/n_0)$ és una funció creixent de n_0 i decreixent de n (reducció de possibilitats, de n_0 passem a n)
- regla de composició:

$$I(p_1, \dots, p_r, p_{r+1}, \dots, p_n; p_1^0, \dots, p_r^0, p_{r+1}^0, \dots, p_n^0) = I(q_1, q_2; q_1^0, q_2^0) +$$

$$q_1 I(p_1 / q_1, \dots, p_r / q_1; p_1^0 / q_1^0, \dots, p_r^0 / q_1^0) + q_2 I(p_{r+1} / q_2, \dots, p_n / q_2; p_{r+1}^0 / q_2^0, \dots, p_n^0 / q_2^0),$$

(la informació es pot calcular per passos, pesats per la seva probabilitat)

$$(q_1 = p_1 + \dots + p_r \quad q_2 = p_{r+1} + \dots + p_n)$$

Informació de Shannon

- Si totes les p_i^0 són idèntiques tindrem:

$$I_K = \sum_{i=1}^k p_i \ln p_i - \ln p_i^0 \sum_{i=1}^k p_i = \sum_{i=1}^k p_i \ln p_i - \ln p_i^0$$

- per això es defineix la entropia o falta d'informació de Shannon com:

$$S(\vec{p}) = -\sum_{i=1}^k p_i \ln p_i$$

- de fet és l'única expressió que compleix les propietats:
 - és màxima quan totes les possibilitats són equiprobables
 - és mínima quan no hi ha incertesa $S(1,0,\dots,0)=0$
 - és simètrica amb els seus arguments
 - si les possibilitats són equiprobables, incrementa amb el número de possibilitats (augmenta l'entropia).
 - propietat additiva (com abans):

$$S(p_1, \dots, p_m, p_{m+1}, \dots, p_n) = S(q_a, q_b) + q_a S(p_1 / q_a, \dots, p_m / q_a) + q_b S(p_{m+1} / q_b, \dots, p_n / q_b)$$

Exemples de l'ús de la informació

- Tenim successos que poden caure en k llocs amb probabilitats p_i que volem estimar-les
- si experimentalment observem un total de n successos repartits de forma que trobem n_i en el lloc i , utilitzarem el likelihood per estimar les p_i

$$L(n_1, \dots, n_k; \vec{p}) = n! \prod_{i=1}^k \frac{p_i^{n_i}}{n_i!} \Rightarrow \ln L = \sum_{i=1}^k n_i \ln p_i + \text{constant} \quad (p_k = 1 - \sum_{i=1}^{k-1} p_i)$$

$$0 = \frac{\partial \ln L}{\partial p_i} = \frac{n_i}{p_i} - \frac{n_k}{p_k} \Rightarrow p_i = \frac{n_i}{n_k} p_k \quad 1 = \frac{n}{n_k} p_k \Rightarrow p_i = \frac{n_i}{n}$$

- però també podem donar l'error:

$$V(p_i) = \frac{p_i(1-p_i)}{n}$$

Exemple Shannon

- Però què farem si no coneixem les n_i i només coneixem $\bar{E} = \sum_{i=1}^k E_i p_i$?
- Maximitzarem la incertesa o minimitzarem la informació que ens passa de la distribució plana a les p_i

$$S = -\sum_{i=1}^k p_i \ln p_i + \lambda \left(\sum_{i=1}^k p_i - 1 \right) + \mu \left(\sum_{i=1}^k E_i p_i - \bar{E} \right)$$

$$\frac{\partial S}{\partial p_i} = -\ln p_i - 1 + \lambda + \mu E_i = 0 \Rightarrow p_i = e^{\lambda-1} e^{\mu E_i}$$

$$\frac{\partial S}{\partial \lambda} = \sum_{i=1}^k p_i - 1 = 0 \Rightarrow e^{\lambda-1} \sum_{i=1}^k e^{\mu E_i} = 1 \Rightarrow p_i = \frac{e^{\mu E_i}}{\sum_{i=1}^k e^{\mu E_i}}$$

$$\frac{\partial S}{\partial \mu} = \sum_{i=1}^k E_i p_i - \bar{E} = 0 \Rightarrow \frac{\sum_{i=1}^k E_i e^{\mu E_i}}{\sum_{i=1}^k e^{\mu E_i}} = \bar{E} \Rightarrow \mu = \mu(\bar{E})$$

$$p_i = \frac{e^{\mu(\bar{E}) E_i}}{\sum_{i=1}^k e^{\mu(\bar{E}) E_i}}$$

Exemple Kullback

- El mateix problema que abans però ara tenim un coneixement a priori de les p_i^0
- minimitzarem la informació que ens passa de les $p_i^0 \rightarrow p_i$

$$I = \sum_{i=1}^k p_i \ln p_i / p_i^0 - \lambda (\sum_{i=1}^k p_i - 1) - \mu (\sum_{i=1}^k E_i p_i - \bar{E})$$

$$\frac{\partial I}{\partial p_i} = + \ln p_i + 1 - \lambda - \mu E_i - \ln p_i^0 = 0$$



$$p_i = \frac{p_i^0 e^{\mu(\bar{E})E_i}}{\sum_{i=1}^k p_i^0 e^{\mu(\bar{E})E_i}}$$

Exemple likelihood

- Si en el cas anterior coneixem les n_i i les probabilitats a priori amb el seu error, utilitzarem el likelihood per estimar les noves probabilitats amb el seu nou error

$$L(n_1, \dots, n_k; \vec{p}) = n! \prod_{i=1}^k \frac{p_i^{n_i}}{n_i!} P(\vec{p}^0)$$



Normal en k dimensions

$$\ln L(\vec{p}) = \sum_{i=1}^k n_i \ln p_i - \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k (p_i - p_i^0) V_{ij}^{-1} (p_j - p_j^0) + \text{constant}$$

- de $\partial \ln L(\vec{p}) / \partial p_i = 0$ obtindrem les noves probabilitats i els nous errors