REMINDER ON PROBABILITIES

- When the result of an experiment (or trial) cannot be predicted with certainty, the result (or outcome, measurement) is described by a random variable X taking its values in the "sample space" Ω
  Coin tossing $\Omega = \{ \text{heads}, \text{tails} \}$
  throwing a die $\Omega = \{1, 2, 3, 4, 5, 6\}$
  Time of decay of a particle $\Omega = \{ x \in [0, \infty[\}$
- Any subspace of Ω is called an event.
  $A \subset \Omega$ the event A is true if the outcome $x \in A$.
- X is given a probability law by defining for all events their probability $P(A) \in [0, 1]$ of being true, satisfying
  $$P(\Omega) = 1 \quad P(\emptyset) = 0$$
  and $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
- Definition A and B are exclusive events if $A \cap B = \emptyset$
  For exclusive events, $P(A \cup B) = P(A) + P(B)$
- Definition. The conditional probability $P(A|B)$ of A being true, knowing that B is true, is
  $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes theorem $P(A|B)\,P(B) = P(B|A)\,P(A)$

- Definition: A is independant of B if $P(A|B) = P(A)$
  it implies that B is independant of B
  and $(P(A \cap B) = P(A) \cdot P(B)$

PROBABILITY LAWS

1. X is discrete : $\Omega = \{x_i\}$

   Probability law for X totally specified by $p_i = P(x_i)$

   with $p_i \in [0,1]$ $\sum_i p_i = 1$ (sum is finite or infinite)

2. X is continuous

   Probability law specified by $F(x_0) = P(x \leq x_0)$

   $F(x)$ is the cumulative fonction, with $F(-\infty) = 0$, $F(+\infty) = 1$

   F monotonous $\nearrow$

   The probability density fonction is $f(x)$ such as

   $f(x)dx = P(X \in [x, x+dx]) = F(x+dx) - F(x)$

   $$\Rightarrow f(x) = \frac{dF(x)}{dx}$$

3. X continuous, bidimensional $\{X, Y\}$

   The pdf is $f(x,y)$, positive fonction with $\int f(x,y)dxdy = 1$

   The marginal p.d.f. for X is $f_X(x) = \int f(x,y)dy$

   The conditional p.d.f for X, knowing that $y = y_0$ is

   $$f_c(x|y_0) = \frac{f(x,y_0)}{\int f(x,y_0)dx}$$

   If X and Y are independant, $f_c(x|y_0) = f_X(x)\ \forall y_0$

   $$\Rightarrow f(x,y) = f_X(x) f_Y(y)$$

4. Change of variables $X \to Y$ (or $\vec{X} \to \vec{Y}$)

$y = H(x)$ or $\vec{y} = \vec{H}(\vec{x})$ analytical relation.

4a x, y of dimension 1 – p.d.f of x is $f(x)$, of y is $g(y)$

If H is bijective, $g(y)\,dy = f(x)\,dx$

$$\Rightarrow g(y) = \frac{f(x)}{\left|\frac{dy}{dx}\right|} = \frac{f(x)}{|H'(x)|}$$

If H is non bijective, one sums on all parent values of x for a given value of x.

Exemples: $g(|x|) = f(x) + f(-x)$

$$y = x^2 \qquad g(y) = \frac{f(\sqrt{y})}{2\sqrt{y}} + \frac{f(-\sqrt{y})}{2\sqrt{y}}$$

4b $\vec{x}, \vec{y}$ of dimension k – Bijective relation.

$$g(\vec{y}) = \frac{f(\vec{x})}{\left|\text{dét}\left(\frac{\partial \vec{H}}{\partial \vec{x}}\right)\right|} \longleftarrow \text{Jacobian}$$

5. Some moments of the probability law

5a) Mean value (or expectation) of x

$$\langle x \rangle = \bar{x} = E(x) = \begin{cases} \sum p_i x_i & \text{(discrete case)} \\ \int x\, f(x)\, dx & \text{(continuous case)} \end{cases}$$

If $y = a(x)$

$$\langle y \rangle = \begin{cases} \sum p_i\, y_i \\ \int y\, g(y)\, dy = \int a(x)\, f(x)\, dx. \end{cases}$$

5b) Variance of x :

$$V_x = \sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2 = \langle (x-\bar{x})^2 \rangle$$

$\sigma = \sqrt{\sigma^2}$ is called standard deviation (or uncertainty, error ...)

5c) Multidimensional case : Covariance.

$$f(x,y) \quad : \quad C_{xy} = \langle (x-\bar{x})(y-\bar{y}) \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

If x and y are independant, $C_{xy} = 0$.

THE RECIPROCAL IS WRONG.

Correlation coefficient $r_{xy} = \dfrac{C_{xy}}{\sigma_x \sigma_y} \in [-1, +1]$

$\vec{x} = x_1, x_2, \ldots, x_k$: Variance-covariance matrix

$$V_{\vec{x}} = \begin{pmatrix} \sigma_{x_1}^2 & C_{x_1 x_2} & \cdots & C_{x_1 x_k} \\ C_{x_1 x_2} & \sigma_{x_2}^2 & & \vdots \\ \vdots & & \ddots & \\ C_{x_1 x_k} & \cdots & & \sigma_{x_k}^2 \end{pmatrix}$$

V is symetric, semi-definite positive

(eigenvalues are ≥ 0)

6 - Variance on new variables:

6a) linear change:

$y_i = \sum m_{ij} x_j$ & $\vec{Y} = M\vec{X}$

not necessarily the same dimension

$$V_{\vec{Y}} = M \, V_{\vec{X}} \, M^T$$

matrix of m_{ij}

Applications • $y = kx \rightarrow \sigma_y^2 = k^2 \sigma_x^2$.

• $f(x,y)$ with $V_{xy} = \begin{pmatrix} \sigma_x^2 & C_{xy} \\ C_{xy} & \sigma_y^2 \end{pmatrix}$

$s = x + y$

$$V_s = \sigma_s^2 = \sigma_x^2 + \sigma_y^2 + 2C_{xy}$$

If x and y are uncorrelated ($C_{xy}=0$) $\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2$

• N outcomes of X for N independant trials pdf: $f(x)$ for each trial; with mean $\bar{x}$ and variance σ^2

$$s_N = \sum x_i \qquad \sigma_{s_N}^2 = N\sigma^2$$

$$z_N = \frac{\sum x_i}{N} \qquad \sigma_{z_N}^2 = \frac{\sigma^2}{N}$$

6b) Non linear variable change $\vec{Y} = \vec{H}(\vec{X})$

No exact formula for $V_{\vec{Y}}$

However, $V_{\vec{Y}} \simeq D \, V_{\vec{X}} \, D^T$

with D the matrix of derivatives computed at $\vec{x} = \langle \vec{x} \rangle$ is often used YOU NEED TO BE CAREFUL!

SOME USUAL LAWS

1) Exponential law (particle decay)

$$f(t) = \frac{1}{\tau} e^{-t/\tau} \qquad \langle t \rangle = \tau \qquad \sigma_t^2 = \tau^2$$

2) Gaussian (or normal) law

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} = G_{\mu,\sigma^2}(x) \qquad \langle x \rangle = \mu \qquad \sigma_x^2 = \sigma^2$$

Multidimensional normal : $\vec{x}$ of dimension k

$$f(\vec{x}) = \frac{1}{(2\pi)^{k/2} |\det V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu}) V^{-1} (\vec{x}-\vec{\mu})^T\right]$$

$$\langle \vec{x} \rangle = \vec{\mu} \qquad V_{\vec{x}} = V$$

Each component x_i is normally distributed

3) Binomial law

X taking values in Ω. Consider $A \subset \Omega$ with $\mathbb{P}(A) = p$.

N independant trials $\rightarrow$ N outcomes : n successes ($x \in A$)
$N-n$ failures ($x \notin A$)

n follows a binomial law :

$$\mathcal{B}_{N,p}(n) = C_N^n \, p^n (1-p)^n \qquad C_N^n = \frac{N!}{n!(N-n)!}$$

$$\langle n \rangle = Np \qquad \sigma_n^2 = Np(1-p)$$

<u>Application</u> N events filling an histogram

each bin corresponds to a probability p_i $(\Sigma p_i = 1)$

and contains n_i events (n_i are random variables with $\Sigma n_i = N$)

$$\left.\begin{array}{l} n_1 \rightsquigarrow B_{N,p_1}(n_1) \\ n_2 \rightsquigarrow B_{N,p_2}(n_2) \\ n = n_1 + n_2 \rightsquigarrow B_{N,p_1+p_2}(n) \end{array}\right\}$$

Exercise: deduce

$$C_{n_1 n_2} = -N p_1 p_2$$

multinomial law $P(n_1, \cdots, n_k) = \dfrac{N!}{n_1! \ldots n_k!} p_1^{n_1} \cdots p_k^{n_k}$

4) Poisson law

$\mathcal{P}_a(n)$ is the limit of $B_{N,p}(n)$ when $\begin{array}{l} N \to \infty \\ p \to 0 \end{array}$ $Np \to a$.

$$P(n) = \mathcal{P}_a(n) = e^{-a} \frac{a^n}{n!} \quad \text{(Exercise: prove it)}$$

$$\langle n \rangle = a \quad \sigma_n^2 = a.$$

<u>Application</u> N events filling an histogram, N poisson distributed

The populations of each bin, n_i, are independant random variables ~~which~~ are Poisson distributed

ASYMPTOTIC THEOREMS

1) Binomial law $B_{N,p}(n) = C_N^n \, p^n (1-p)^{N-n}$

$f_N = \frac{n}{N}$ frequency of success

$\langle f_N \rangle = p \quad \sigma_f^2 = \frac{p(1-p)}{N} \longrightarrow 0$ when $N \to \infty$

Constructive way of defining probability law on Ω

2) X with p.d.f $f(x)$ N independant trials giving $x_1 \cdots x_N$

$\langle x \rangle = \bar{x} \quad \sigma_x^2 = \sigma^2$

$s_N = \sum_1^N x_i$ has a mean $N\bar{x}$ and a variance $\sigma_x^2 N$

$\frac{s_N}{N}$ has a mean $\bar{x}$ and a variance $\frac{\sigma^2}{N} \longrightarrow 0$ when $N \to \infty$

p.d.f for s_2 : $g(s_2) = \int f(s_2 - x) f(x)\, dx = f * f$

p.d.f for $s_N = f^{*N}$

Central limit theorem:

When $N \to \infty$, the p.d.f for $s_N \to G_{N\bar{x}, N\sigma^2}$

the p.d.f for $\frac{s_N}{N} \to G_{\bar{x}, \frac{\sigma^2}{N}}$

$$t_N = \frac{1}{\sigma \sqrt{N}} \sum_1^N (x_i - \bar{x}) \quad \text{p.d.f for } t_N \to G_{0,1}$$

Note: This theorem will give the asymptotic properties of the estimates of unknown parameters.

PARAMETER ESTIMATION

X is a random variable with p.d.f $f(x;\theta_0)$, where θ_0 is the true value of an unknown parameter θ.

N independant trials x_i are summarized into a statistic $t_N = h(x_1 \cdots x_N)$ (defined for any N)

t_N will be an unbiased estimator of θ_0 if $\langle t_N \rangle = \theta_0$

If $\langle t_N \rangle = \theta_0 + b_N$, t_N is biased

t_N will be a convergent estimator of θ_0 if, as $N \to \infty$, $b_N \to 0$ like $\frac{1}{N}$ and $\sigma^2_{t_N} \to 0$ like $\frac{1}{N}$

t_N will be an optimal estimator if it is unbiased and of minimal variance.

t_N will be an efficient estimator if it is unbiased and if its variance reaches the theoretical lower bound (given by the theory of information)

<u>Note</u> Efficient estimators do NOT necessarily exist.

USUAL ESTIMATORS

A- Moment method

X with p.d.f $f(x;\theta_0)$

build $Y = a(X)$ $\begin{cases} E(Y) = \int a(x)\, f(x;\theta_0)\, dx = g(\theta_0) \\ \sigma^2(Y) = \int [a(x) - g(\theta_0)]^2 f(x;\theta_0)\, dx \end{cases}$

N trials $x_i \Rightarrow$ build $z_N = \frac{1}{N}\sum_1^N a(x_i)$

z_N is an unbiased convergent estimator of $g(\theta_0)$

z_N is in general not optimal

$\langle z_N \rangle = g(\theta_0) \quad \sigma^2_{z_N} = \frac{\sigma_Y^2}{N}$

From central limit theorem, z_N is asymptotically normally distributed.

Note: $g^{-1}(z_N)$ is a biased and convergent estimator of θ_0

The asymptotic variance of $\hat{\theta} = g^{-1}(z_N)$ is $\frac{1}{N}\left|\frac{\partial g}{\partial \theta}(\hat{\theta})\right|^{-2} V_Y$

B- Maximum of likelihood

X of p.d.f $f(x;\theta_0)$ N independant trials

$\hat{\theta}_N$ is the value of θ maximizing

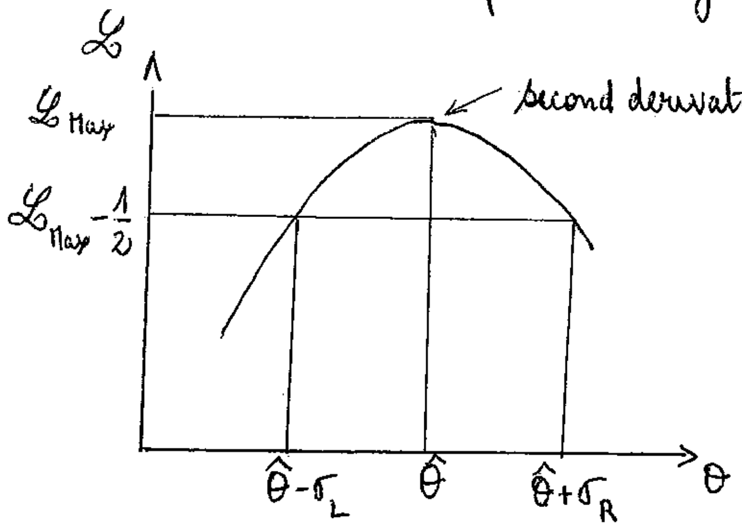
$$\ln \mathcal{L} = \sum_1^N \ln f(x_i;\theta)$$

$\hat{\theta}_N$ is an asymptotically efficient estimator of θ_0

N finite: $\hat{\theta}_N$ is a biased, non optimal, convergent estimator of θ_0

$$V(\hat{\theta}) \xrightarrow{N\to\infty} \frac{1}{N}\left[-E\left(\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}\Big|_{\theta=\theta_0}\right)\right]^{-1}$$

estimated in practice by $\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial\theta^2}\Big|_{\theta=\hat{\theta}}\right]^{-1} = -D_2^{-1}$



$N\to\infty \quad \hat{\theta}\to\theta_0$

σ_L^2, σ_R^2 and $D_2^{-1} \to V_{\hat{\theta}} \to 0$

$\boxed{\mathcal{L}_{max}-\frac{1}{2} \text{ rule}}$

Special case: Darmois theorem

If $f(x;\theta) = \exp\{\alpha(x)\,a(\theta) + \beta(x) + c(\theta)\}$

and if Ω_x independant of θ

Then $\frac{1}{N}\sum_1^N \alpha(x_i)$ is an efficient estimator

of $\tau(\theta_0) = -\dfrac{\frac{dc}{d\theta}(\theta_0)}{\frac{da}{d\theta}(\theta_0)}$

C – Least-squares estimator

k measurements x_i with mean $\mu_i(\theta_0)$ and variance $\sigma_i^2(\theta_0)$

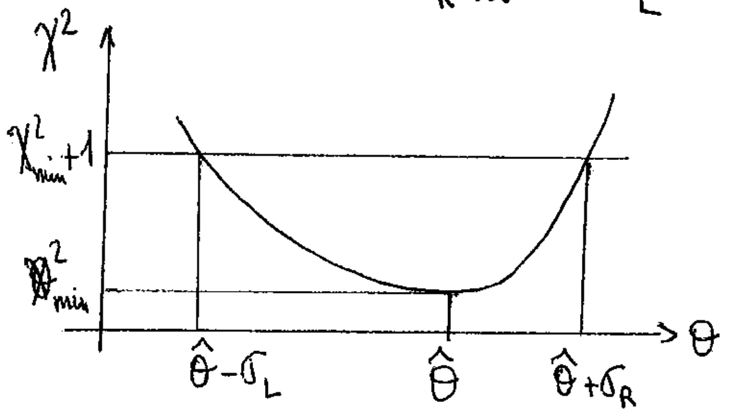
θ_0 true value of unknown parameter θ.

$\hat{\theta}$ is the value of θ which minimizes

$$\chi^2 = \sum_1^k \frac{(x_i - \mu_i(\theta))^2}{\sigma_i^2(\theta)}$$

$\hat{\theta}$ is a biased, convergent, non optimal estimator of θ_0

$$V_{\hat{\theta}} \xrightarrow[k\to\infty]{} 2\left[\left\langle \frac{\partial^2 \chi^2}{\partial \theta^2} \right\rangle_{\theta=\theta_0}\right]^{-1} \sim 2\, D_2^{-1} \text{ at } \hat{\theta}$$



$\boxed{\chi^2_{min}+1 \text{ rule}}$

$k\to\infty \quad \sigma_R^2, \sigma_L^2, 2D_2^{-1} \to V_{\hat{\theta}} \to 0$

Special case : Linear model

When 1) $\mu_i(\theta)$ are linear fonctions of θ

2) $V_{\vec{x}}$ is independant of θ.

The minimization of $\chi^2 = (\vec{x} - \vec{\mu}(\theta))\, V_{\vec{x}}^{-1} (\vec{x} - \vec{\mu}(\theta))^T$

gives an optimal, convergent estimator of θ_0

$$V_{\hat{\theta}} \equiv 2\left[\frac{\partial^2 \chi^2}{\partial \theta^2}\right]^{-1}$$

χ^2 law

When observations x_i are normally distributed with mean μ_i and variance σ_i^2

$$Q_0 = \chi^2(\theta_0) = \sum_{i=1}^{k} \frac{[x_i - \mu_i(\theta_0)]^2}{\sigma_i^2(\theta)}$$ is distributed

according to a χ^2 law at k degrees of freedom

$$f(\chi_k^2) = \frac{(\chi_k^2)^{\frac{k}{2}-1}}{2\,\Gamma(\frac{k}{2})} e^{-\frac{\chi_k^2}{2}} \qquad \langle \chi_k^2 \rangle = k$$

$$V_{\chi_k^2} = 2k$$

<u>When the model is linear</u>, $Q_{min} = \chi^2(\hat{\theta})$

follows a χ^2_{k-r} law and $Q_0 - Q_{min}$ a χ^2_r law

(where r is the dimension of θ)

and $\hat{\theta}$ has a k-dimensional normal law (see page 6)

with mean θ_0 and variance $2D_2^{-1}$ $\left(D_2 = \frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right)$

Histogram fitting

NOT a linear χ^2

(1) $\chi^2 = \sum_1^k \frac{(n_i - \bar{n}_i(\theta))^2}{\bar{n}_i(\theta)}$ non linear

(2) $\chi^2 = \sum_1^k \frac{(n_i - \bar{n}_i(\theta))^2}{n_i}$ linear but approximate

(3) $\ln \mathcal{L} = C + \sum_{i=1}^{k} n_i \ln \bar{n}_i(\theta)$ likelihood of multinomial law

3 different estimators of θ_0.

Empirical rule

(3) should be preferred

ESTIMATION BELTS

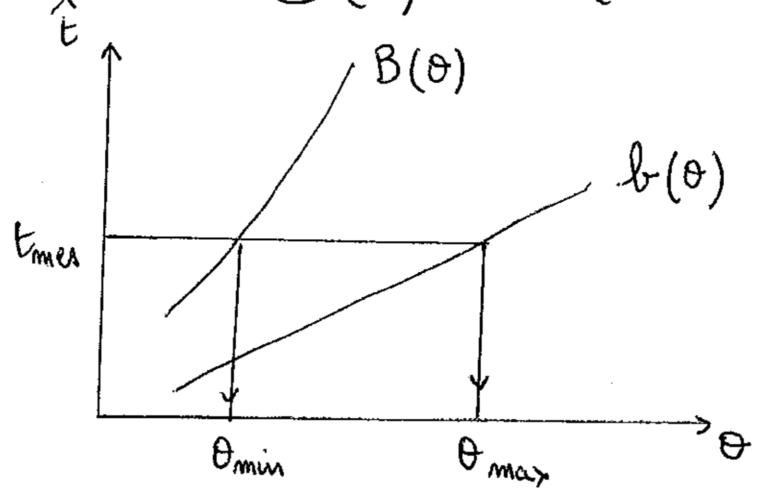
$\hat{t}$ being an estimator of θ_0 of known p.d.f $f(\hat{t};\theta)$

dimension 1

one defines 2 curves:

$$b(\theta) \rightarrow P(\hat{t} < b(\theta);\theta) = \alpha$$

$$B(\theta) \rightarrow P(\hat{t} > B(\theta);\theta) = \alpha$$



From the value t_{mes} found for $\hat{t}$, one derives θ_{min} and θ_{max} (Neyman construction)

$$P(\theta_0 < \theta_{max}) = 1-\alpha$$

$$P(\theta_0 > \theta_{min}) = 1-\alpha$$

$$P(\theta_{min} < \theta_0 < \theta_{max}) = 1-2\alpha$$

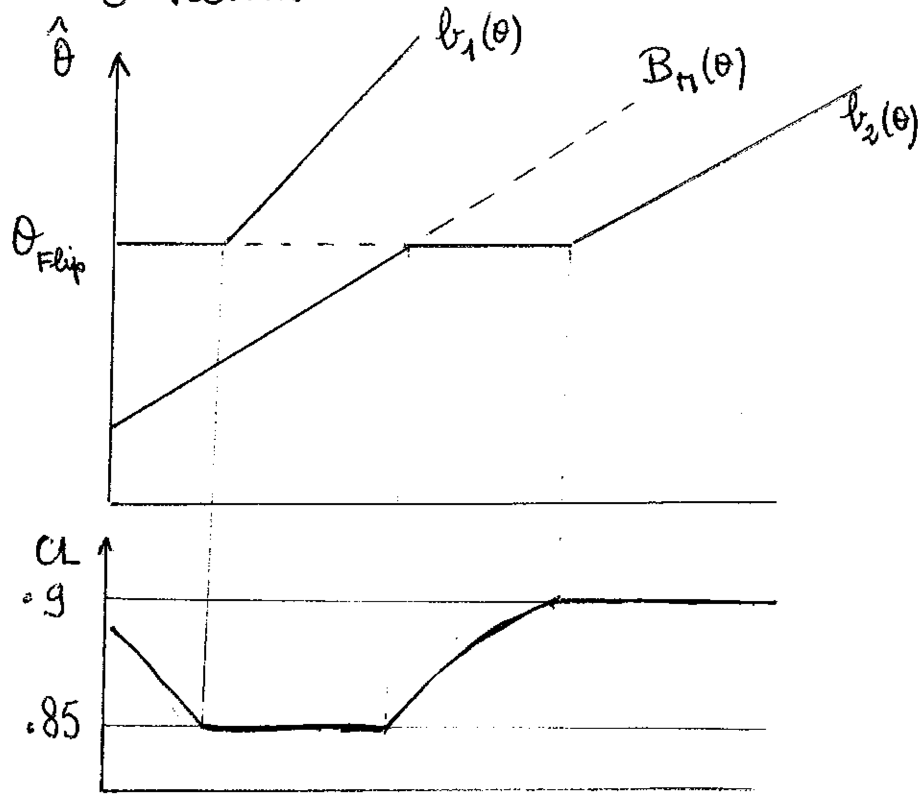
These probabilities refer to the random variables θ_{min} and θ_{max}, and are called confidence levels

Note Bayesian statisticians reinterpret these probabilities as referring to the unknown parameter θ_0

Problems with intervals and upper bounds

A. Flip-flop problem

If I decide to publish a 90% CL interval if $\hat{\theta} > \theta_{Flip}$ and a 90% CL upper bound when $\hat{\theta} < \theta_{Flip}$, the true confidence level is smaller (i.e. 85%) for some θ values



$P(\hat{\theta} > b_1(\theta)) = 5\%$

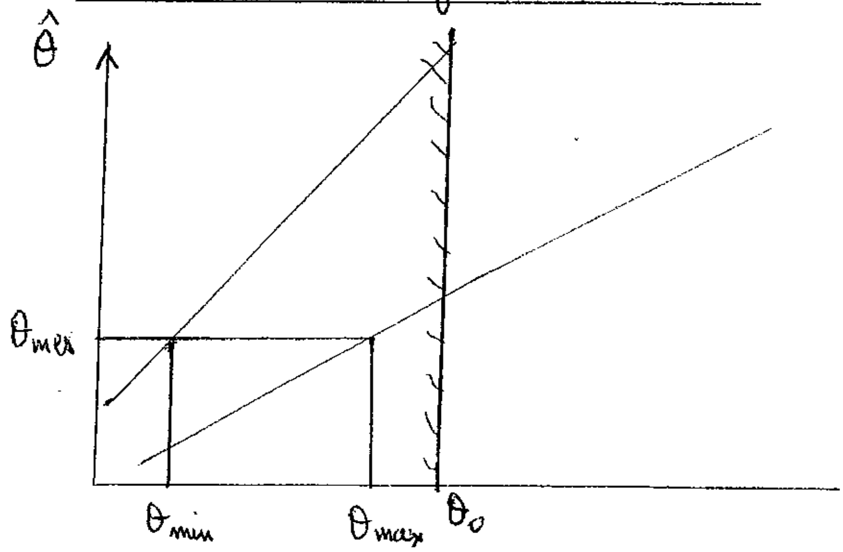
$P(\hat{\theta} < b_2(\theta)) = 5\%$

$P(\hat{\theta} < B_M(\theta)) = 10\%$

Conclusion

One has to decide beforehand whether to publish an interval or an upper bound, independantly of the value found for $\hat{\theta}$

B. Null results for bounded θ



I publish

$\theta_{min} < \theta < \theta_{max}$

while I know that $\theta > \theta_0$!

I am victim of a large fluctuation of $\hat{\theta}$ towards low value

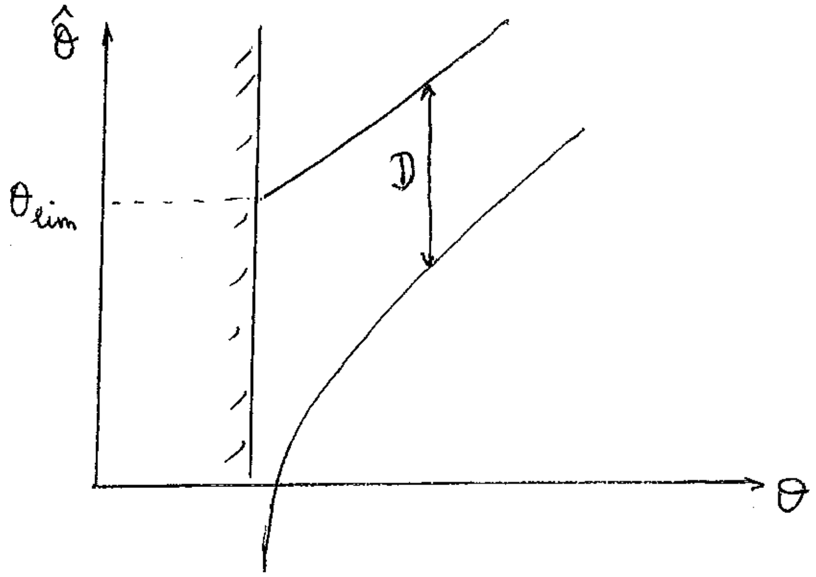
Cure : Feldman-Cousins construction

New way of building $b_1(\theta)$ and $b_2(\theta)$

For every physical value of θ, build the domain D of $\hat{\theta}$ values such that

$$\int_{r>r_c} f(\hat{\theta};\theta)\, d\hat{\theta} = 1-\alpha \quad \text{with } r = \frac{f(\hat{\theta};\theta)}{f(\hat{\theta};\theta_{best})}$$

θ_{best} is the physical value of θ maximizing $f(\hat{\theta};\theta)$



1) Solves flip-flop. The CL of the upper bound if $\hat{\theta} < \theta_{lim}$ or of the interval if $\hat{\theta} > \theta_{lim}$ is exactly $1-\alpha$

2) Solves null results (actually, not always true)

3) gives a straightforward generalization for multidimensional θ [$r(\hat{\theta};\theta)$ gives an ordering for $\hat{\theta}$ whatever the dimension of θ]

This method is now recommended by P.D.G.

$\hat{\theta}$ estimator of θ_0 with variance $\sigma_{\hat{\theta}}^2$

I shall write $\theta_0 = \hat{\theta} \pm \sigma_{\hat{\theta}}$ (statistical uncertainty)

$\sigma_{\hat{\theta}} \to 0$ like $N^{-1/2}$

But due to some limitations, the physical parameter is $\theta_0 + b$, and I have an estimator $\hat{b}$ of b of mean 0 and variance $\sigma_{\hat{b}}^2$.

Thus I will write $\theta_0 = \hat{\theta} \pm \sigma_{\hat{\theta}} \pm \sigma_{\hat{b}}$ (systematic uncertainty)

which means $\theta_0 = \hat{\theta} \pm \sqrt{\sigma_{\hat{\theta}}^2 + \sigma_{\hat{b}}^2}$

$\sigma_{\hat{b}}$ is constant when $N \to \infty$

Example 2 experiments at LEP measure M_{Z_0}

But the energy scale is known up to a translation of mean 0 and variance σ_E^2. Thus

Exp^t 1 $\quad M_{Z^0} = M_1 \pm \sigma_1$ (stat) $\pm \sigma_E$ (syst)

Exp^t 2 $\quad M_{Z^0} = M_2 \pm \sigma_2$ (stat) $\pm \sigma_E$ (syst)

$$V_{M_1 M_2} = \begin{pmatrix} \sigma_1^2 + \sigma_E^2 & \sigma_E^2 \\ \sigma_E^2 & \sigma_2^2 + \sigma_E^2 \end{pmatrix}$$

$\sigma_1^2 = \sigma_2^2 = \sigma^2$ Combining expts:

$$M_{Z^0} = \frac{M_1 + M_2}{2} \pm \frac{\sigma}{\sqrt{2}} \text{ (stat) } + \sigma_E \text{ (syst)}$$

HYPOTHESIS TESTING

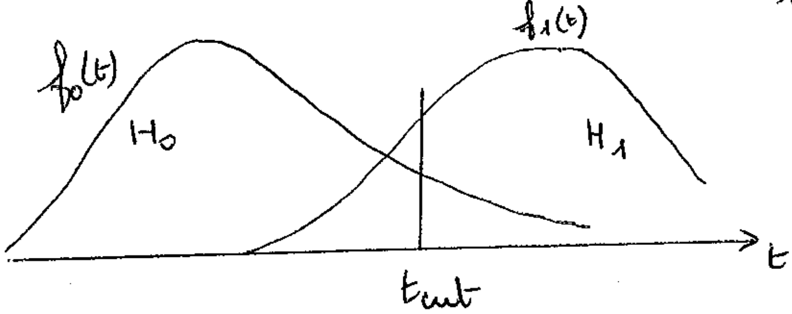
I Classifying observations between 2 categories

Each observation summarized by a few "discriminating" random variables x_i

Build a test t as a function of these x_i

If the observation belongs to category H_0, pdf of t is $f_0(t)$

" " " H_1 " $f_1(t)$



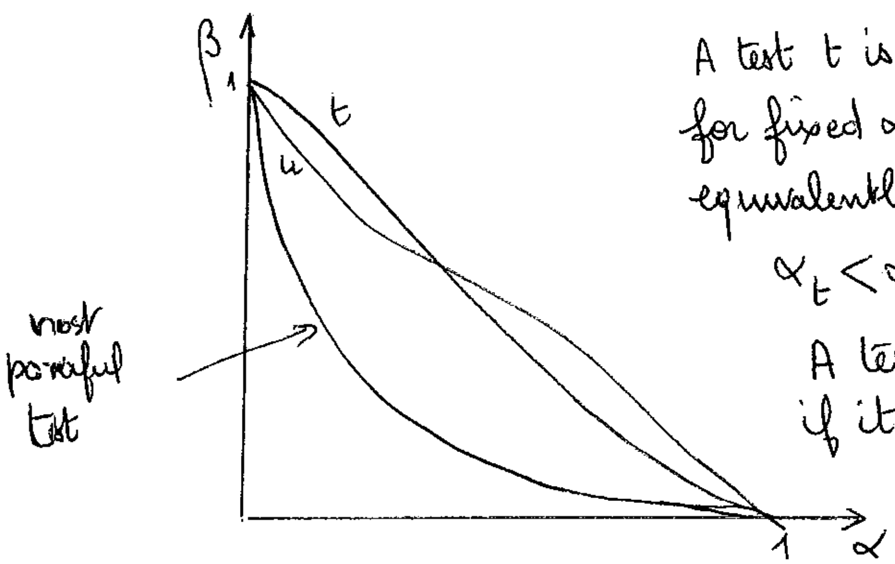
Choose t_{cut} Reject H_0 if $t > t_{cut}$ and put in category 1

Reject H_1 if $t < t_{cut}$ and put in category 0.

For category H_0 the loss of the test is $\alpha = P(t > t_{cut} | H_0)$

and the contamination is $\beta = P(t < t_{cut} | H_1)$

For category H_1, loss is β and contamination is α



A test t is better than test u if for fixed α, $\beta_t < \beta_u$ or equivalently if for fixed β, $\alpha_t < \alpha_u$

A test is said most powerful if it is better than any other for all values of α and β

<u>Theorem</u> The most powerful test for totally specified hypotheses (no unknown parameter) H_0 and H_1 is the likelihood ratio

$$t(\vec{x}) = \frac{f(\vec{x}\,|H_0)}{f(\vec{x}\,|H_1)} = \frac{f_0(\vec{x})}{f_1(\vec{x})}$$

<u>Parametric hypotheses</u>

$$f(\vec{x};\theta) \qquad \begin{array}{ll} H_0 & \theta \in \Omega_0 \\ H_1 & \theta \in \Omega_1 \end{array} \qquad \Omega_0 \cap \Omega_1 = \emptyset$$

$$t(\vec{x}) = \frac{\operatorname{Max}_{\theta \in \Omega_0} f(\vec{x};\theta)}{\operatorname{Max}_{\theta \in \Omega_0 \cup \Omega_1} f(\vec{x};\theta)}$$

$t \in [0,1]$ It is not a most powerful test

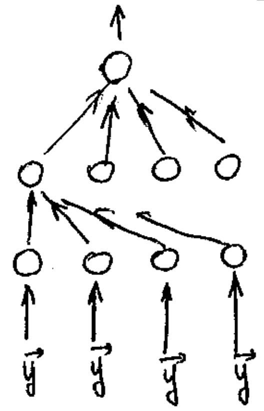
If $\Omega_0 = \{\theta_0\}$ and $\Omega_0 \cup \Omega_1 = \Omega$, then $-2\ln t$ will asymptotically be distributed according to a χ^2_k law (k being the dimension of θ) when H_0 is true.

Application = search for a small potential signal

2 categories : Signal, background

1. Study characteristics of each category using MC simulation
2. Determine a set of "discriminating variables" for each event (topology, missing energy, ...), noted $\vec{y}$
3. Tune the MC in areas where there is no signal
4. With MC simulations, determine the p.d.f of $\vec{y}$ for signal events $f_S(\vec{y})$ and for background events $f_B(\vec{y})$. In general too difficult = approximate f_S and f_B keeping only the higher correlations
5. Determine the distributions of the (pseudo) likelihood ratio $r(\vec{y}) = \dfrac{f_S(\vec{y})}{f_B(\vec{y})}$ for signal and for background.
6. Determine the best value at which to cut on r to get the best sensitivity on the signal, i.e. maximize $\varepsilon_S / \sqrt{\varepsilon_B}$ for the cut.
7. Finally, look at real data, selecting those events with $r > r_{cut}$, and compare to the background predicted by the MC simulation
8. Conclude on the presence/absence of ~~a~~ signal and give an estimate of the signal size.

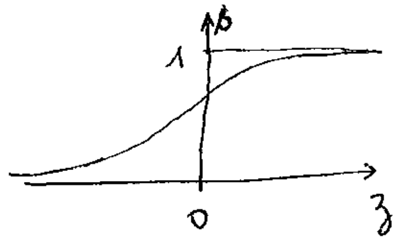
Approximating the likelihood ratio with a neural network

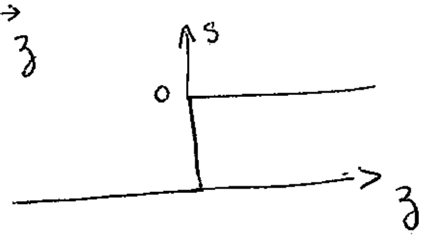


each neuron in one row receive as inputs the outputs of the previous row

The neuron j builds a linear combination of its inputs e_i $\quad z_j = \sum_k w_{jk} e_k + c_j$

and outputs $s_j = \underbrace{\left(1 - e^{-\lambda z_j}\right)^{-1}}_{\text{sigmoid of temperature } \lambda}$





at 0 temperature

$\boxed{\text{Perceptron}}$

o

o o o o → each ~~...~~ neuron defines an hyperplane in $\vec{y}$ space

→ this neuron selects a volume in $\vec{y}$ space limited by the hyperplanes

<u>Training</u>: determine w_{jk} so that the output s is near 0 for background events and near 1 for signal events by minimizing $\displaystyle\sum_{N_s} (s_i - 1)^2 + \sum_{N_B} s_i^2$

At finite λ, the boundary is made non planar and the transition between 0 and 1 smoother, s approximating in a non linear way the likelihood ratio

"GOODNESS OF FIT" TESTS

Are my observations compatible with hypothesis H_0?

Likelihood is useless.

$\Rightarrow$ Use a χ^2 test on observations. It gives χ^2_0

If H_0 true the p.d.f. for χ^2 is known or can be determined by M.C. simulation, so that

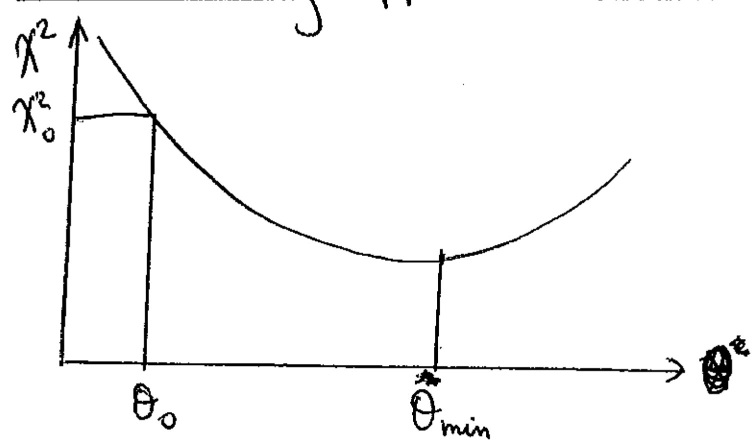
$$P(\chi^2 > \chi^2_0 \mid H_0) = \alpha$$

α is called the probability of the test. If α is small,

One will reject H_0 ~~at~~ at $(1-\alpha)$ CL

(the loss for H_0 from this test being α)

Understanding apparent paradoxes



Testing $\theta = \theta_0$ $\chi^2(\theta_0)$ follows a χ^2_k law if $\theta = \theta_0$

Testing θ family $\chi^2(\theta_{min})$ follows a χ^2_{k-1} law if some θ describes data

Estimating θ: One takes for granted that Nature has chosen the θ family for some value θ_v

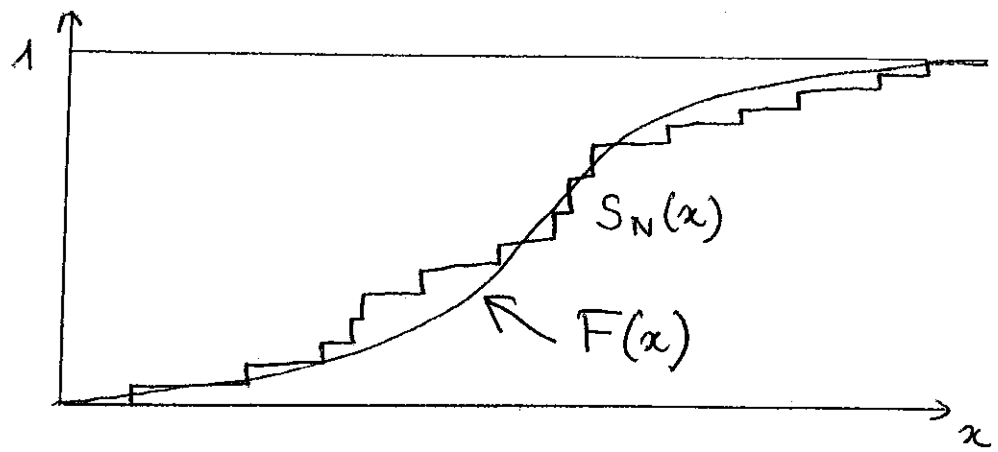
Estimation based on $\chi^2(\theta) - \chi^2(\theta_{min})$

3 different ways of using $\chi^2(\theta) \Rightarrow$ Different answers.

Tests better than χ^2

N observations x_i. Are they compatible with p.d.f $f(x)$?

Build $S_N(x)$ as a staircase fonction with steps of height $\frac{1}{N}$ at each value x_i, and compare $S_N(x)$ with the cumulative fonction $F(x)$



<u>SMIRNOV test</u> $W_N^2 = \int_{-\infty}^{+\infty} (S_N - F)^2 f(x)\, dx$

if H_0 true W_N p.d.f is a universal tabulated function

<u>KOLMOGOROV test</u> $K_N = \mathrm{Max}\, |F - S_N|$

if H_0 true, K_N p.d.f is asymptotically ($N > 180$) a universal tabulated law.

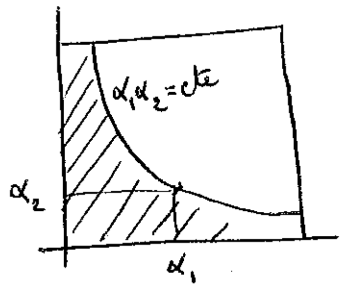
COMBINING 2 INDEPENDANT TESTS OF H_0

Test 1 has a probability α_1 for H_0

Test 2 has a probability α_2 for H_0

Note: If H_0 true α_1 p.d.f is flat on $[0,1]$
α_2 " "

$\alpha = \alpha_1\alpha_2$ $P(\alpha < a) = a(1 - \ln a)$



The probability of the combined test is therefore

$$\alpha_1\alpha_2\,(1 - \ln \alpha_1\alpha_2)$$

Application Test 1 rejects H_0 at 83% CL $[\alpha_1 = 0.17]$
Test 2 rejects H_0 at 75% CL $[\alpha_2 = 0.25]$

$\Rightarrow$ Combined test rejects H_0 at 82% CL

Reference

BOOKS

- Kendall & Stuart   The bible!
  "The advanced theory of statistics" 2 volumes
  ed. GRIFFIN , London

- Eadie et al.   Difficult (and old)
  "Statistical methods in experimental physics"
  Ed. North Holland, Amsterdam (1971)

- Byron Roe
  "Probability & statistics in experimental physics
  Ed. Springer

In french

- Jean-Pierre Lecoutre
  "Statistique et probabilité"
  Ed. DUNOD