Tests Estadísticos

Curso de Estadística

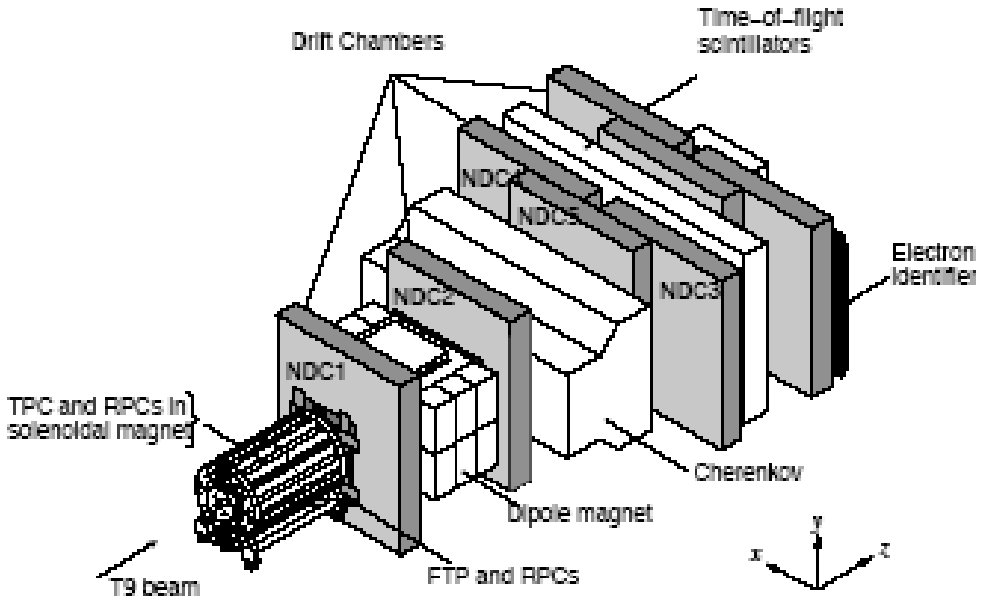
TAE, 2005

J.J. Gómez-Cadenas

Introducción

Considerar el experimento HARP:

pAl →p,π,K,e...



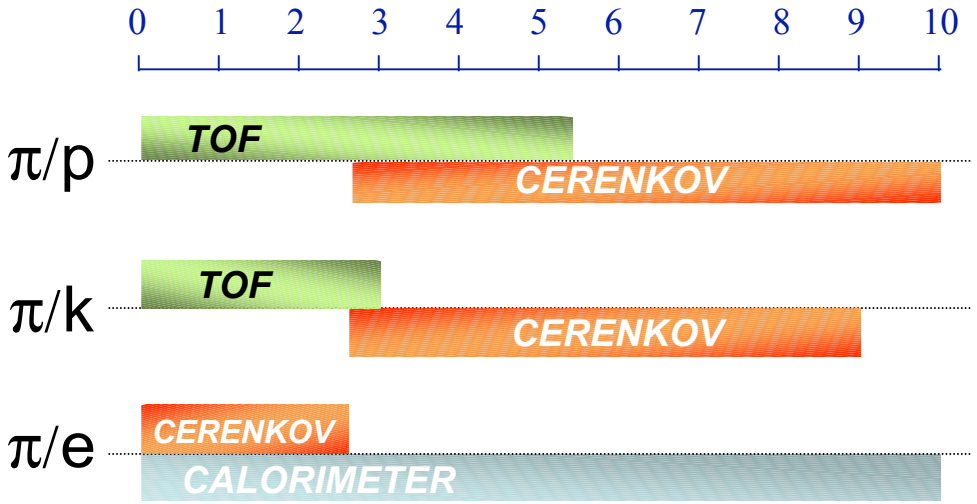
Objetivo: Medir la sección eficaz de producción de piones en función del momento y el ángulo sólido.

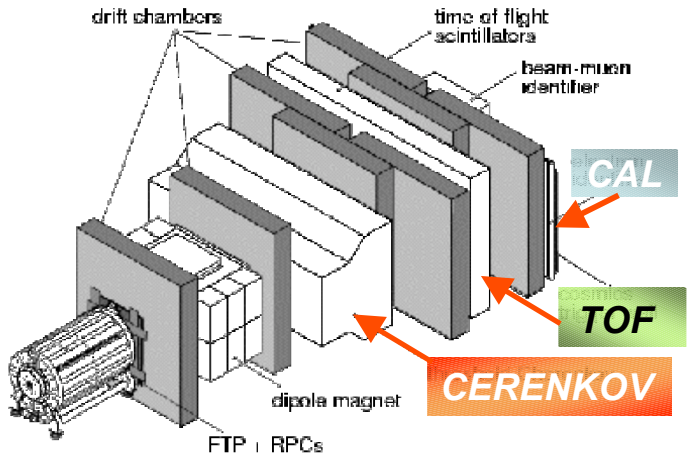
Para ello:

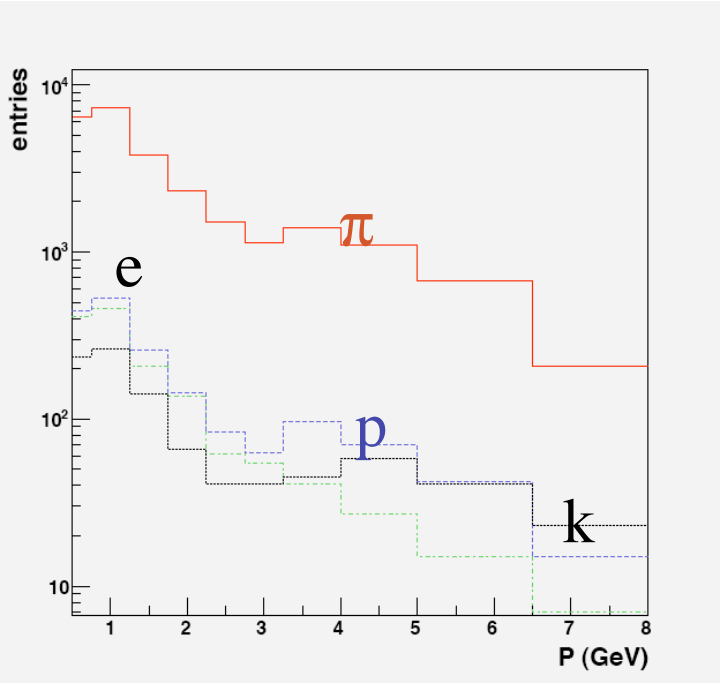
Contamos partículas en bines de (p,θ,ϕ)

Aplicamos criterios que permitan separar los π del resto de las partículas.

Detectores de PID: TOF,CHE, ECAL







Simulación MC de la distribución de partículas en función del momento

Detectores de PID en HARP

TOF →Mide la β de la partícula

CHE →Da señal para piones y electrones, no da señal para kaones y protones

ECAL →Mide la energía electromagnética. Separa electrones de hadrones

Suponer que el resultado de una medida en HARP es $x=(x_1,...,x_n)$. Por ejemplo:

x_1 = p (momento de la partícula)

$x_2 = \beta$ (velocidad estimada de la partícula con el TOF)

$x_3 = \alpha$ (respuesta del CHE: si $\alpha > 0$ la partícula es un pion o un electrón)

$x_4 = \gamma$ ($g = E_{ecal}/E_{tot}$)

x sigue una pdf conjunta, que en el caso de HARP depende del tipo de partícula producida (distinta distribución angular y de momento, distintas propiedades para β, α y γ).

El objetivo de un test estadístico es pronunciarse respecto a cuán bien un conjunto de datos observados (el vector **x**) describe un conjunto de probabilidades predichas, es decir una ***HIPÓTESIS.***

La hipótesis bajo estudio se llama tradicionalmente la hipótesis nula H_0 y a menudo especifica cierta pdf f(x) del vector **x** (hipótesis simple). Generalmente, para establecer la validez de H_0 es necesario compararla con hipótesis alternativas H_1, H_2...

Específicamente en el ejemplo que nos ocupa (HARP):

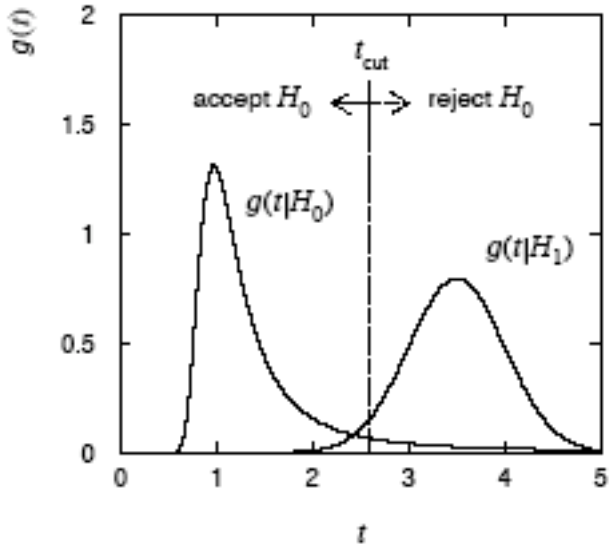
H_0 →La partícula observada, caracterizada por el conjunto x de medidas es un pion (señal)

H_1 →La partícula observada, caracterizada por el conjunto x de medidas no es un pion (ruido de fondo, no es importante si se trata de un protón, kaon, electrón, etc.)

A fin de investigar el grado de acuerdo entre la medida y la hipótesis, podemos construir un test estadístico $t(x)$. Cada una de las hipótesis implicará una pdf para t, e.g, $g(t|H_0)$, $g(t|H_1)$, etc.

t puede ser un vector multidimensional, aunque en general, por simplicidad, es conveniente que las dimensiones de t sean pequeñas como sea posible sin perder la capacidad de discriminar entre las hipótesis.

En nuestro ejemplo, supongamos que hemos construido una función escalar, $t(x)$ a partir del vector de observaciones x, que sigue la pdf $g(t|H_0)$ si H_0 es verdadera y la pdf $g(t|H_1)$ si H_1 es verdadera



A menudo expresamos la compatibilidad entre una hipótesis y los datos en términos de aceptar o rechazar la hipótesis nula H_0 (la partícula es un pion).

Para ello definimos un valor $t=t_{cut}$ tal que la hipótesis se rechaza si $t>t_{cut}$ (región crítica) o se acepta si $t< t_{cut}$ (región de tolerancia).

t_{cut} se escoge de tal manera que la probabilidad de que $t> t_{cut}$ siendo H_0 verdadera es un cierto valor α llamado el nivel de relevancia del test.

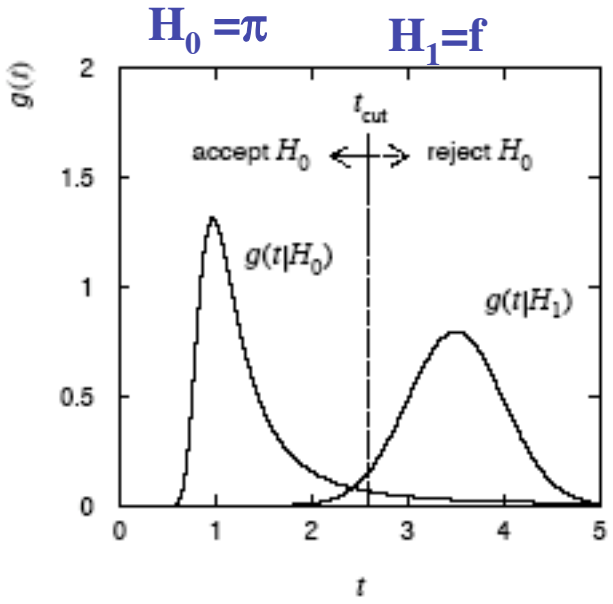
$$\alpha = \int_{tcut}^{\infty} g(t \mid H_0)dt$$

α es la probabilidad de rechazar H_0, siendo H_0 verdadera. La probabilidad de que un suceso x distribuido de acuerdo a H_1 (ruido de fondo) resulte en un test *t* tal que $t<tcut$ (por lo tanto H_1 se acepta) es:

$$\beta = \int_{-\infty}^{tcut} g(t \mid H_1)dt$$

$1-\beta \rightarrow$Poder del test para discriminar en contra de H_1

Ejemplo: Selección de partículas en HARP cortando en t



$$\varepsilon_\pi = \int_{-\infty}^{tcut} g(t \mid \pi) dt = 1 - \alpha$$

$$\varepsilon_f = \int_{-\infty}^{tcut} g(t \mid f) dt = \beta$$

t(x) = test estadístico construido a partir del vector de observaciones x. Combina de alguna manera no especificada la información de los diferentes detectores de PID. En el caso más sencillo: → Corte en una variable.

Seleccionamos piones requiriendo $t<t_{cut}$:

Si relajamos el corte →aceptamos más π pero también más ruido de fondo

Si lo hacemos más estricto→perdemos eficiencia, pero la muestra de π es más pura

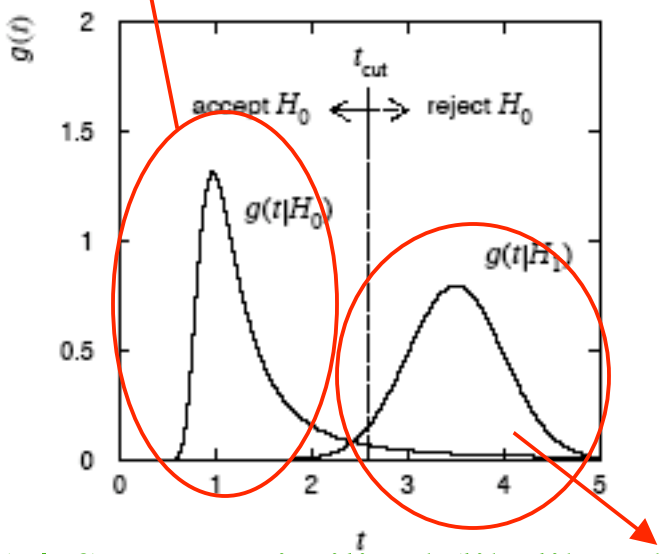
EFICIENCIA ⇔ PUREZA

NB: El valor de tcut viene determinado por nuestra decisión A PRIORI de los valores para α y β (o lo que es lo mismo ε_π y ε_f). Los cortes no se deciden a ojo!

Ejemplo: Selección de partículas en HARP usando el Teorema de Bayes

En lugar de cortar en t, podemos calcular la probabilidad de que una partícula con un valor observado de t sea un pión o bien fondo a partir de las pdfs g(t|π) y g(t|f) usando el teorema de Bayes

$g(t \mid \pi)$ = Verosimilitud (likelihood) de π



$g(t \mid f)$ = Verosimilitud (likelihood) de f

$$h(\pi \mid t) = \frac{n_\pi g(t \mid \pi)}{n_\pi g(t \mid \pi) + (1 - n_\pi) g(t \mid f)}$$

$$h(f \mid t) = \frac{(1 - n_\pi) g(t \mid f)}{n_\pi g(t \mid \pi) + (1 - n_\pi) g(t \mid f)}$$

Donde n_π y (1- n_π) son los priores para las hipótesis $H_0(\pi)$ y $H_1(f)$ (en nuestro caso: las abundancias relativas de piones y fondo)

Problema en Harp: n_π no se conoce (precisamente es lo que queremos medir)!

El problema de los Priores

NB: Hemos planteado el problema en términos de estadística Bayesiana. La aproximación es muy potente. Podemos pesar cada partícula por su probabilidad de ser pion y medir así directamente la sección eficaz, sin tener que corregir por eficiencias y purezas.

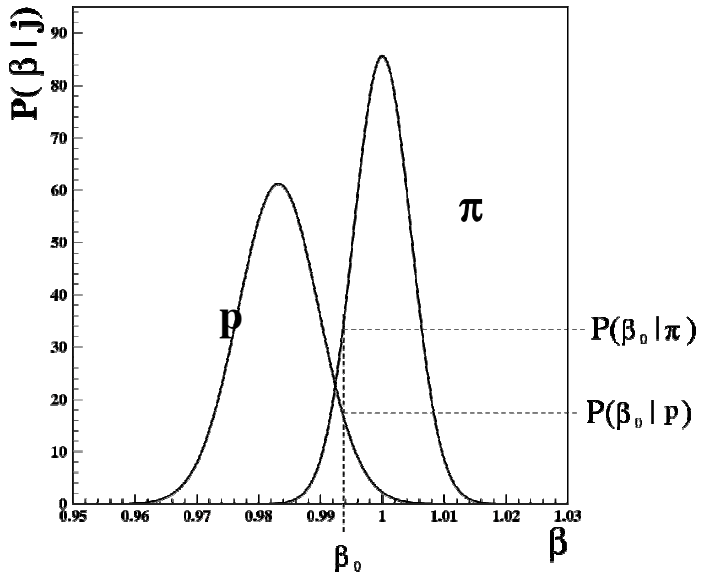
El problema es que no conocemos el prior (se trata de un caso corriente).

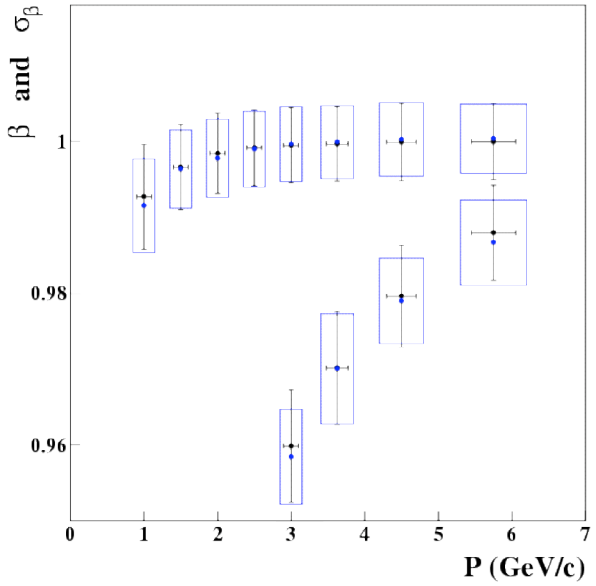
Sin embargo es posible utilizar una estimación del prior. Por ejemplo podríamos usar el Monte Carlo para predecir la fracción relativa de n_π. Dos puntos deben tomarse en cuenta para estimar el prior

- El prior debe maximizar la entropía tomando en cuenta cualquier información o ligaduras existentes.
- El resultado no puede depender del prior en el límite de estadística infinita (y debe depender sólo débilmente si la estimación es robusta)

En general, cuando sustituimos el prior por un estimador en la fórmula de Bayes, el resultado ya no es una probabilidad estrictamente, sino un estimador (una cantidad en la que podemos cortar!)

Construcción de un test estadístico. Un solo detector





¿quiénes son las funciones g(tlp) y g(tlf)?

HARP: A partir del TOF medimos β para piones y ruido de fondo (protones).

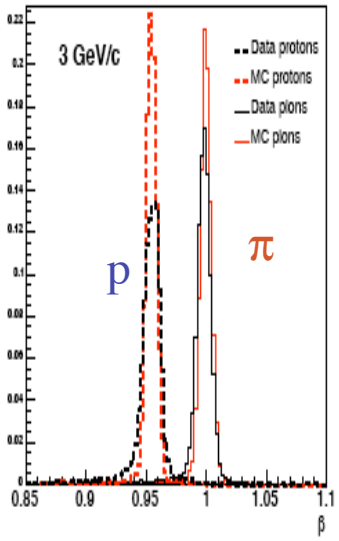
Para ello usamos piones y protones monocromáticos que nos proporcional el haz del SPS. Sabemos pues exactamente, cuál es la respuesta del TOF a esta partículas En la figura se muestra la distribución de β, a 5 GeV para ambos tipos.

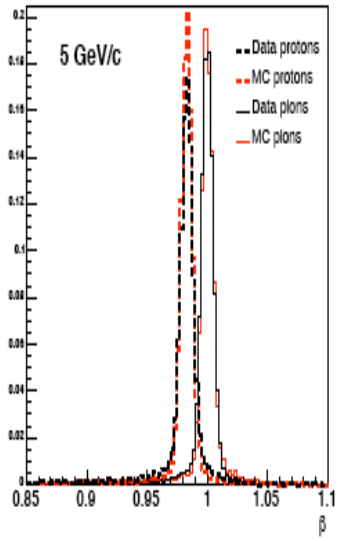
Reptiiendo las medidas a diferentes energías del haz obtenemos la dependencia en **p**

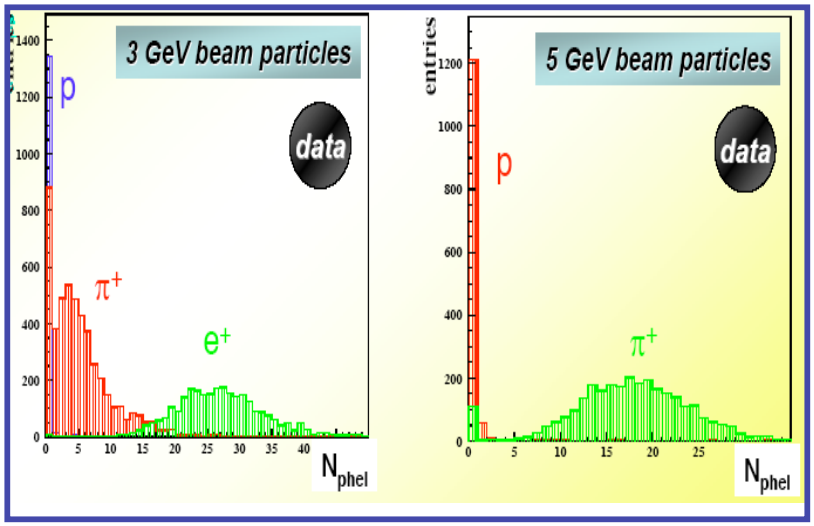
$g(\beta \mid \pi, p)$ = Verosimilitud (medida) de π para β(p)

$g(f \mid \pi, p)$ = Verosimilitud (medida) de fondo para β(p)

Combinación de varios detectores







A menudo se dispone de varios detectores, para separar la señal y el ruido. En Harp:

TOF (β)

CHE (α = Nphe)

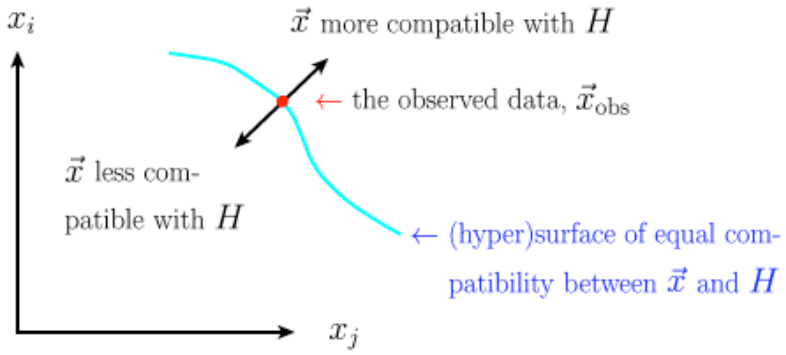
ECAL (γ = E1/Etot)

Por lo tanto la verosimilitud de ser un pion es una función multidimensional g(**x**|π,p), donde **x**= (α,β,γ…)

Si los detectores son independientes, la verosimilitud total no es más que el producto de las verosimilitudes individuales

Tests de calidad de un ajuste

Problema: Especificar cuán compatibles son los datos con una hipótesis nula H sin hacer referencia a una hipótesis alternativa.



Hipótesis H (modelo) predice f(**x**|H) para un cierto vector de datos $\boldsymbol{x} = (x1, \ldots, x_n)$

Observamos el punto x_{obs}

Qué podemos decir sobre la validez de H a partir de los datos?

Calidad de un ajuste: Construir un test estadístico t(x) cuyo valor refleje el grado de acuerdo entre los datos **x** y el modelo H.

Cuando t decrece →los datos son más compatibles con H

Cuando t aumenta →los datos son menos compatibles con H

Puesto que la pdf $f(x|H)$ se conoce, la pdf $g(t|H)$ puede determinarse

Nivel de confianza

La calidad de un ajuste puede determinarse en términos de una cantidad llamada valor-P, también conocida como nivel de confianza, definida como:

P = probabilidad de observar datos x (o t(x)) con un nivel de compatibilidad con el modelo H igual o menor que la observación realizada x_{obs} (o $t(x_{obs})$).

NB: P no especifica la probabilidad de que H sea cierta.

De hecho, en estadística clásica, P(H) no está bien definido (si lo está en estadística Bayesiana) y la forma en que lo utilizamos es bastante vaga. Refleja la verosimilitud de que los datos x sigan el modelo H.

P es una variable aleatoria tal que:

Si H es cierto entonces para x continuo P es uniforme en [0,1]

Si H es falsa entonces la pdf de P tiene (habitualmente) un pico cerca del 0

Un ejemplo académico: Monedas falsas

Nos proponemos estudiar si una determinada moneda es auténtica (igual probabilidad de dar cara y cruz) o falsa (descompensada para dar más o menos caras que cruces).

Para ello recordamos que la probabilidad de observar n_c caras cuando lanzamos la moneda N veces sigue una distribución binomial:

$$f(n_c; p_c, N) = \frac{N!}{n_c!(N-n_c)!} p_c^{n_c} (1-p_c)^{N-n_c} \qquad H : p_c = p_{cr} = 0.5$$

Tomamos como test estadístico de la calidad del ajuste:

$$t = \left| n_c - \frac{N}{2} \right| \begin{cases} t \to 0 & \text{H es probablemente cierta (moneda auténtica)} \\ t >> 0 & \text{H es probablemente falsa (moneda falsa)} \end{cases}$$

Suponer que lanzamos la moneda 20 veces y obtenemos 17 caras. ¿Cual es la verosimilitud del modelo (H : La moneda es buena)?

$$t_{obs} = \left| 17 - \frac{20}{2} \right| = 7$$

Para calcular el nivel de confianza, consideramos la región del espacio (en t) con la misma o menor compatibilidad con el modelo. Es decir, $t \geq 7$, que se corresponde a $n_c^i = (0,1,2,3,17,18,19,20)$

A continuación aplicamos la distribución binomial para calcular la probabilidad suma de las probabilidades correspondientes a n_c^i

$$P = P_{t \geq t_{obs}} = f(0;0.5,20) + f(1;0.5,20) + \ldots + f(19;0.5,20) + f(20;0.5,20) = 0.0026$$

¿Significa este resultado que la moneda es falsa? En estadística clásica, la pregunta no tiene respuesta. El nivel de confianza o valor P sólo da la probabilidad de que el nivel de discrepancia sea igual o mayor, *por azar* al obtenido comparando la hipótesis (moneda buena) y el resultado observado (17 caras). En otras palabras, si efectuáramos un número muy grande de lanzamientos, un 0.26 % de estos podrían arrojar este resultado *para una moneda buena*

Relevancia (significado) de una señal observada

Suponer que observamos n sucesos, que no ajustan bien a un cierto modelo. Por ejemplo, el número de desintegraciones del Z a tau, en los primeros datos de Mark-II era mucho más alto del que se correspondería a la predicción de un BR de 3%.

Los sucesos que observamos contienen, presumiblemente :

n_b sucesos de ruido de fondo (procesos conocidos, Z-->$\tau^+\tau^-$)

n_s sucesos de señal (¿nueva física resultante en un exceso de taus?)

Si n_b y n_s están distribuidos Poisson con medias ν_b y ν_s, entonces $n=n_s+n_b$ también está distribuido Poisson, con media $\nu = \nu_b + \nu_s$,

$$P(n;\nu_s,\nu_b) = \frac{(\nu_s+\nu_b)^n}{n!} e^{-(\nu_s+\nu_b)}$$

Supongamos que ν_b=0.5 y que observamos n_{obs} =5 ¿Podemos afirmar que tenemos evidencia de nueva física?

Para responder: La hipótesis H (modelo) es que *no* hay nueva física, es decir que $\nu_s = 0$.

El nivel de confianza P se corresponde a la probabilidad de observar por azar $n > n_{obs}$

$$P(n \geq n_{obs}) = \sum_{n=n_{obs}}^{\infty} P(n;\nu_s = 0,\nu_b = 0.5) = \sum_{n=n_{obs}}^{\infty} \frac{(\nu_b)^n}{n!} e^{-\nu_b}$$

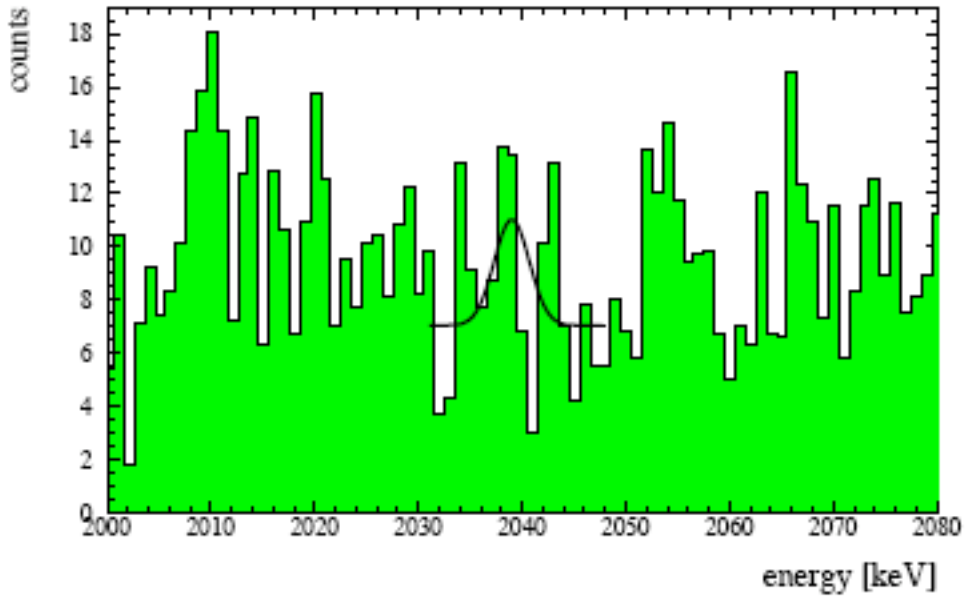
$$= 1 - \sum_{n=0}^{n_{obs}-1} \frac{\nu_b{}^n}{n!} e^{-\nu_b} = 1 - \sum_{n=0}^{4} \frac{0.5^n}{n!} e^{-0.5} = 1.7 \times 10^{-4}$$

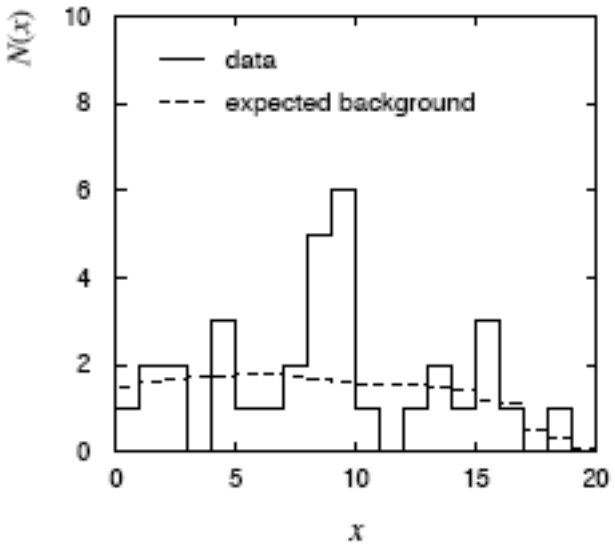
Es decir, la probabilidad, si $\nu_s = 0$ (si no hay señal) de observar por azar 5 o más sucesos es del orden de 10^{-4}. Con lo cual, sería posiblemente correcto asumir que hemos descubierto nueva física, es decir que $\nu_s \neq 0$

PERO ATENCIÓN: Hemos asumido que νb se conoce sin error, lo cual en general no es cierto. Por ejemplo para ν_b=0.8, P aumenta casi un orden de magnitud. Es por lo tanto esencial cuantificar los errores sistemáticos en el ruido de fondo para evaluar el nivel de significado de un nuevo efecto.

Relevancia de un pico

Ejemplo: “Evidencia” de la existencia de desintegración doble-β (Klapdor et al)





Binamos los datos en un histograma.
Conocemos la forma del ruido de fondo.
Cada bin está distribuido Poisson

Suponer que en los 2 bines en los que encontramos el pico hay 11 entradas y que la media (estimada) del fondo es ν_b=3.2.
Entonces:

$$P(n \geq 11; \nu_s = 0, \nu_b = 3.2) = 5 \times 10^{-4}$$

¡Premio Nobel!

Pero... ¿Cómo sabemos dónde encontrar el pico? (podría ser una simple fluctuación, debida a nuestro binado). Algunas precauciones que debemos tomar:

¿Cuál es la probabilidad P($n \geq 11$) para cualquier par de bines adyacentes?
debería ser mucho más pequeña de la que hemos encontrado

¿Es la anchura del pico consistente con la resolución esperada?

asegurarnos que el pico tiene una anchura varias veces superior a la resolución que esperamos

¿Es posible que los cortes de selección estén "forzando" accidentalmente un pico?

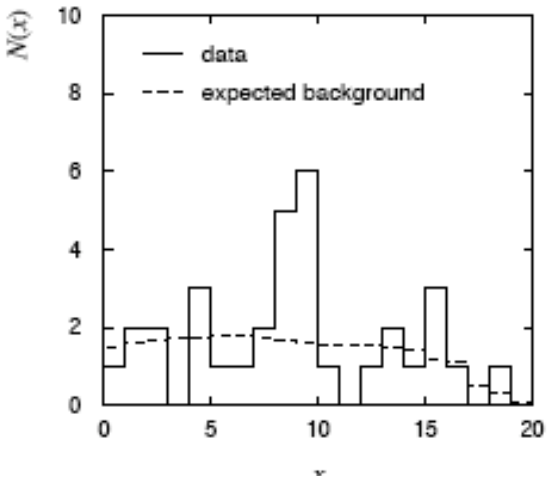
congelar cortes, repetir el análisis con una submuestra diferente

¿Qué forma tienen los bines adyacentes al pico?

quizás están deprimidos, lo que podría indicar una fluctuación

En resumen: Es mucho más difícil convencerse de que un pico (pequeño) implica nueva física de lo que parece!

El test chi2 de Pearson



$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

$$= 29.8 \text{ for } N = 20 \text{ dof.}$$

Supongamos que tenemos un histograma de la variable x, con N bins, de tal manera que en cada bin observamos $n=(n_1,n_2,...,n_N)$.

Los valores esperados en cada bin son

$$\nu = (\nu_1,\nu_2,...,\nu_N).$$

El test Chi2 de Pearson refleja el nivel de acuerdo entre el histograma observado y el histograma esperado:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

Si el número de entradas por bin no es demasiado pequeño (típicamente 5 mayor) entonces el test de Pearson sigue la distribución χ^2 con N grados de libertad.

El χ^2 observado, arroja entonces un nivel de confianza:

$$P(\chi^2 \geq \chi_0^2) = \int_{\chi_0^2}^{\infty} f(z;N)dz$$

donde f(z;N) es la pdf de la distribución χ^2 con N grados de libertad.

A menudo, en lugar de utilizar P se utiliza χ^2 /N. Se considera el ajuste bueno para valores de χ^2 /N próximos a uno. La justificación es que E[z]=N.

Sin embargo, en general, es preferible dar el valor P, obtenido a partir de χ^2 y N. Por ejemplo:

$$\chi^2 = 15,\ N = 10 \rightarrow P\text{-value} = 0.13$$
$$\chi^2 = 150,\ N = 100 \rightarrow P\text{-value} = 9.0 \times 10^{-4}$$