Probabilidad y estadística para la física experimental

José Ocariz

Université Paris Cité

ocariz@in2p3.fr

June 25, 2024

Latin American alliance for Capacity buildiNG in Advanced physics

LA-CoNGA physics

Cofinanciado por el programa Erasmus+ de la Unión Europea

- El libro clásico de referencia (912 páginas) :
  - Stuart, K. Ord, S. Arnold, Kendall's Advanced theory of statistics Volume 2A : Classical Inference and the Linear Model, John Wiley & Sons, 2009
- Libros de estadísticas, escritos por físicos de partículas :
  - L. Lyons, Statistics for Nuclear and Particle Physics, Cambridge, 1986
  - G. Cowan, Statistical Data Analysis, Clarendon Press, Oxford 1998
    (ver también `http://www.p.rhul.ac.uk/~cowan/stat_course.htm`)
  - R.J. Barlow, A Guide to the Use of Statistical Methods in the Physical Sciences John Wiley & Sons, 1989
  - F. James, Statistical Methods in Experimental Physics, World Scientific, 2006
- El PDG también es una fuente conveniente para referencias rápidas :
  - 2020 Review of Particle Physics, P.A. Zyla et al. (Particle Data Group), Prog. Theor. Exp. Phys. 2020, 083C01 (2020) "Mathematical Tools" section
    (ver también `https://pdg.lbl.gov/`)
- Las grandes colaboraciones internacionales tienen foros y grupos de trabajo, con muchos enlaces y referencias.

Algunas consideraciones iniciales

- Este curso se inspira ampliamente de mi experiencia personal como físico experimental de partículas
  - el lenguaje, las notaciones, etc... reflejan los usos y costumbres de mi área temática
  - ¡otras comunidades tienen convenciones y definiciones diferentes!
  - en ocasiones indicaré diferencias en notaciones (cuando las conozco...)
- De la misma manera, las herramientas a las que estoy más acostumbrado provienen de la física de partículas
  - nunca he usado **R**
  - trabajo más a menudo con `C++` que con `python`, pero ambos me van
  - mi herramienta cotidiana para análisis y visualización es **ROOT** (`https://root.cern/`)
    - existe una interfaz `PyROOT` que enlaza `python` con el `C++` nativo de **ROOT**
  - para análisis estadísticos más elaborados (funciones de verosimilitud) utilizo el paquete `RooFit`
  - y en cuanto a Machine Learning, yo uso sobre todo **TMVA**, a `Toolkit for MultiVariate data Analysis`
    - nota : **TMVA** no incluye ciertas herramientas modernas tipo GAN, GNN, etc...
  - instalé `TensorFlow` en mi móvil, pero más por curiosidad que otra cosa...
  - pienso también probar con `PyTorch`, pero eso sigue en mi lista de tareas pendientes...
- Los ejercicios y tareas del curso pueden ser trabajados bajo forma de `notebooks Jupyter`
  - el uso de **ROOT** no es obligatorio (aunque por supuesto me sería más fácil a mí...)
  - me esfuerzo en que (buena parte de) los ejercicios no requieran funciones súper-específicas de HEP...
- Este documento retoma el material preparado para los cursos *LA-CoNGA physics* de los dos años opasados
  - Si bien ya cubre todos los capítulos del programa, sin duda ampliaré algunos segmentos a lo largo del curso

- Este curso *no* es un curso de Machine Learning (aunque hablaremos de Machine Learning)
- Este curso *no* es un curso sobre Teoría de Decisiones (que es una aplicación muy importante de la estadística)

Este curso *sí* trata de la inferencia científica, con el objetivo de pasar de lo intuitivo hacia una lógica más formal

- Cuando lees un artículo con un resultado así:

$$m_{\text{top}} = 173.34 \pm 0.27(\text{stat.}) \pm 0.71(\text{syst.})\text{GeV}$$

  ¿Cómo interpretas ese resultado?
- Cuando en otro artículo, te topas con una afirmación del estilo

  Gluino and squark masses up to 1.5 TeV are excluded at 95% C.L.

  ¿Cómo interpretas esa afirmación?
- Cuándo en otro artículo, lees una frase del estilo

  The most significant deviation with respect to the background-only hypothesis is observed for a mass of 19.35 GeV, corresponding to a local significance of 3.1σ

  ¿Cómo interpretas esa frase?

CAPITULO I

PROBABILIDAD MATEMATICA

Probabilidad matemática (I)

- ¿Qué es para tí la probabilidad? ¿Cómo la definirías?
- La *probabilidad matemática* es un concepto axiomático abstracto, desarrollado por Kolmogorov (1933) y otros
- La *teoría de la probabilidad* es el marco conceptual para el estudio de los *procesos aleatorios*
- Un proceso es llamado aleatorio si satisface dos condiciones :
  - su realización (un "evento") no puede ser predicha con total certeza ;
  - si el proceso se repite bajo las mismas condiciones, cada nueva realización puede ser diferente
- Es usual clasificar las fuentes de incertidumbre según su origen :
  - *reducibles* : errores en la la medición, p.e. limitaciones prácticas que en principio pueden ser mejoradas (mejores instrumentos, mejor control de las condiciones experimentales) ;
  - *cuasi-irreducibles* : errores aleatorios en la medición, como efectos térmicos o de turbulencia ;
  - *fundamentales* : cuando el proceso físico es intrínsecamente incierto (mecánica cuántica).

En física subatómica experimental, los tres tipos de incertidumbre deben ser considerados. Notar en particular :

- los eventos resultances de colisiones de partículas son independientes, y son un ejemplo perfecto de procesos aleatorios de origen cuántico
- las partículas inestables obedecen probabilidades de desintegración descritas por la mecánica cuántica

Ejercicio : dar ejemplos de procesos físicos para cada una de las fuentes de incertidumbre mencionadas arriba.

Sea Ω el universo total de posibles realizaciones de un proceso aleatorio, y sean $X, Y \ldots$ elementos de Ω
Una función de probabilidad $\mathcal{P}$ se define como un mapa en los números reales :

$$\begin{aligned} \mathcal{P} : \{\Omega\} &\rightarrow [0:1] \ , \\ X &\rightarrow \mathcal{P}(X) \ . \end{aligned}$$

Ese mapeo debe satisfacer los siguientes axiomas :

$$\begin{aligned} \mathcal{P}(\Omega) &= 1 \ , \\ \text{si } X \cap Y &= \oslash \ , \text{ entonces } \mathcal{P}(X \cup Y) = \mathcal{P}(X) + \mathcal{P}(Y) \ , \end{aligned}$$

de los cuales se pueden derivar varias propiedades útiles, p.e. (donde $\overline{X}$ es el complemento de X)

$$\begin{aligned} \mathcal{P}(\overline{X}) &= 1 - \mathcal{P}(X) \ , \\ \mathcal{P}(X \cup \overline{X}) &= 1 \ , \\ \mathcal{P}(\oslash) &= 1 - \mathcal{P}(\Omega) = 0 \ , \\ \mathcal{P}(X \cup Y) &= \mathcal{P}(X) + \mathcal{P}(Y) - \mathcal{P}(X \cap Y) \ , \end{aligned}$$

La probabilidad condicional $\mathcal{P}(X \mid Y)$ se define como la probabilidad de X, dado Y

- equivale a restringir el universo Ω a la muestra Y.

El ejemplo más sencillo de probabilidad condicional es para realizaciones independientes :

- dos elementos X e Y son independientes (sus realizaciones no estén relacionadas en ninguna manera) si

$$\mathcal{P}(X \cap Y) = \mathcal{P}(X)\mathcal{P}(Y).$$

- por lo tanto, si X e Y son independientes, se satisface la condición

$$\mathcal{P}(X \mid Y) = \mathcal{P}(X)$$

El teorema de Bayes cubre el caso general : en vista de la relación $\mathcal{P}(X \cap Y) = \mathcal{P}(Y \cap X)$, se obtiene que

$$\mathcal{P}(X \mid Y) = \frac{\mathcal{P}(Y \mid X)\mathcal{P}(X)}{\mathcal{P}(Y)} \ .$$

Un corolario útil del teorema de Bayes : si Ω puede dividirse en un número de submuestras disjuntas X_i (una "partición"), entonces

$$\mathcal{P}(X \mid Y) = \frac{\mathcal{P}(Y \mid X)\mathcal{P}(X)}{\sum_i \mathcal{P}(Y \mid X_i)\mathcal{P}(X_i)} \ .$$

CAPITULO II

VARIABLES ALEATORIAS

FUNCIONES DE DENSIDAD DE PROBABILIDAD

El escenario más relevante para nosotros es cuando la realización de un proceso aleatorio se presenta en forma numérica (p.e. corresponde a una medición) : a cada elemento X corresponde una variable x (real o entera). Para x continuo, su función de densidad de probabilidad (PDF) $P(x)$ se define como :

$$\mathcal{P}(X \text{ en } [x, x+dx]) \;=\; P(x)dx \;,$$

donde $P(x)$ es definida-positiva para todo de x, y satisface la condición de normalización

$$\int_{-\infty}^{+\infty} dx' P(x') \;=\; 1 \;.$$

Para x_i discreto, la definición es similar :

$$\mathcal{P}(X \;\; \text{en } x_i) \;=\; p_i \;,$$
$$\text{con } \sum_j p_j = 1 \text{ y } p_k \geq 0 \;\forall k \;.$$

Probabilidades finitas se obtienen por integración sobre un rango no-infinitesimal,

$$\mathcal{P}(a \leq X \leq b) \;=\; \int_a^b dx' P(x') \;.$$

Ejercicio: determinar el coeficiente de normalización de la función Gaussiana, $g(x;\mu,\sigma) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Y mejor aún: normalizar la Gaussiana en un intervalo truncado, con $a \leq x \leq b$.

En ocasiones es conveniente referirse a la función de densidad cumulativa (CDF) :

$$C(x) \;=\; \int_{-\infty}^{x} dx' P(x') \;,$$

de modo que las probabilidades finitas corresponden a evaluar la CDF en los bordes del rango de interés :

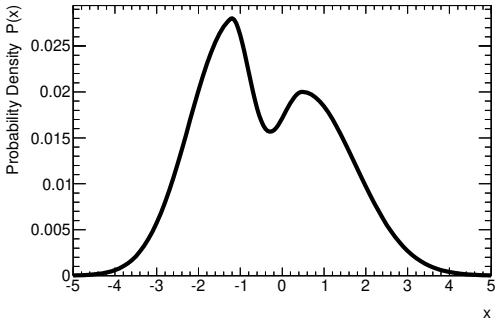
$$\mathcal{P}(a \leq X \leq b) \;=\; \int_{a}^{b} dx' P(x') \;=\; C(b) - C(a) \;.$$

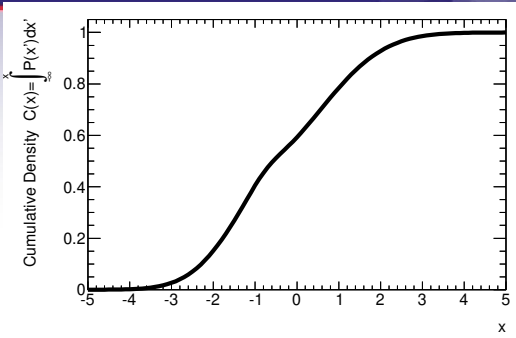
Una PDF no puede ser completamente arbitraria :

- debe satisfacer la condición de normalización previamente indicada
- debe ser definida positiva
- debe ser de soporte acotado, con valores despreciables fuera de una región finita

Fuera de esas condiciones, una PDF puede ser arbitraria, p.e. exhibir uno o varios máximos locales, tener discontinuidades...
En contraste, la CDF es una función monotónicamente creciente de x.
(ver ejemplo en la lámina siguiente)

Variables aleatorias, funciones de densidad de probabilidad (III)





Un ejemplo arbitrario de PDF con un máximo global y un segundo máximo local, y su CDF correspondiente. Fuera del intervalo en el gráfico, el valor de la PDF es totalmente despreciable.

Ejercicio : Suponer que nuestra PDF $P(x)$ es la suma de dos PDFs Gaussianas

$$P(x; \mu_1, \sigma_1, \mu_2, \sigma_2, f) = fG(x; \mu_1, \sigma_1) + (1 - f)G(x; \mu_2, \sigma_2) \ .$$

1. verificar que el parámetro f está limitado al intervalo $[0 : 1]$;
2. verificar que $P(x)$ está normalizada;
3. para los valores siguientes: $\mu_1 = -1$, $\sigma_1 = 1$, $\mu_2 = 1$, $\sigma_2 = 2$, $f = 0.5$. Representar gráficamente $P(x)$ y su CDF.

Para un evento descrito por un conjunto n-dimensional de elementos $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ y su correspondiente conjunto de variables aleatorias $\vec{x} = \{x_1, x_2, \ldots, x_n\}$, tenemos su PDF multidimensional :

$$P(\vec{x})d\vec{x} \;=\; P(x_1, x_2, \ldots, x_n)dx_1 dx_2 \ldots dx_n \;.$$

PDFs de menor dimensionalidad pueden derivarse por integración de ciertas variables. Por ejemplo, para una variable específica $x = x_j$ su densidad de probabilidad marginal unidimensional $P_X(x)$ es :

$$P_X(x)dx = dx \int_{-\infty}^{+\infty} dx_1 \ldots \int_{-\infty}^{+\infty} dx_{j-1} \int_{-\infty}^{+\infty} dx_{j+1} \ldots \int_{-\infty}^{+\infty} dx_n P(x_1, x_2, \ldots, x_n) \;.$$

Caso bidimensional, con elementos X, Y y variables aleatorias $\vec{X} = \{x, y\}$. La probabilidad finita en un rango bidimensional rectangular es

$$\mathcal{P}(a \leq X \leq b \;;\; c \leq Y \leq d) \;=\; \int_a^b dx \int_c^d dy P(x, y) \;.$$

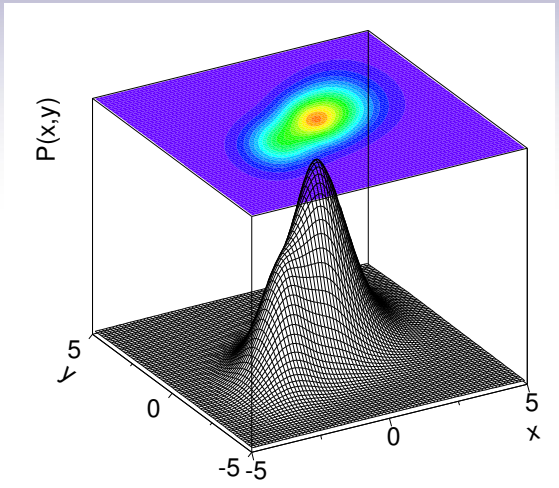
Para un valor fijo de Y, la función de densidad condicional de X es

$$P(x \mid y) = \frac{P(x, y)}{\int dx P(x, y)} = \frac{P(x, y)}{P_Y(y)} \;.$$

De nuevo, la relación $P(x, y) = P_X(x) \cdot P_Y(y)$ solamente es válida para X, Y independientes.

Ejemplo de una función de densidad bidimensional con variables no-independientes, $P(x, y) \neq P_X(x) \cdot P_Y(y)$.

Ejercicio : Producir un gráfico que represente una PDF bidimensional razonablemente similar a la ilustrada aquí.



Modelo : descripción de un proceso aleatorio
Modelo analítico = descripción de un proceso aleatorio con funciones analíticas para las PDFs
Modelo paramétrico : sus PDFs pueden describirse completamente usando un número finito de parámetros

- este requisito no es obligatorio; las PDFs pueden también ser no-paramétricas (equivalente a suponer que se necesita un número infinito de parámetros), o pueden ser mixtas

Una implementación sencilla de una PDF paramétrica es cuando sus parámetros son argumentos analíticos de la función de densidad ; la notación

$$P\left(x, y, \ldots ; \theta_1, \theta_2, \ldots\right)$$

indica la dependencia funcional o *forma* de la PDF en términos de variables $x_1, y_2, \ldots$ y parámetros $\theta_1, \theta_2, \ldots$.

Consideremos una variable aleatoria X con PDF $P(x)$. Para una función genérica $f(x)$, su *valor de expectación* $E[f]$ es su promedio ponderado sobre el rango cubierto por x :

$$E[f] \;=\; \int dx P(x) f(x) \;=\; \frac{\int dx P(x) f(x)}{\int dx P(x)} \; .$$

Como describiremos más adelante, los parámetros de una PDF pueden ser estimados a partir de ciertos valores de expectación.

Por ser de uso frecuente, algunos valores de expectación tienen nombre propio.
Para PDFs unidimensionales, la *media* y la *varianza* se definen así :

$$\text{Media} \quad : \mu = \quad E[x] = \int dx P(x) x \ ,$$

$$\text{Varianza} \quad : \sigma^2 = \quad V[x] = E[x^2] - \mu^2 = E[(x-\mu)^2] \ ;$$

y la *desviación estándar* σ es la raíz cuadrada de la varianza.
Para PDFs multidimensionales, la matriz de *covarianza* $C_{ij} = C(x_i, x_j)$ y la matriz adimensional de correlación linear ρ_{ij} se definen así :

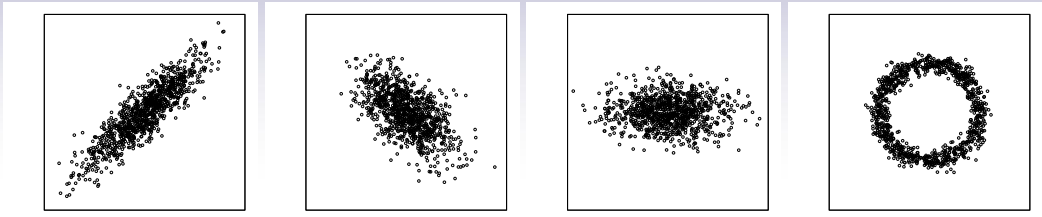
$$C_{ij} = E[x_i x_j] - \mu_i \mu_j = E[(x_i - \mu_i)(x_j - \mu_j)] \ , \ \rho_{ij} = \frac{C_{ij}}{\sigma_i \sigma_j} \ .$$

Los coeficientes de correlación lineal tienen valores en el rango $-1 \leq \rho_{ij} \leq 1$, e indican la tendencia dominante de densidad en el patrón $(x_i; x_j)$: se habla de correlaciones positivas y negativas (o anti-correlaciones).
Para variables aleatorias X_i, X_j independientes, es decir con $P(x_i, x_j) = P_{X_i}(x_i) P_{X_j}(x_j)$, se tiene

$$E[x_i x_j] = \int\int dx_i dx_j P(x_i, x_j) x_i x_j = \mu_i \mu_j \ , \longrightarrow \ \rho_{ij} = 0 \ .$$

(pero la conversa no es necesariamente cierta, p.e. ejemplo en la lámina siguiente)

Ejemplos de correlaciones



De izquierda a derecha :

- $\rho = +0.9$,
- $\rho = -0.5$;
- $\rho = 0$, para variables independientes ,
- variables fuertemente correlacionadas con un patrón no lineal de correlación que "conspira" para arrojar una correlación lineal nula, $\rho = 0$.

CAPITULO III

CARACTERIZACION DE FORMAS

ESTIMACION DE PARAMETROS

PROPAGACION DE ERRORES

Caracterización de la forma de una PDF (I)

Estamos suponiendo que nuestros objetos de estudio de procesos aleatorios, que se manifiestan bajo forma de realizaciones aleatorias a partir de una PDF subyacente

- En la práctica, la verdadera dependencia funcional de una PDF es a menudo desconocida
- La información sobre su forma solamente puede extraerse a partir de una muestra de talla finita (digamos que contiene N eventos), es decir suponemos que la muestra disponible es una realización aleatoria a partir de una PDF desconocida.
- Si consideramos que esa PDF subyacente is de tipo paramétrico, la *caracterización de su forma* es un procedimiento para estimar los valores numéricos de sus parámetros, partiendo de una hipótesis "razonable" sobre la dependencia funcional sobre sus variables.
- Ahora, solamente un número finito de valores de expectación independientes pueden extraerse de una muestra de talla finita.
- No existe una receta única para la elección de los parámetros a ser estimados, con lo que el proceso es intrínsecamente incompleto.
- Se puede sin embargo confirmar en la práctica que el proceso de caracterización de forma es bastante poderoso, si los parámetros seleccionados proveen información útil y complementaria.

Ilustramos estas consideraciones con un ejemplo unidimensional para una variable aleatoria única x.

Ilustramos las consideraciones anteriores con un ejemplo unidimensional para una variable aleatoria única x.
Consideremos el promedio empírico $\overline{x}$, definido como

$$\overline{x} \;=\; \frac{1}{N}\sum_{i=1}^{N} x_i \;.$$

Mostraremos más adelante que $\overline{x}$ es un buen *estimador* de la media μ de la PDF subyacente $P(x)$.
De modo análogo, la media cuadrática RMS (del inglés "root-mean-squared"), definida como

$$\mathrm{RMS}^2 \;=\; \overline{x^2} - (\overline{x})^2 \;\text{, con } \overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \;,$$

es un estimador razonable de la varianza σ^2 (veremos más adelante una mejor definición).
En términos intuitivos, el promedio y el RMS reflejan información útil y complementaria sobre la "localización" y la "dispersión" de la región de x mayor densidad de eventos, y esta región debe corresponder de manera aproximada a los intervalos de x donde la PDF tiene valores más grandes.

- obviamente, en un caso general esos dos parámetros son insuficientes para caracterizar una PDF más genérica, que require un procedimiento más sistemático.

Caracterización de la forma de una PDF (III)

Para un procedimiento más sistemático, partamos de la definición de los *momentos* de la distribución μ,

$$\mu_n = \int dx \mathcal{P}(x) x^n \ ,$$

que pueden ser usados para caraterizar la forma de la PDF. Es usual definir una variable reescalada $x' = (x - \mu)/\sigma$, cuyo efecto es de desplazar la PDF $\mathcal{P}(x')$ para que tenga promedio nulo, y reescalarla para tener varianza unitaria. De esta manera, los dos primeros momentos son $\mu_1 = 0$ y $\mu_2 = 1$.
En principio, mientras mayor el número de momentos μ_j sean estimados, más detaillada será la caracterización de la forma de la PDF (pero una muestra finita solamente permite medir un número finito de momentos).
Los momentos 3 y 4 tienen nombres específicos, y sus valores son sencillos de interpretar en términos de la forma:

- el tercer momento es llamado oblicuidad (o *skewness* en inglés)
  - una distribución simétrica tiene skewness nula,
  - un valor negativo (positivo) indica una "anchura" mayor a la izquierda (derecha) de su media.
- el cuarto momento es llamado *kurtosis*
  - cantidad definida positiva, relacionada con cuán "picante" es la distribución
  - un valor pequeño indica un pico estrecho y "colas" de largo alcance: es una distribución leptokúrtica
  - un valor grande indica un pico central ancho y colas poco prominentes: es una distribución platykúrtica

Caracterización de la forma de una PDF (III)

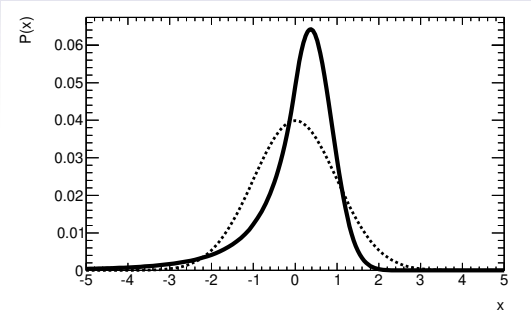
Para un procedimiento más sistemático, transformamos la x-dependencia de la PDF $P(x)$ en una k-dependencia de la *función característica* $C[k]$, definida como

$$C[k] \;=\; E\left[e^{ik\frac{x-\mu}{\sigma}}\right] \;=\; \sum_j \frac{(ik)^j}{j!}\mu_j \;.$$

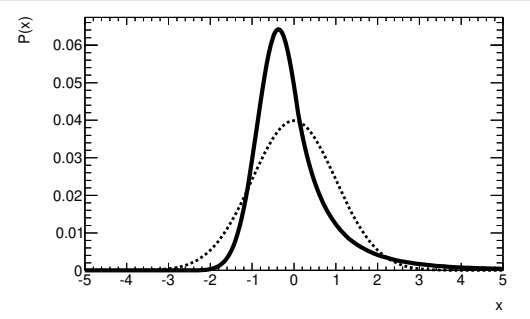
Como se puede notar, la función característica es la transformación de Fourier de la PDF. Los coeficientes μ_j de la expansión are se llaman *momentos reducidos*; por construcción, los primeros momentos son $\mu_1 = 0$ y $\mu_2 = 1$; en términos de la variable reescalada $x' = (x-\mu)/\sigma$, la PDF fue desplazada para tener promedio nulo, y escalada para tener varianza unitaria.
En principio, mientras mayor el número de momentos μ_j sean estimados, más detaillada será la caracterización de la forma de la PDF (pero una muestra finita solamente permite medir un número finito de momentos).
Los momentos 3 y 4 tienen nombres específicos, y sus valores son sencillos de interpretar en términos de la forma:

- el tercer momento es llamado oblicuidad (o *skewness* en inglés)
  - una distribución simétrica tiene skewness nula,
  - un valor negativo (positivo) indica una "anchura" mayor a la izquierda (derecha) de su media.
- el cuarto momento es llamado *kurtosis*
  - cantidad definida positiva, relacionada con cuán "picante" es la distribución
  - un valor pequeño indica un pico estrecho y "colas" de largo alcance: es una distribución leptokúrtica
  - un valor grande indica un pico central ancho y colas poco prominentes: es una distribución platykúrtica

(la curva punteada es una Gaussiana reducida, con $\mu_1 = 0$, $\mu_2 = 1$, y $\mu_i = 0\ \forall i > 2$)

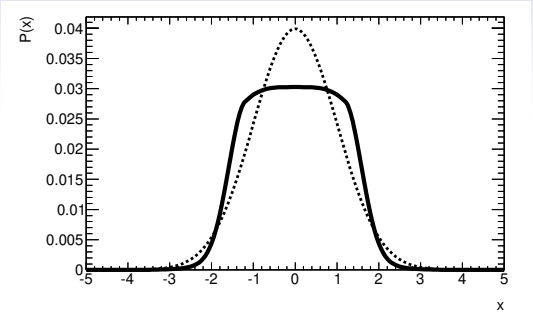


Cola "no-Gaussiana" a la izquierda del pico: skewness positiva, $\mu_3 > 0$

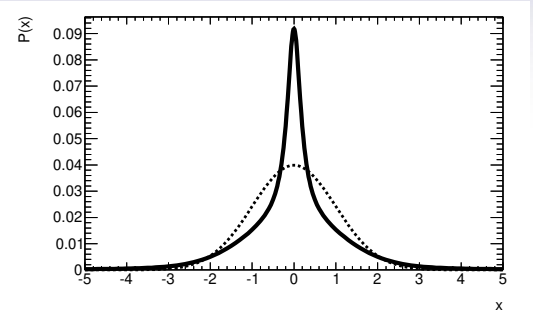


Cola "no-Gaussiana" a la derecha del pico: skewness negativa, $\mu_3 < 0$

(la curva punteada es una Gaussiana reducida, con $\mu_1 = 0$, $\mu_2 = 1$, y $\mu_i = 0\ \forall i > 2$)



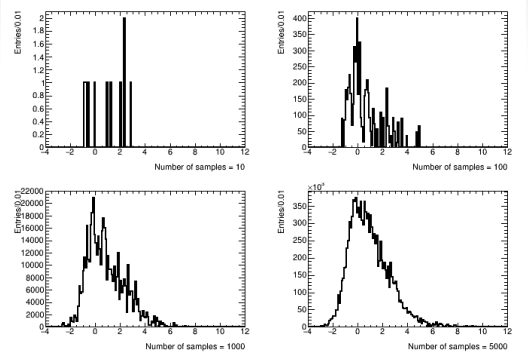
Pico ancho, colas de corto alcance :
kurtosis pequeña (“platykurtic”)

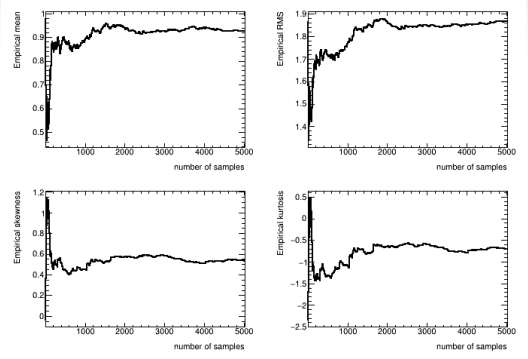


Pico estrecho, colas de largo alcance :
kurtosis grande (“leptokurtic”)

Un ejemplo : a partir de una misma PDF, se realizan $N = 10, 100, 1000, 5000$ realizaciones de una variable x:

- se grafican las distribuciones de frecuencia correspondientes (figuras a la izquierda),
- se estiman los valores de los 4 primeros momentos : media, RMS, skewness y kurtosis (figuras a la derecha).





La caracterización de la forma de una PDF a través de una estimación sequencial de parámetros de forma nos permitió introducir de manera cualitativa al concepto de estimación de parámetros (también llamado en inglés "point estimation"). Una receta más general sería la siguiente :
Consideremos una PDF n-dimensional, k-paramétrica,

$$P(\ x_1, x_2, \ldots, x_n\ ;\ \theta_1, \theta_2, \ldots, \theta_k\ )\ ,$$

para la cual queremos estimar los valores $\theta_1, \ldots, \theta_k$ a partir de una muestra de talla finita, utilizando un conjunto de estimadores $\hat{\theta_1}, \ldots, \hat{\theta_k}$.
Esos estimadores son también variables aleatorias, con sus propias medias y varianzas: sus valores diferirían al ser estimados sobre otras muestras.
(Nota: estamos implícitamente suponiendo que las PDFs de los estimadores son Gaussianas, totalmente caracterizadas por sus media y varianza)
Esos estimadores deben satisfacer dos propiedades clave:

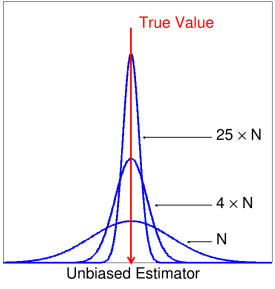
- *ser consistente* : asegura que, en el límite de una muestra de talla infinita, el estimador converge al verdadero valor del parámetro;
- *ser no sesgado* : la ausencia de sesgo asegura que el valor de expectación del estimador es el verdadero valor del parámetro, para toda talla de la muestra.

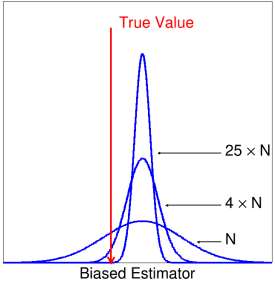
Un estimador sesgado pero consistente (también llamado asintóticamente no-sesgado) es tal que el sesgo disminuye al aumentar la talla de la muestra.

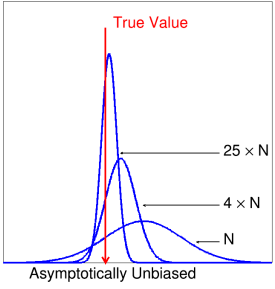
Otros criterios son útiles para caracterizar la calidad de los estimadores; por ejemplo

- eficiencia: un estimador de pequeña varianza es más eficiente que uno de mayor varianza ;
- robusteza: este criterio describe la "sensibilidad" del estimador a incertidumbres en la forma de la PDF. Por ejemplo, el promedio es robusto contra incertidumbres sobre los momentos de orden par, pero es menos robusto contra incertidumbres en los momentos de orden impar.

Nota : estos criterios son en ocasiones mutuamente contradictorios; por razones prácticas, puede ser preferible tener un estimador eficiente pero sesgado, a uno no sesgado pero de pobre convergencia.







El promedio empírico $\overline{x}$ es un estimador convergente, no sesgado de la media μ de la PDF subyacente: $\hat{\mu} = \overline{x}$. Esto se demuestra facilmente, evaluando el valor de expectación y la varianza de $\overline{x}$:

$$E\left[\overline{x}\right] = \frac{1}{N}\sum_{i=1}^{N} E\left[x\right] = \mu \ ,$$

$$V\left[\overline{x}\right] = E\left[(\overline{x}-\mu)^2\right] = \frac{\sigma^2}{N} \ .$$

Al contrario, el RMS empírico de una muestra es un estimador sesgado (aunque asintóticamente no-sesgado) de la varianza σ^2. Esto se demuestra facilmente también, reescribiendo su cuadrado en terminos de la media:

$$\mathrm{RMS}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 = \left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\right] - (\overline{x}-\mu)^2 \ ,$$

de manera que su valor de expectación es

$$E\left[\mathrm{RMS}^2\right] = \sigma^2 - V\left[\overline{x}\right] = \frac{N-1}{N}\sigma^2 \ ,$$

que si bien converge a la verdadera varianza σ^2 en el límite $N \to \infty$, subestima sistemáticamente su valor para muestras de talla finita.

Pero es inmediato definir un estimador modificado

$$\frac{N}{N-1}\mathrm{RMS}^2 \;=\; \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2 \;\;,$$

que es, para muestras de talla finita, un estimador no sesgado de la varianza.
En resumen: para una PDF desconocida, tenemos estimadores consistentes y no sesgados de su media μ y su varianza σ^2, que pueden ser extraídos de muestras de talla finita:

$$\hat{\mu} = E\left[\overline{x}\right] = \frac{1}{N}\sum_{i=1}^{N} x_i \;, \hat{\sigma}^2 = \frac{N}{N-1}E\left[\mathrm{RMS}^2\right] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\hat{\mu})^2 \;\;,$$

que junto con sus varianzas

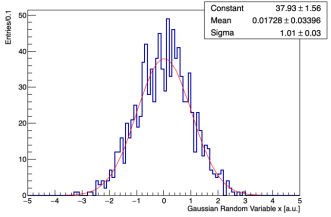
$$V\left[\overline{x}\right] = \frac{\hat{\sigma}^2}{N} \;,\; V\left[\mathrm{RMS}^2\right] = \frac{\hat{\sigma}^2}{2N} \;,$$

determinan completamente los estimadores de la media μ y la varianza σ de una PDF.

El resultado de la estimación de μ y σ se reporta usualmente bajo la forma del estimador y su "incertidumbre" :

$$\begin{aligned} \hat{\mu} &= \left(\overline{x} \pm \frac{\mathrm{RMS}}{\sqrt{N}} \right) , \\ \hat{\sigma} &= \left(\frac{N}{N-1} \mathrm{RMS} \pm \frac{\mathrm{RMS}}{\sqrt{2N}} \right) \simeq \left(\mathrm{RMS} \pm \frac{\mathrm{RMS}}{\sqrt{2N}} \right) . \end{aligned}$$

- Los factores $1/N$ y $1/(N-1)$ para $\hat{\mu}$ y $\hat{\sigma}^2$ se entienden intuitivamente: el promedio empírico puede medirse incluso en la muestra más pequeña posible de un solo evento, mientras que al menos dos eventos son necesarios para estimar la dispersión empírica de una muestra.



Ejercicio : completar los detalles relacionando la media y la varianza con sus estimadores empíricos.

El ejemplo clásico anterior trataba de una única variable aleatoria.
En presencia de múltiples variables aleatorias $\vec{x} = \{x_1, \ldots, x_n\}$, la generalización del resultado anterior lleva a definir la covarianza empírica, cuyos elementos $\hat{C}_{ab}$ son estimados en una muestra de N eventos de la manera siguiente:

$$\hat{C}_{ab} \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(x_{a,i} - \hat{\mu_a}\right)\left(x_{b,i} - \hat{\mu_b}\right) \;\; .$$

(los índices a, b recorren la lista de variables aleatorias, $1 \leq a, b \leq n$). (Nota: para N pequeños, se debe corregir un sesgo en este estimador, c.f. el ejemplo del RMS)
Suponiendo que la verdadera covarianza es conocida, la varianza de una función arbitraria $f(\vec{x})$ de las variables aleatorias se evalúa a partir de la expansión de Taylor alrededor de las medias de sus parámetros $\hat{\vec{\mu}}$ según

$$f\left(\vec{x}\right) \;=\; f\left(\hat{\vec{\mu}}\right) + \sum_{a=1}^{n}\left.\frac{\partial f}{\partial x_a}\right|_{\vec{x}=\hat{\vec{\mu}}}\left(x_a - \hat{\mu}_a\right) \;\; ,$$

en otras palabras, $E\left[f(\vec{x})\right] \;\simeq\; f(\hat{\vec{\mu}})$.

De manera similar,

$$E\left[f^2(\vec{x})\right] \simeq f^2(\hat{\vec{\mu}}) + \sum_{a,b=1}^{n} \frac{\partial f}{\partial x_a}\frac{\partial f}{\partial x_b}\bigg|_{\vec{x}=\hat{\vec{\mu}}} \hat{C}_{ab} \ ,$$

con lo que la varianza de f se estima como

$$\hat{\sigma}_f^2 \simeq \sum_{a,b=1}^{n} \frac{\partial f}{\partial x_a}\frac{\partial f}{\partial x_b}\bigg|_{\vec{x}=\hat{\vec{\mu}}} \hat{C}_{ab} \ .$$

Esta expresión, llamada *fórmula de propagación de incertidumbres*, estima la varianza de una función genérica $f(\vec{x})$ a partir de los estimadores de las medias y covarianzas de los argumentos de la función.
(Nota: estamos implícitamente suponiendo que la PDF es una Gaussiana multidimensional, totalmente caracterizada por sus medias y la matriz de covarianza)
Ejemplos particulares de propagación de incertidumbres:

- cuando todas las variables aleatorias $\{x_a\}$ son no-correlacionadas, la matriz de covarianza es diagonal, $C_{ab} = \sigma_a^2 \delta_{ab}$ y la covarianza de $f(\vec{x})$ se reduce a

$$\hat{\sigma}_f^2 \simeq \sum_{a=1}^{n} \left(\frac{\partial f}{\partial x_a}\right)^2\bigg|_{\vec{x}=\hat{\vec{\mu}}} \hat{\sigma}_a^2 \ .$$

- para la suma de dos variables aleatorias $S = x_1 + x_2$, la varianza es

$$\sigma_S^2 \;=\; \sigma_1^2 + \sigma_2^2 + 2C_{12} \;=\; \sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2\rho_{12},$$

- la generalización a más de dos variables es:

$$\sigma_S^2 \;=\; \sum_{a,b} \sigma_a \sigma_b \rho_{ab} \;.$$

En ausencia de correlaciones, se dice que *los errores absolutos se suman "en cuadratura"*:

$$\sigma_S \;=\; \sigma_1 \oplus \sigma_2 \;= \sqrt{\sigma_1^2 + \sigma_2^2}\,,$$

y si la correlación vale 1, el error en la suma es la suma directa de los errores de cada término.
Ejemplos :

- para errores absolutos similares, $x_1 = (6.0 \pm 0.3)$ (5% relativo) y $x_2 = (2.0 \pm 0.4)$ (20% relativo), la suma $S = x_1 + x_2$ es $S = (8.0 \pm 0.5)$ (contribuciones similares de ambos errores absolutos) ;
- para errores absolutos muy diferentes, $x_1 = (6.0 \pm 0.9)$ (15% relativo) y $x_2 = (2.5 \pm 0.1)$ (4% relativo), la suma es $S = (8.5 \pm 0.9)$ (el error absoluto total está dominado por el mayor).

- para el producto de dos variables aleatorias $P = x_1 x_2$, la varianza relativa es

$$\left(\frac{\sigma_P}{P}\right)^2 = \left(\frac{\sigma_1}{x_1}\right)^2 + \left(\frac{\sigma_2}{x_2}\right)^2 + 2\frac{\sigma_1}{x_1}\frac{\sigma_2}{x_2}\rho_{12} \ ,$$

- la generalización a más de dos variables:

$$\left(\frac{\sigma_P}{P}\right)^2 = \sum_{a,b} \frac{\sigma_a}{x_a}\frac{\sigma_b}{x_b}\rho_{ab} \ .$$

En ausencia de correlaciones, se dice que *los errores relativos se suman en cuadratura*:

$$\frac{\sigma_P}{P} = \frac{\sigma_1}{x_1} \oplus \frac{\sigma_2}{x_2} \ ,$$

y si la correlación vale 1, el error relativo sobre el producto es la suma directa de los errores relativos de cada término.
Ejemplos :

- para errores relativos similares, $x_1 = (2.0 \pm 0.3)$ (15% relativo) y $x_2 = (5.0 \pm 1.0)$ (20% relativo) el producto $P = x_1 x_2$ es $P = (10.0 \pm 2.5)$ (25% relativo);
- para errores relativos muy diferentes, $x_1 = (10.0 \pm 3.0)$ (30% relativo) y $x_2 = (2.5 \pm 0.1)$ (4% relativo) el producto es $P = (25.0 \pm 6.6)$ (30.2% relativo).

- para una función genérica en ley de potencia, $Z = x_1^{n_1} x_2^{n_2} \ldots$, si todas las variables son no-correlacionadas, el error relativo es

$$\frac{\sigma_Z}{Z} = n_1 \frac{\sigma_1}{x_1} \oplus n_2 \frac{\sigma_2}{x_2} \oplus \ldots$$

En otras palabras: en la suma en cuadratura, los errores relativos son ponderados por los exponentes. Por tanto los términos con exponentes importantes (cuadrados, cubos, etc...) dominarán el *presupuesto de error*, mientras que términos con raíces (cuadradas, cúbicas, etc...) tendrán un impacto subdominante.
Ejemplo :

- para la función $Z = x_1^2 / \sqrt[3]{x_2}$, si tenemos $x_1 = (8.0 \pm 1.6)$ (20% relativo) y $x_2 = (64.0 \pm 19.2)$ (30% relativo), 3l resultado es $Z = (16.0 \pm 6.6)$ (41.2% relativo) : la contribución dominante al error proviene de x_1, a pesar de ser el término con menor error relativo.

En muchas situaciones importantes, las correlaciones entre variables tienen que ser tomadas en cuenta. Ilustramos ese caso con un ejemplo es la medida de la *eficiencia*, o fracción de eventos que pasan un criterio de selección :

$$\varepsilon = \frac{N_{\text{pass}}}{N} = \frac{N_{\text{pass}}}{N_{\text{fail}} + N_{\text{pass}}} \text{ , con } N = N_{\text{pass}} + N_{\text{fail}} \ .$$

Una equivocación usual es creer que N y N_{pass} están descorrelacionados, y que por tanto la incertidumbre en la eficiencia $\sigma(\varepsilon)$ sería la propagación simple de las incertidumbres $\sigma(N_{\text{pass}})$ y $\sigma(N)$. Por el contrario, los eventos que pasan/fallan el criterio, N_{pass} y N_{fail} sí son variables independientes, pero la eficiencia no es una expresión en términos de productos de leyes de potencia. La propagación de incertidumbres es por tanto :

$$\sigma(\varepsilon) = \left(\left| \frac{\partial \varepsilon}{\partial N_{\text{pass}}} \right| \sigma(N_{\text{pass}}) \right) \oplus \left(\left| \frac{\partial \varepsilon}{\partial N_{\text{fail}}} \right| \sigma(N_{\text{fail}}) \right) \ ,$$

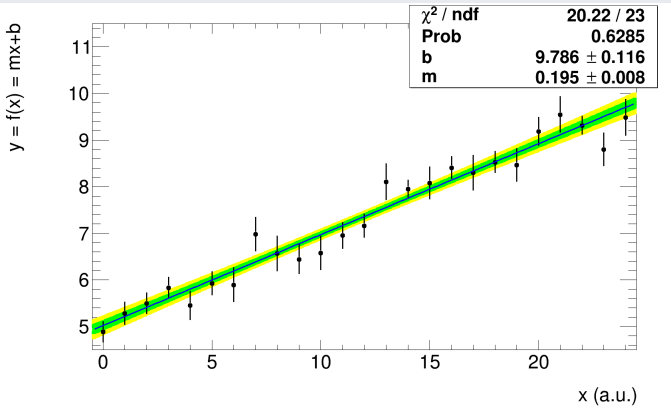
que en el caso particular $\sigma(N_{\text{pass}}) = \sqrt{N_{\text{pass}}}$ y $\sigma(N_{\text{pass}}) = \sqrt{N_{\text{pass}}}$ (aproximación válida en un límite que estudiaremos más adelante) se reduce a una forma sencilla :

$$\sigma(\varepsilon) = \sqrt{\frac{\varepsilon(1-\varepsilon)}{N}} \ .$$

Nota : en física de partículas usamos *eficiencia* y *pureza*, mientras que en otras disciplinas se usa *sensibilidad* y *especificidad*, conceptos definidos en términos de fracciones de *falsos positivos* y *falsos negativos*.
Ejercicio : reproducir la expresión del error de la eficiencia, y derivar las expresiones para la pureza, la sensibilidad y la especificidad.

Propagación de incertidumbres (VI)

Otra situación importante en la que se debe tomar en cuenta las posibles correlaciones entre variables : los ajustes (fits en inglés). El resultado de un ajuste nos da, además de los valores de los parámetros del ajuste, sus incertidumbres y las correlaciones entre parámetros. Un ejemplo sencillo : ajustar a una recta 25 puntos que exhiben una tendencia creciente.



Un ajuste a la función $f(x) = mx + b$ resulta en una correlación *negativa* entre m y b...
... mientras que un ajuste a la función $f(x) = m(x - 25) + b$ (un simple cambio del origen de la escala horizontal) da el mismo resultado para la función (con b reescalado claro) pero una correlación *positiva* entre m y b
... y correlaciones casi nulas si se usa $f(x) = m(x - 12) + b$
Eso se entiende perfectamente en términos de la "palanca" del ajuste, y explica por qué los *intervalos de confianza* del ajuste (bandas verdes y amarillas) son más amplios en los bordes que en el centro (diferencia entre interpolar y extrapolar)

CAPITULO IV

FUNCIONES DE DENSIDAD DE PROBABILIDAD DE USO COMUN

Distribución	Uso(s) en altas energías	Otro(s) nombre(s)
Binomial	Tasa de decaimiento, eficiencias	Bernouilli
Poisson	Conteo de eventos	"ley de eventos raros"
Uniforme	Integración Monte-Carlo	
Exponencial	Vida media, tiempos de relajación	
Gaussiana	Resolución	Normal
Breit-Wigner	Resonancia	Cauchy, Lorentz
χ^2	"Goodness-of-fit"	"bondad de ajuste"

Lista no exhaustiva; otras de uso frecuente incluyen la distribución de Student, Galton o lognormal...

Ejemplo de una distribución discreta: Binomial (I)

Escenario con dos únicas realizaciones posibles: "éxito" y "fracaso", y con una probabilidad fija p de "éxito". Nos interesamos solamente en el número k de "éxitos" después de n intentos $0 \leq k \leq n$; (suponemos que la secuencia de intentos es irrelevante)
El número entero k sigue la distribución binomial $P(k; n, p)$:

$$P_{\text{binomial}}\,(k; n, p) \;=\; \frac{n!}{k!(n-k)!} p^k\,(1-p)^{n-k} \;\;,$$

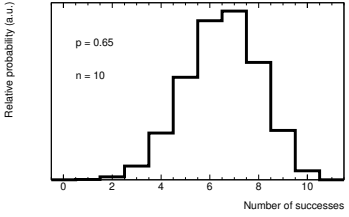
donde k es la variable aleatoria, mientras que n y p son parámetros.
Ejemplo típico: el número de eventos en una sub-categoría específica de eventos (p.e. una tasa de decaimiento)

Ejercicios :

- Verificar que la binomial está normalizada
- mostrar que la media y la varianza de una distr. binomial son

$$\begin{aligned} E\,[k] &= \sum_{k=0}^{n} k P\,(k; n, p) \;=\; np\;, \\ V\,[k] &= np(1-p)\;. \end{aligned}$$

- Sean dos binomiales con números de intentos n_1 y n_2 y de misma probabilidad p. Mostrar que la suma es también una binomial.



Teorema de la varianza de suma de binomiales:
La varianza de la suma de variables aleatorias que siguen distribuciones binomiales con probabilidades p_i diferentes,

$$k = \sum_i k_i \ ,$$

viene dada por

$$V\left[k\right] \quad = \quad n\overline{p}(1-\overline{p}) - ns^2 \ , \text{con } s^2 = \frac{1}{n}\sum_i (p_i - \overline{p})^2 \ ,$$

(donde $\overline{p}$ es el promedio de las probabilidades p_i), y es por tanto inferior o igual a la varianza de una variable binomial de probabilidad $\overline{p}$.
Ejercicio:
Verificar numéricamente la validez de este teorema, estimando con una simulación sencilla la varianza de la suma de dos variables aleatorias binomiales con probabilidades p_1 y p_2, y comparándola con la varianza de una variable aleatoria binomial con $p = (p_1 + p_2)/2$. Elegir varios valores de p_1 y p_2 que ilustren casos extremos del teorema.

Ejemplo de una distribución discreta: Poisson

Para la distribución binomial en el límite $n \to \infty$, $p \to 0$ (con $\lambda = np$ finito y no-nulo) la variable aleatoria k sigue la llamada distribución de Poisson $P(k;\lambda)$,

$$P_{\text{Poisson}}(k;\lambda) \;=\; \frac{\lambda^k e^{-\lambda}}{k!}\;,$$

que tiene λ por único parámetro. Para esta distribución de Poisson, la media y la varianza tienen el mismo valor :

$$E[k] \;=\; V[k] \;=\; \lambda\;.$$

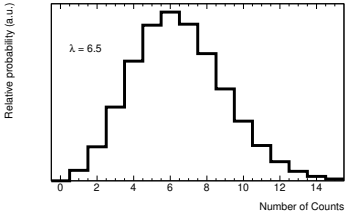
Esta distribución es también llamada "ley de los eventos raros" (debido al límite $p \to 0$), describe el número de observaciones en condiciones fijas si la tasa de ocurrencia es constante. Ejemplo: el número de decaimientos en un intervalo de tiempo fijo, para una fuente de actividad constante.

Ejercicios :

- mostrar que la media y la varianza de una Poisson son

$$E[k] \;=\; V[k] \;=\; \lambda\;.$$

- Sean dos distribuciones de Poisson con parámetros λ_1 y λ_2. Mostrar que la distribución de la suma es también una Poisson, con parámetro $\lambda = \lambda_1 + \lambda_2$.



Ejemplo de una distribución continua: Uniforme

Una variable aleatoria continua x, con densidad de probabilidad $P(x;a,b)$ constante y no-nula únicamente dentro de un intervalo finito $[a,b]$:

$$P_{\text{uniform}}(x;a,b) = \begin{cases} \frac{1}{b-a} & , \quad a \leq x \leq b \ , \\ 0 & , \quad x < a \text{ o } x > b \ . \end{cases}$$

Ejercicio : mostrar que para esta distribución uniforme, la media y la varianza son

$$E[x] = \frac{a+b}{2} \ , \ V[x] = \frac{(b-a)^2}{12} \ .$$

Uso cumún de la distribución uniforme : generación de distribuciones aleatorias

- "acepto-rechazo" para simular un proceso de probabilidad $p(x)$: utilizar dos variables aleatorias: x uniforme en el intervalo de interés, y x_0, uniforme en el intervalo $[0,1]$. Aceptar el evento si $x_0 < p(x)$. Intuitivo y sencillo, pero usualmente poco eficaz.
- Transformada inversa: si se conoce la CDF de la distribución deseada, a partir de x uniforme en el intervalo $[0,1]$, la variable aleatoria $z = CDF(x)$ sigue la distribución correspondiente.
- Transformación de Box-Muller: a partir de dos variables aleatorias x_1, x_2 uniformes en $[0,1]$ e independientes, las variables z_1, z_2 dadas por

  $$z_1 = \sqrt{-2\ln x_1}\cos(2\pi x_2) \ , \ z_2 = \sqrt{-2\ln x_1}\sin(2\pi x_2) \ ,$$

  son independientes y siguen cada distribuciones Gaussianas, ambas de media nula y varianza unitaria.

Ejemplo de una distribución continua: Exponencial

Una variable aleatoria que sigue una densidad de probabilidad $P(x;\xi)$ dada por

$$P_{\text{exponential}}(x;\xi) \;=\; \begin{cases} \frac{1}{\xi} e^{-x/\xi} & , \quad x \geq 0 \;, \\ 0 & , \quad x < 0 \;, \end{cases}$$

y tiene por media y varianza

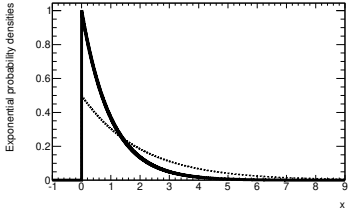
$$E[x] = \xi \;,\; V[x] = \xi^2 \;.$$

Ejercicio: evaluar la media y la varianza de una distribución exponencial truncada, que es no-nula solamente en un intervalo finito $a \leq x \leq b$. (Nota: verificar que la PDF que usen está correctamente normalizada)
La aplicación más común de esta distribución exponencial es la descripción de fenómenos independientes que se realizan a una tasa constante, como el tiempo de vida de una partícula inestable.

En vista del carácter auto-similar de la función exponencial:

$$P(t - t_0 | t > t_0) \;=\; P(t) \;,$$

se dice que esta distribución es "sin memoria".
La figura muestra dos PDF exponenciales, con parámetros $\xi = 1$ (línea sólida) y $\xi = 2$ (punteada).
Ejercicio: Obtener numéricamente una realización aleatoria de una distribución exponencial usando el método de la transformada inversa. Determinar el promedio y el RMS empíricos resultantes.

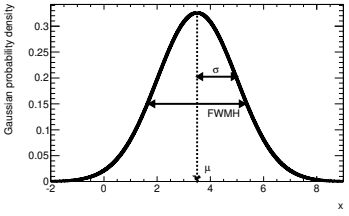


$$P_{\text{Gauss}}(x;\mu,\sigma) \;=\; \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \;.$$

Para la distribución Normal (o Gaussiana), su media y varianza vienen dadas por

$$E[x] \;=\; \mu \;,\; V[x] \;=\; \sigma^2 \;.$$

Usamos deliberadamente los símbolos μ y σ tanto para los parámetros de la PDF Gaussiana como para su media y su varianza: la Gaussiana está caracterizada unívocamente por sus dos primeros momentos; $\mu_i = 0 \;\forall\; i > 2$.



La dispersión de una distribución a un solo máximo es en ocasiones caracterizada en términos de su *anchura a media altura* o FWHM (full width at half-maximum); para Gaussianas, la relación con la varianza es as $\text{FWHM} = 2\sqrt{2\ln 2}\sigma \simeq 2.35\sigma$.

La Gaussiana es también la distribución límite para las binomiales y Poisson, en los límites a gran n y gran λ, respectivamente:

$$P_{\text{binomial}}(k; n\to\infty, p) \quad\to\quad P_{\text{Gauss}}(k; np, np(1-p)) \;,$$

$$P_{\text{Poisson}}(k; \lambda\to\infty) \quad\to\quad P_{\text{Gauss}}\left(k; \lambda, \sqrt{\lambda}\right) \;.$$

Ejercicio:
Verificar que los momentos superiores de la Gaussiana son todos nulos.

Nota :, una corrección de continuidad es requerida: el rango de la Gaussiana se extiende a valores negativos, mientras que Binomial y Poisson están solamente definidas en el rango positivo.

La primacía de la Gaussiana en términos de su relevancia conceptual y sus aplicaciones prácticas proviene en gran parte del *teorema del límite central* : si tenemos n variables aleatorias independendientes $\vec{x} = \{x_1, x_2, \ldots, x_n\}$, cada una con medias y varianzas μ_i y σ_i^2, la resultante $S(\vec{x})$ de sumarlas todas

$$S \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_i}{\sigma_i} \;,$$

sigue una distribución que, en el límite de gran n, tiende a una distribución normal reducida.
(El caso particular $\mu = 0$, $\sigma = 1$ es llamado en ocasiones "normal reducida").
Por ello, una gran variedad de procesos, sujetos a múltiples fuentes independientes de incertidumbre, pueden describirse en buena aproximación con distribuciones Gaussianas, sin necesidad de conocer los detalles específicos de cada fuente de incertidumbre.
(ejemplo gráfico en la lámina siguiente, ilustrando la convergencia rápida del teorema del valor central para suma de distribuciones uniformes)
Ejercicio : verificar con una aplicación numérica la validez del teorema para otras distribuciones, por ejemplo sumas de exponenciales.

La distribución continua de mayor ubicuidad: la Gaussiana (III)

Teorema del valor central :

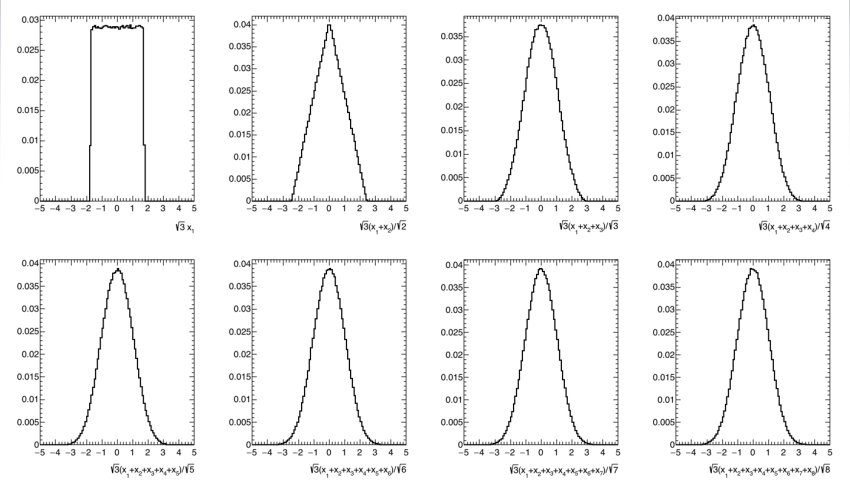


Ilustración: sumas de $2, \cdots 8$ variables aleatorias, todas provenientes de distribuciones uniformes.

Ejemplo de una distribución continua: probabilidad de χ^2

Una distribución importante es llamada "probabilidad de χ^2, definida como sigue:

$$P_{\chi^2}(x;n) = \begin{cases} \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(\frac{n}{2})} & , \quad x \geq 0\ , n \text{ entero} \\ 0 & , \quad x < 0 \text{ o } n \text{ no} - \text{entero}\ , \end{cases}$$

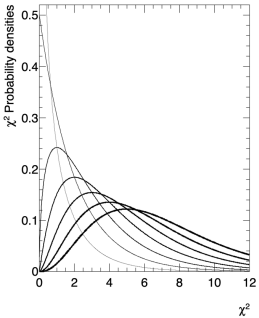
con único parámetro n, y donde $\Gamma(n/2)$ es la función Gamma. Su media y su varianza vienen dadas por

$$E[x] = n\ ,\ V[x] = 2n\ .$$

Una propiedad interesante de esta distribución es que sus momentos μ_k $(\forall k > 2)$ se anulan para grandes valores de n; en particular su skewness y kurtosis vienen dadas por:

$$\mu_3 = \sqrt{\frac{8}{n}}\ ,\ \mu_4 = \frac{12}{n}\ ;$$

en otras palabras para n grande, $P_{\chi^2}(x;n)$ se acerca a una Gaussiana de media $\mu \to n$ y varianza $\sigma^2 \to 2n$.
La forma de la distribución de P_{χ^2} depende de n, como se ve en la gráfica que representa P_{χ^2} para $n = 1, \cdots, 7$.
El parámetro n es también llamado "número de grados de libertad", y como veremos se refiere al comportamiento de los ajustes por el método de los cuadrados mínimos: cuando n_d puntos son utilizados para estimar n_p parámetros, el número correspondiente de grados de libertad es $n_d - n_p$.

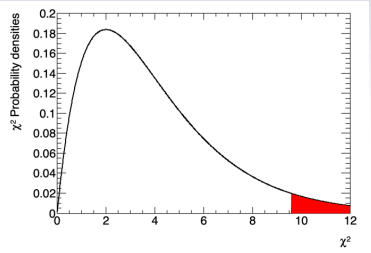


Ejemplo de una distribución continua: χ^2

Definición importante:

Si tenemos n variables aleatorias x_i, cada una de ellas distribuidas siguiendo normales de media μ_i y varianza σ_i^2, definimos una nueva variable aleatoria $\chi^2(n)$:

$$\chi^2(n) = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 .$$



El parámetro n es también llamado "número de grados de libertad", y se refiere al comportamiento de los ajustes por el método de los cuadrados mínimos: cuando n_d puntos son utilizados para estimar n_p parámetros, el número correspondiente de grados de libertad es $n_d - n_p$.

Si el modelo utilizado para ajustar es adecuado, los valores de los χ^2 obtenidos deben seguir la distribución P_{χ^2}.

Este criterio es llamado *goodness-of-fit* o "bondad del ajuste":

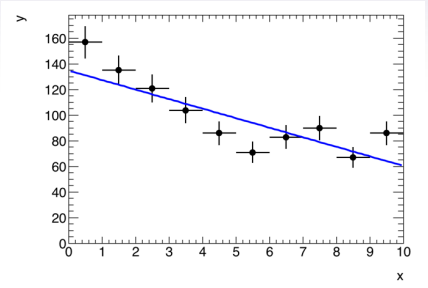
Definimos la "p-value" o "probabilidad de χ^2" como la probabilidad de obtener un valor de χ^2 más grande que el observado; un χ^2 muy superior al valor de expectación (n), es decir una p-value muy pequeña, se interpreta como indicador de la pobre calidad del ajuste.

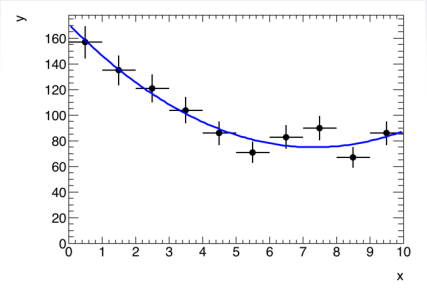
En el ejemplo aquí graficado, un ajuste con 4 grados de libertad arrojó un $\chi^2 = 9.1$. La p-value se obtiene integrando

$$\mathrm{p-value} = \int_{\chi^2=9.1}^{\infty} dx P_{\chi^2}(x;4) = 0.061 .$$

Un ejemplo sencillo ("regresión") de un ajuste de χ^2

Tomemos una serie de 10 medidas de observables x e y, que muestran una clara anticorrelación. Suponemos que x es la variable *independiente*, y es la dependiente. Suponemos también que solamente las incertidumbres sobre y son pertinentes en este ejemplo.



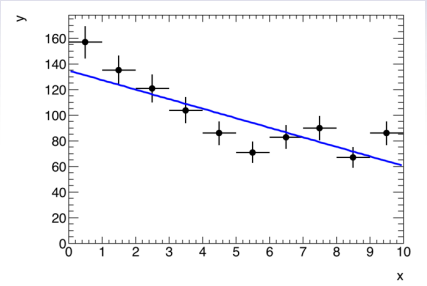


Discusión: ¿a qué corresponden las dos curvas azules?

Ejercicio: reproducir los ajustes mostrados en la lámina siguiente. Los valores precisos son: $\overline{y} = [157; 135; 121; 104; 86; 71; 83; 90; 67; 86]$ y los errores en y son la raíz cuadrada de su valor.

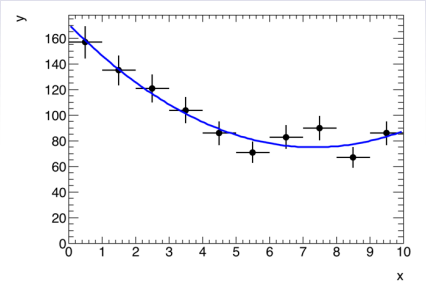
Un ejemplo sencillo ("regresión") de un ajuste de χ^2

Tomemos una serie de 10 medidas de observables x e y, que muestran una clara anticorrelación. Suponemos que x es la variable *independiente*, y es la dependiente. Suponemos también que solamente las incertidumbres sobre y son pertinentes en este ejemplo.



Un ajuste a un polinomio de orden 1 da : $\chi^2 = 23.1$ para 8 grados de libertad.

```
**********************************
Minimizer is Linear
Chi2                      =       23.1088
NDf                       =             8
p0                        =       134.861   +/-   7.02278
p1                        =      -7.43446   +/-   1.13342

**********************************
Prob(23.1,8) = 0.00322733
```

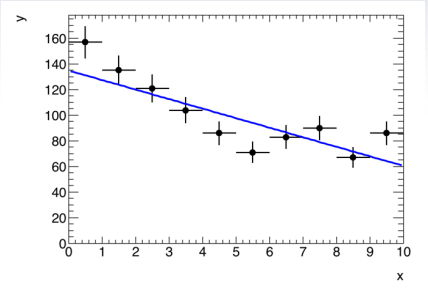


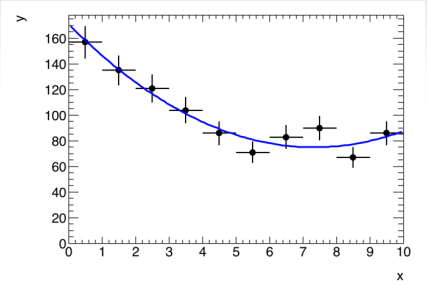
Mientras que un ajuste a un polinomio de orden 2 da : $\chi^2 = 6.5$ para 7 grados de libertad.

```
**********************************
Minimizer is Linear
Chi2                      =       6.51604
NDf                       =             7
p0                        =       170.778   +/-   11.2722
p1                        =      -26.1437   +/-   4.73078
p2                        =       1.78313   +/-   0.43774

**********************************
Prob(6.5,7) = 0.480943
```

Un ejemplo sencillo ("regresión") de un ajuste de χ^2

Tomemos una serie de 10 medidas de observables x e y, que muestran una clara anticorrelación. Suponemos que x es la variable *independiente*, y es la dependiente. Suponemos también que solamente las incertidumbres sobre y son pertinentes en este ejemplo.





Comentarios:

- claramente, mientras más "flexible" sea la función utilizada, "mejor" será el ajuste
- pero, ¡ese es rara vez el objetivo! (claro, puedo imaginar algún contraejemplo...)
- idealmente, la decisión sobre la función a utilizar debe ser tomada, previamente, **sin mirar los datos**

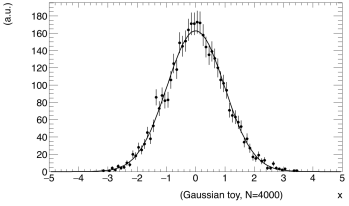
Otro ejemplo de ajuste de χ^2: un histograma

De manera general, la función de χ^2 que se minimiza es de esta forma:

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i; \vec{\theta})}{\sigma_i} \right)^2 ,$$

donde $\vec{\theta} = \{\theta_1, \theta_2, \cdots\}$ son los parámetros que deben ser determinados de manera a minimizar el valor de χ^2. En los ejemplos anteriores, teníamos

- p0 y p1 para el ajuste a un polinomio de orden uno
  - en este caso la minimización consiste en resolver un sistema lineal de 2 ecuaciones con dos incógnitas, la solución es analítica
- p0, p1 y p2 para el ajuste a un polinomio de orden dos



Otra situación usual es la de un histograma:

- aquí las incertidumbres asignadas a cada "bin" corresponden a la varianza en la altura del bin
- si se trata de conteos simples lo usual es asignar $\sigma_i = \sqrt{y_i}$
- cuidado sin embargo con los bines con pocas entradas...
- ¡y sobre todo con los bines vacíos!

Ejemplo de una distribución continua: Breit-Wigner (I)

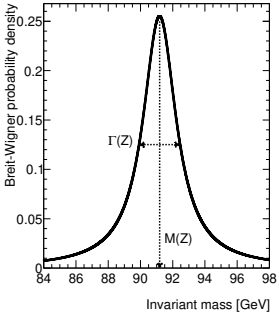
También llamada distribución de Cauchy, o Lorentziana, la distribución

$$P_{\mathrm{BW}}\left(x;\Gamma,x_0\right) \;=\; \frac{1}{\pi}\frac{\Gamma/2}{\left(x-x_0\right)^2+\Gamma^2/4}\;,$$

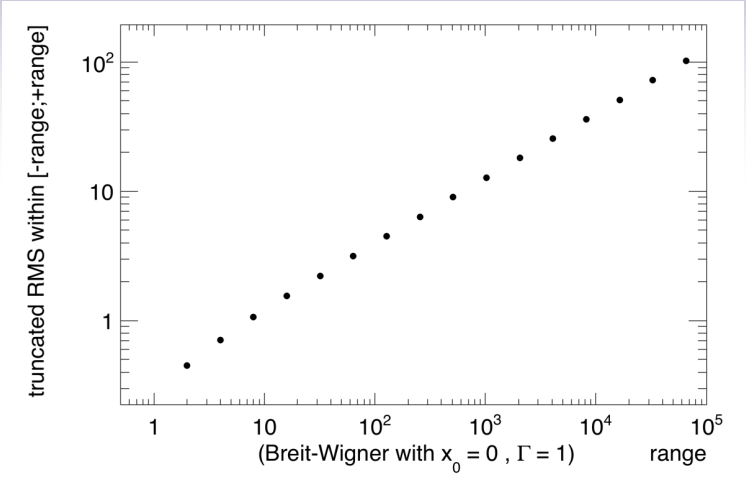
es a menudo utilizada para describir un proceso resonante (p.e. la masa invariante de productos del decaimiento de un estado intermedio resonante), con lo que x_0 y Γ son la masa y la anchura natural, que es su FWHM.
La BW es un ejemplo de una distribución *"fat-tailed"*:

- promedio empírico y RMS mal definidos (idem momentos superiores)
- sus varianzas aumentan con el tamaño de la muestra
- ejercicio : verificar con una simulación numérica que el RMS empírico de la Breit-Wigner aumenta (siguiendo una ley de potencia) con el rango en el que es estimado (ver gráfica en la lámina siguiente)
- un teorema curioso (que no demostraremos) el estimador de x_0 menos ineficaz es una media truncada sobre el 24% central de la muestra : otras truncaciones son menos eficaces
- ejercicio opcional : verificar este teorema con una simulación numérica

Gráfico : una Breit-Wigner con la masa y la anchura natural del bosón Z : $M_Z = 91.1876 \pm 0.0021\ \mathrm{GeV}/c^2$, $\Gamma_Z = 2.4952 \pm 0.0023\ \mathrm{GeV}/c^2$.



Ejemplo de una distribución continua: Breit-Wigner (II)



Ejemplo de una distribución continua: Voigtiana (I)

La Voigtiana es la convolución de una Breit-Wigner con una Gaussiana,

$$P_{\mathrm{Voigt}}\left(x; x_0, \Gamma, \sigma\right) \;=\; \int_{-\infty}^{+\infty} dx' \, P_{\mathrm{Gauss}}\left(x'; 0, \sigma\right) P_{\mathrm{BW}}\left(x - x'; x_0, \Gamma\right) \;,$$

- La Voigtiana es una distribución a tres parámetros : masa x_0, anchura natural Γ y resolución σ.
- Es un modelo de medidas de procesos resonantes, suponiendo que la resolución instrumental sea Gaussiana.
- No hay una forma analítica cerrada sencilla para la Voigtiana, pero hay implementaciones numéricas precisas y eficientes, p.e. la función miembro `TMath::Voigt` en ROOT, o la clase `RooVoigtian` en `RooFit`.
- Para valores de Γ y σ razonablemente similares, la FWHM de la Voigtiana se aproxima a una combinación en suma y en cuadratura :

$$\mathrm{FWHM}_{\mathrm{Voigt}} \;\simeq\; \left[(\Gamma/2) \oplus 2\sqrt{2\ln 2}\sigma\right] + \Gamma/2 \;.$$

Ejercicio :

- La figura en la lámina siguiente muestra la distribución de masa invariante dielectrón alrededor de la masa del bosón Z, obtenida con datos ATLAS a 13 TeV.
- Suponiendo que la distribución puede ser descrita por una Voigtiana, determinar aproximadamente cuál es la resolución en masa dielectrón del detector ATLAS.
- La parte inferior de la figura muestra una comparación entre los datos y la simulación. Comentar.

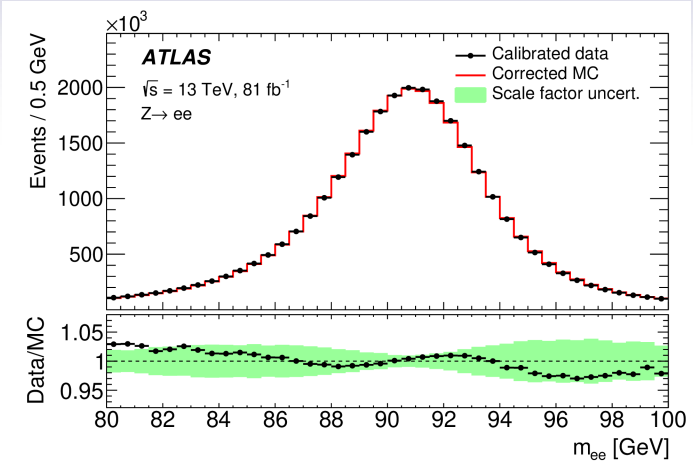
Ejemplo de una distribución continua: Voigtiana (II)

$$\mathrm{FWHM}_{\mathrm{Voigt}} \simeq \Gamma/2 + \left[(\Gamma/2) \oplus 2\sqrt{2\ln 2}\sigma\right]$$

$M_Z = 91.1876 \pm 0.0021\ \mathrm{GeV}/c^2$
$\Gamma_Z = 2.4952 \pm 0.0023\ \mathrm{GeV}/c^2$
$\sigma(m_{ee}) = ?$
En palabras simplificadas :

$$\mathrm{FWHM}_{Z\to ee} \simeq \Gamma_Z \otimes \sigma_{\mathrm{exp.}}$$

- para la física, el parámetro de interés es Γ_Z
- pero lo que se mide es $\mathrm{FWHM}_{Z\to ee}$
- !'se requiere efectuar un "unfolding" o *desconvolución* de la resolución experimental $\sigma_{\mathrm{exp.}}$!



Otros ejemplos de distribuciones continuas

La "Gaussiana Bifurcada" : útil para describir distribuciones asimétricas

$$P_{\mathrm{BG}}\left(x;\mu,\sigma_L,\sigma_R\right) = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma_L+\sigma_R}\times\begin{cases} e^{-\frac{(x-\mu)^2}{2\sigma_L^2}} & , \quad x\leq\mu \\ e^{-\frac{(x-\mu)^2}{2\sigma_R^2}} & , \quad x\geq\mu \end{cases}$$

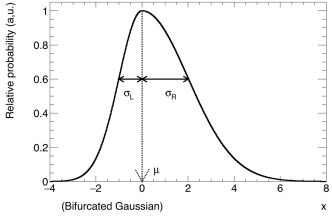
La "Crystal Ball" : en ocasiones utilizada para describir resoluciones con una componente de "fuga" (leakage)

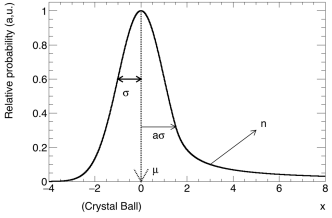
$$P_{\mathrm{CB}}(x;\mu,\sigma,\alpha,n) = N\times\begin{cases} e^{-\frac{(x-\mu)^2}{2\sigma^2}} & , \quad \frac{x-\mu}{\sigma}\leq\alpha \\ \left(\frac{n}{n-\alpha^2+\alpha\sigma^{-1}(x-\mu)}\right)^n e^{\left(-\frac{\alpha^2}{2}\right)} & , \quad \frac{x-\mu}{\sigma}\geq\alpha \end{cases}$$

Funciones implementadas p.e. en `RooFit`
Interpretación "intuitiva" de los parámetros, pero

- los estimadores simples están sesgados
- correlaciones muy grandes entre ellos
- utilizar, pero con prudencia y ojo crítico

Ejercicio numérico : evaluar las matrices de correlación
Otras funciones similares : DSCB, ACB ...





Otro ejemplo proveniente de una convolución

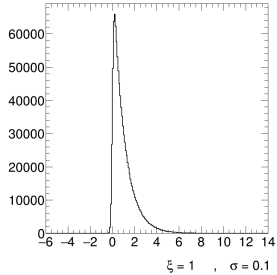
Una situación usual es cuando una distribución de tipo exponencial "a nivel de la verdad" es observada "a nivel experimental" con una resolución σ comparable a su escala característica ξ :

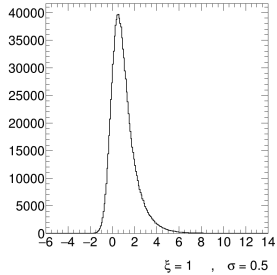
$$P_{\text{obs}}(x;\xi,x_0,\sigma) = P_{\text{true}}(x;\xi,x_0) \otimes P_{\text{Gauss}}(x;0,\sigma) \ , \ \text{con} \ \ P_{\text{true}}(x;\xi,x_0) = \frac{1}{\xi} e^{-(x-x_0)/\xi} \ ,$$

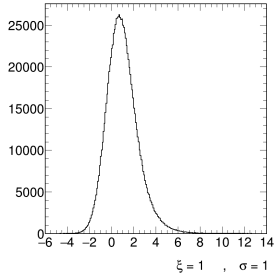
que puede mostrarse fácilmente (¡ejercicio!) que tiene la forma siguiente:

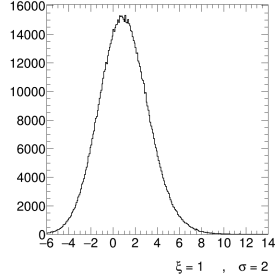
$$P_{\text{obs}}(x;\xi,x_0,\sigma) \propto e^{-\frac{\left(x-\frac{\sigma^2}{2\xi}\right)}{\xi}} \left[1 + \text{erf}\left(\frac{x-\sigma^2/\xi}{\sqrt{2}\sigma}\right)\right] \ , \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x dz e^{-z^2} \ .$$

donde $\text{erf}(x)$ es llamada "función de error".









CAPITULO V

EL METODO DE LA VEROSIMILITUD MAXIMA

En las previas láminas describimos una lógica "intuitiva" o "caso a caso" de la estimación de parámetros. Eso, por supuesto, no es generalizable (ni robusto) : de manera general, y más allá de los ejemplos sencillos

- los estimadores sencillos de los momentos de la función característica están sometidos a sesgos
- una descripción perfecta de la PDF require a priori un número infinito de momentos, lo cual no es posible por supuesto

Una descripción más general del problema es la siguiente :

- tenemos una muestra compuesta por N realizaciones independientes de variables aleatorias $\vec{x}$
- suponemos que esas realizaciones resultan de muestrear una PDF n-paramétrica

$$P\left(\vec{x};\theta_1,\ldots,\theta_n\right) \ ,$$

- suponemos también que la dependencia funcional de la PDF es conocida, y que solamente ignoramos los valores numéricos de los parámetros

Fisher, 1921: el *teorema de la verosimilitud máxima* (maximum likelihood) es una herramienta poderosa para la estimación de parámetros $\theta_1,\ldots,\theta_n$ de nuestra PDF.

Definimos la función de verosimilitud $\mathcal{L}$, evaluada sobre una muestra compuesta por N eventos :

$$\mathcal{L}(\theta_1, \ldots, \theta_n) \;=\; \prod_{i=1}^{N} P(\vec{x}_i; \theta_1, \ldots, \theta_n) \;\; .$$

Teorema : los valores $\hat{\theta_1}, \ldots, \hat{\theta}_n$ que maximizan la función $\mathcal{L}$ son estimadores de los parámetros $\theta_1, \ldots, \theta_n$,

$$\mathcal{L}\left(\hat{\theta}_1, \ldots, \hat{\theta}_n\right) \;=\; \max|_\theta \left\{\mathcal{L}(\theta_1, \ldots, \theta_n)\right\} \;\; ,$$

con varianzas $\hat{\sigma}_\theta$ que se extraen a partir de la matriz de covarianza de $\mathcal{L}$ alrededor de su máximo. Este método de estimación es de parámetros es llamado MLE.
En palabras intuitivas: para una muestra dada, ¡el MLE corresponde a los valores que maximizan la probabilidad de realizar esa muestra!
No es un pleonasmo : la función $\mathcal{L}$ debe satisfacer ciertas condiciones:

- ser derivable al menos dos veces con respecto a los parámetros $\theta_1, \ldots, \theta_n$,
- ser (asintóticalmente) no sesgada y eficiente (condición llamada "Cramer-Rao bound"),
- seguir una distribución (asintóticamente) multi-normal,

$$f\left(\hat{\vec{\theta}}, \vec{\theta}, \mathbf{\Sigma}\right) \;=\; \frac{1}{\sqrt{2\pi\,|\mathbf{\Sigma}|}} \exp\left\{-\frac{1}{2}\left(\hat{\vec{\theta}}_i - \vec{\theta}_i\right) \Sigma^{-1}_{ij} \left(\hat{\vec{\theta}}_j - \vec{\theta}_j\right)\right\} \;\; .$$

- seguir una distribución (asintóticamente) multi-normal,

$$f\left(\hat{\vec{\theta}}, \vec{\theta}, \mathbf{\Sigma}\right) \;=\; \frac{1}{\sqrt{2\pi}\,|\mathbf{\Sigma}|}\exp\left\{-\frac{1}{2}\left(\hat{\vec{\theta}}_i - \vec{\theta}_i\right)\Sigma^{-1}_{ij}\left(\hat{\vec{\theta}}_j - \vec{\theta}_j\right)\right\},$$

donde la matriz de covarianza $\mathbf{\Sigma}$ es

$$\mathbf{\Sigma}^{-1}_{ij} \;=\; -E\left[\frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j}\right] .$$

Al alejarnos del máximo (es decir, al pasar del punto $\hat{\vec{\theta}}$ a un punto cualquiera $\vec{\theta}$), el valor de la función $\mathcal{L}$ disminuye, a una tasa que depende de los elementos de la matriz de covarianza :

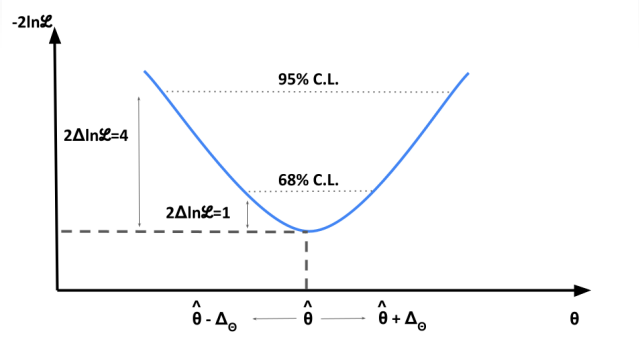
$$-2\Delta \ln \mathcal{L} \;=\; -2\left[\ln \mathcal{L}(\vec{\theta}) - \ln \mathcal{L}(\hat{\vec{\theta}})\right] \;=\; \sum_{i,j}\left(\theta_i - \hat{\theta}_i\right)\mathbf{\Sigma}^{-1}_{ij}\left(\theta_i - \hat{\theta}_j\right) .$$

En otras palabras: la matriz de covarianza define **mapas de contorno** alrededor de su máximo, que corresponden a **intervalos de confianza**.
En el caso de una $\mathcal{L}$ con un parámetro único $\mathcal{L}(\theta)$, el intervalo contenido dentro de $-2\Delta \ln \mathcal{L} < 1$ alrededor de $\hat{\theta}$ define un intervalo de confianza a 68% que corresponde a un rango $-\Delta_\theta \leq \theta - \hat{\theta} \leq \Delta_\theta$ alrededor del punto máximo.
Por ello el resultado de MLE se escribe a menudo como $\left(\hat{\theta} \pm \hat{\Delta}_\theta\right)$.

El teorema de verosimilitud máxima (III)

En el caso de una $\mathcal{L}$ con un parámetro único $\mathcal{L}(\theta)$, el intervalo contenido dentro de $-2\Delta \ln \mathcal{L} < 1$ alrededor de $\hat{\theta}$ define un intervalo de confianza a 68% que corresponde a un rango $-\Delta_\theta \leq \theta - \hat{\theta} \leq \Delta_\theta$ alrededor del máximo. Por ello el resultado de MLE se escribe a menudo como $\left(\hat{\theta} \pm \hat{\Delta}_\theta\right)$.

(y de la misma manera, el intervalo $-2\Delta \ln \mathcal{L} < 4$ define un intervalo de confianza a 95% corresponde a un rango $-2\Delta_\theta \leq \theta - \hat{\theta} \leq 2\Delta_\theta$, etc...)



Ejemplo de estimación por verosimilitud máxima (I) : la Gaussiana

Sea una muestra, compuesta por N realizaciones de una única variable aleatoria x, que suponemos sigue una distribución Gaussiana de media μ y anchura σ. El teorema MLE nos permite estimar los parámetros así:

$$\mathcal{L}(\mu,\sigma) = \prod_{i=1}^{N} P_{\text{Gaus}}(x_i;\mu,\sigma) \; ; -\ln\mathcal{L} = N\ln\sigma + \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2} \; (+\text{constante}).$$

Los estimadores $\hat{\mu}$ y $\hat{\sigma}$ son los ceros de las primeras derivadas de $-\ln\mathcal{L}$ con respecto a μ y σ :

$$\frac{\partial}{\partial\mu}(-\ln\mathcal{L})\Big|_{\hat{\mu},\hat{\sigma}} = 0 \quad \longrightarrow \quad \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \; .$$

$$\frac{\partial}{\partial\sigma}(-\ln\mathcal{L})\Big|_{\hat{\sigma},\hat{\mu}} = 0 \quad \longrightarrow \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \; .$$

Si se ajusta una muestra usando el MLE y una PDF Gaussiana,
¡el estimador de μ es el promedio empírico!
¡y el estimador de σ es el RMS empírico!

Las segundas derivadas nos dan los elementos de la matriz de covarianza:

$$\frac{\partial^2}{\partial\mu^2}(-\ln\mathcal{L})\Big|_{\hat{\mu},\hat{\sigma}} = \frac{N}{\hat{\sigma}^2} \quad \longrightarrow \Sigma_{\mu\mu} = \frac{\hat{\sigma}^2}{N} \quad \longrightarrow \hat{\Delta}_\mu = \frac{\hat{\sigma}}{\sqrt{N}} \; ,$$

$$\frac{\partial^2}{\partial\sigma^2}(-\ln\mathcal{L})\Big|_{\hat{\mu},\hat{\sigma}} = \frac{2N}{\hat{\sigma}^2} \quad \longrightarrow \Sigma_{\sigma\sigma} = \frac{\hat{\sigma}^2}{2N} \quad \longrightarrow \hat{\Delta}_\sigma = \frac{\hat{\sigma}}{\sqrt{2N}} \; ,$$

Nota importante :
¡los errores $\hat{\Delta}_\mu$ y $\hat{\Delta}_\sigma$ escalan con el tamaño de la muestra N siguiendo la relación $1/\sqrt{N}$!
¡ambos errores son proporcionales al RMS empírico!

(y los términos no diagonales de la covarianza son ambos cero).

$\longrightarrow$ el MLE extrae de manera formal y robusta los parámetros y errores $(\hat{\mu}\pm\hat{\Delta}_\mu)$ y $(\hat{\sigma}\pm\hat{\Delta}_\sigma)$ de una Gaussiana.

Consideremos ahora otra muestra, compuesta por N realizaciones de una única variable aleatoria x, que suponemos sigue una distribución exponencial con parámetro de forma ξ. El teorema MLE nos permite estimar ese parámetro ξ analíticamente, al menos para el caso en que la variable x cubre el rango $[0; +\infty]$.

$$\mathcal{L}(\xi) = \prod_{i=1}^{N} P_{\text{exponential}}\,(x_i; \xi) = \prod_{i=1}^{N} \frac{1}{\xi} e^{-x_i/\xi} \ .$$

De manera análoga al ejemplo anterior, determinamos el NLL :

$$-\ln \mathcal{L} = N \ln \xi - \frac{1}{\xi} \sum_{i=1}^{N} x_i \ ,$$

y evaluamos sus primera y segunda derivadas para el valor $\hat{\xi}$ que anula la primera derivada :

$$\frac{\partial}{\partial \xi}(-\ln \mathcal{L})\bigg|_{\hat{\xi}} = 0 \ \longrightarrow \ \hat{\xi} = \frac{1}{N} \sum_{i=1}^{N} x_i \ ,$$

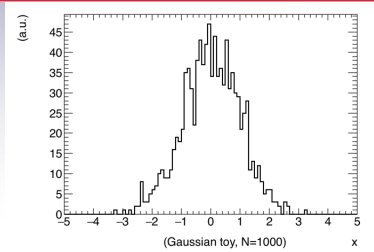
$$\frac{\partial^2}{\partial \xi^2}\bigg|_{\hat{\xi}} = \frac{N}{\hat{\xi}^2} \ \longrightarrow \ \Sigma_{\xi\xi} = \frac{\hat{\xi}^2}{N} \ \longrightarrow \ \hat{\Delta}_\xi = \frac{\hat{\xi}}{\sqrt{N}} \ .$$

Con lo que el MLE nos da en este caso también una solución analítica para la estimación del parámetro de forma exponencial y su correspondiente error $(\hat{\xi} \pm \hat{\Delta}_\xi)$.

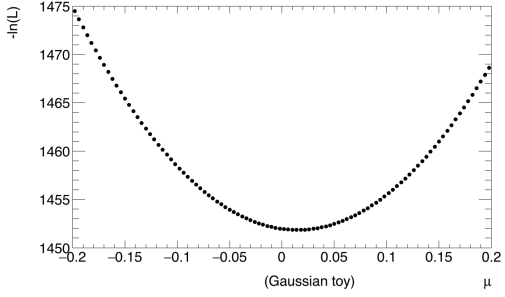
Nota: en este caso como en el anterior, los errores $\hat{\Delta}$ escalan como $1/\sqrt{N}$. Esa es una propiedad muy general.

Un "ejemplo de juguete" (I)

- Generamos una realización aleatoria con N=1000 eventos ("toy sample") a partir de una PDF Gaussiana reducida, con $\mu = 0$, $\sigma = 1$.
- Evaluamos (el negativo del logaritmo de) la función de verosimilitud para esa muestra, para diferentes valores de μ y σ ("escaneamos" los parámetros de interés)



$$-\ln \mathcal{L}(\mu, \sigma) = N \ln \sigma + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \ .$$

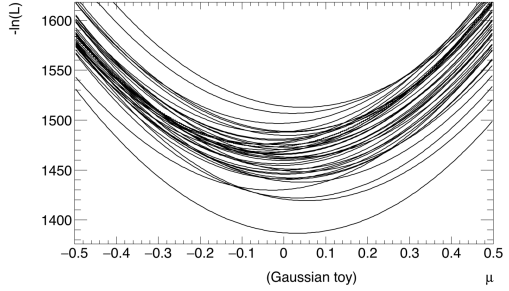


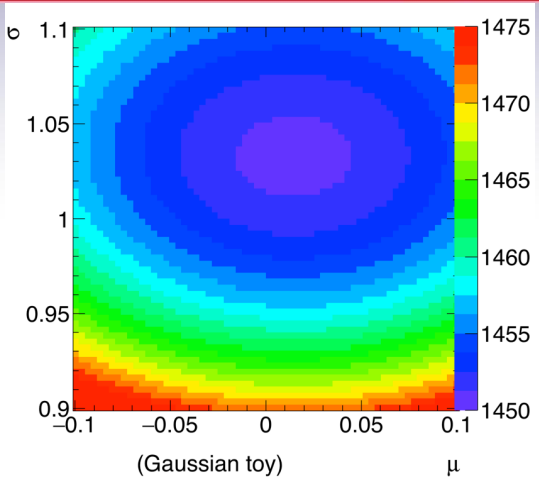
- La forma de $-\ln \mathcal{L}$ en función de μ (para σ fijo) sigue un perfil parabólico, y el mínimo coincide con el promedio empírico
- Por inspección alrededor del mínimo, se observa que el NLL aumenta en 0.5 unidades para $\Delta\mu \sim \pm 0.03$, que corresponde aproximadamente a $\sigma/\sqrt{N}$ para $N = 1000$
- La lámina siguiente muestra el gráfico del escaneo en 2D de μ y σ

Un "ejemplo de juguete" (II)

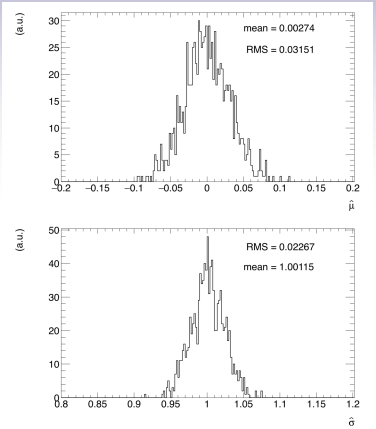
- El perfil bidimensional de $-\ln \mathcal{L}$ es el de un paraboloide, con semi-ejes diferentes y ortogonales
- El mínimo coincide con la posición del promedio y el RMS empíricos
- El semi-eje a lo largo de σ es más estrecho que el de μ, como se espera de la relación $1/\sqrt{2N}$ vs. $1/\sqrt{N}$

Para otras realizaciones aleatorias independientes, la posición y la anchura de los mínimos cambia :



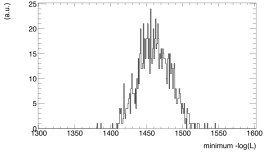


Un "ejemplo de juguete" (III)



Si realizamos el mismo estudio sobre un ensamble de muestras generadas con la misma PDF (aquí N=1000, $\mu = 0$, $\sigma = 1$) obtenemos los resultados siguientes :

- los mínimos de $\hat{\mu}$ fluctúan alrededor de su valor verdadero $\mu = 0$ con una dispersión de $\pm 3.1\%$, que corresponde a $\sigma/\sqrt{N}$
- los mínimos de $\hat{\sigma}$ fluctúan alrededor de su valor verdadero $\sigma = 1$ con una dispersión de $\pm 2.3\%$, que corresponde a $\sigma/\sqrt{2N}$
- los intervalos con $-\Delta \ln \mathcal{L} < 0.5$ alrededor del mínimo corresponden bien a las regiones que cubren 68.3% de la dispersión
- el teorema MLE nos da una definición rigurosa y precisa de la incertidumbre estadística

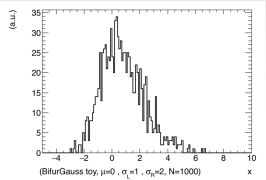


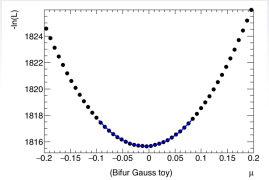
La distribución del $-\ln \mathcal{L}$ no siempre sigue una forma precisa (aquí parece bastante Gaussiana) pero permite hacer una estimación del "goddness-of-fit" :

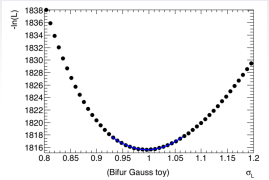
- si el $-\ln \mathcal{L}$ observado en una muestra se aleja significativamente del intervalo cubierto en un "estudio de juguete", la calidad del modelo es sospechosa...

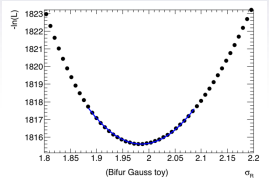
Otro "ejemplo de juguete" (I)

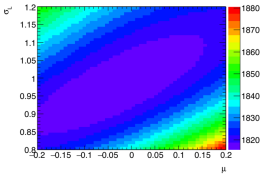
Consideremos ahora una Gaussiana bifurcada, PDF con tres parámetros : μ, σ_L y σ_R. Los mínimos del $-\log \mathcal{L}$ no son del todo parabólicos, y hay correlaciones importantes entre los parámetros.

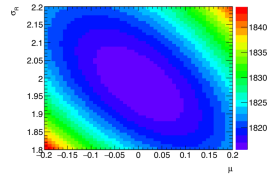


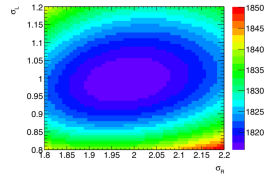










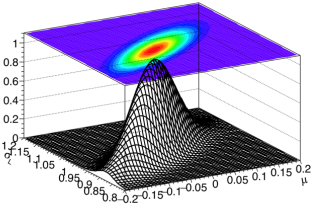


Otro "ejemplo de juguete" (II)

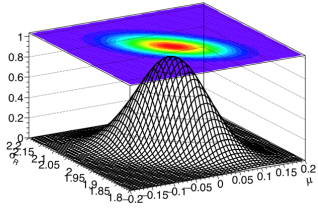
Los parámetros de la PDF son ellos mismos variables aleatorias:

- si efectuamos otras realizaciones aleatorias a partir de la misma PDF, obtendremos valores diferentes de los parámetros
- y éstos se distribuirán siguiendo la matriz de covariancia de los parámetros (y por tanto tomando en cuenta las correlaciones)

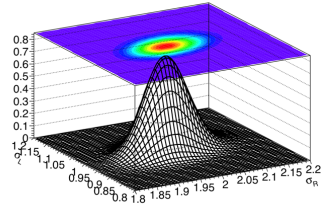
Correlación entre μ y σ_L : +79%



Correlación entre μ y σ_R : -59%



Correlación entre σ_L y σ_R : +19%



Nota: se trata por supuesto de una PDF en 3 dimensiones: por eso se muestran 3 proyecciones bidimensionales.

Los ejemplos de juguete discutidos previamente son casos sencillos, con 2 o 3 parámetros, y pudimos explorar el espacio bi- o tri-dimensional con bucles sencillas.
En situaciones más generales, el número de parámetros puede ser significativamente superior, así que la aproximación de "escanear" los parámetros para identificar el mínimo del $-\log \mathcal{L}$ no es eficaz (y se vuelve rápidamente imposible).
Por ello se utilizan algoritmos de minimización numérica, que optimizan la búsqueda del mínimo de una función. El proceso se llama *ajuste numérico* o "fit". El algoritmo más usado en altas energías es MINUIT diseñado en el CERN en los años 1970. MINUIT está implementado en `ROOT`, en la clase `TMinuit`.

MINUIT

From Wikipedia, the free encyclopedia

MINUIT, now **MINUIT2**, is a numerical minimization computer program originally written in the FORTRAN programming language[1] by CERN staff physicist Fred James in the 1970s. The program searches for a minimum in a user-defined function with respect to one or more parameters using several different methods as specified by the user. In addition to that it can compute confidence intervals for the parameters by scanning the function around the minimum.

The original FORTRAN code was later ported to C++ by the ROOT project; both the FORTRAN and C++ versions are in use today. The program is very widely used in particle physics, and thousands of published papers cite use of MINUIT.[2] In the early 2000s, Fred James started a project to implement MINUIT in C++ using object-oriented programming. The new MINUIT is an optional package (minuit2) in the ROOT release. As of October 2014 the latest version is 5.34.14, released on 24 January 2014.[3] There is also a Java port[4] as well as a Python frontend to the C++ code.[5]

MINUIT is not a program that can be distributed as an executable binary to be run by a relatively unskilled user: the user must write and compile a subroutine defining the function to be optimized, and oversee the optimization process.

Relación entre un fit NLL y un fit de χ^2

Muchos paquetes y programas contienen algoritmos de minimización más sencillos, del tipo "mínimos cuadrados" o χ^2. Estos difieren de un ajuste de verosimilitud máxima: aquí la muestra se compone de un conjunto de n medidas con sus incertidumbres $(y_i \pm \sigma_i)$, que dependen de una serie de medidas independientes x_i, y la función a minimizar es:

$$\chi^2(\theta) = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i;\theta)}{\sigma_i} \right)^2 ,$$

donde $f(x;\theta)$ es la función que pretende describir una dependencia funcional $y = f(x)$. La minimización del χ^2 provee los valores $\hat{\theta}$ estimados por el ajuste.
La relación con la verosimilitud es fácil de establecer :

$$\begin{aligned} e^{-\frac{1}{2}\chi^2(\theta)} &= e^{-\sum_{i=1}^{n} \frac{(y_i - f(x_i;\theta))^2}{2\sigma_i^2}} , \\ &= \prod_{i=1}^{n} e^{-\frac{(y_i - f(x_i;\theta))^2}{2\sigma_i^2}} , \end{aligned}$$

que muestra que minimizar el χ^2 es equivalente (a excepción de una constante global de normalización) a maximizar el likelihood.
Esta relación entre χ^2 y NLL muestra además por qué los intervalos de confianza a 68% vienen determinados por un cambio en media unidad en el likelihood.

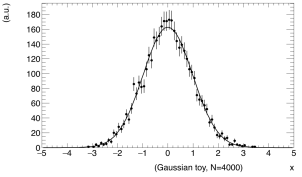
Relación entre un fit NLL y un fit de χ^2

Muchos paquetes y programas contienen algoritmos de minimización más sencillos, del tipo "mínimos cuadrados" o χ^2. Estos difieren de un ajuste de verosimilitud máxima: aquí la muestra se compone de un conjunto de n medidas con sus incertidumbres $(y_i \pm \sigma_i)$, que dependen de una serie de medidas independientes x_i, y la función a minimizar es:

$$\chi^2(\theta) = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i;\theta)}{\sigma_i} \right)^2 ,$$

donde $f(x;\theta)$ es la función que pretende describir una dependencia funcional $y = f(x)$. La minimización del χ^2 provee los valores $\hat{\theta}$ estimados por el ajuste.

El conjunto de puntos puede provenir de medidas individuales, o ser una reducción por histogrameo de una muestra completa: en el ejemplo aquó se tiene una muestra compuesta de 4000 valores de una variable aleatoria x, histogrameada con 100 *bines* de anchura $\delta x = 0.1$ en el intervalo $-5 \leq x \leq +5$. El equivalente a y_i es el número de eventos con valores contenidos en el bin i, y la incertidumbre asociada es $\sigma_i = \sqrt{(y_i)}$.

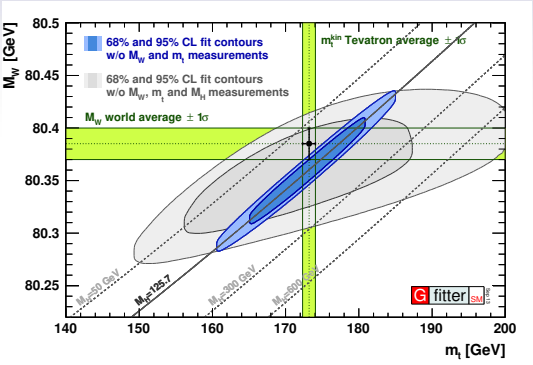


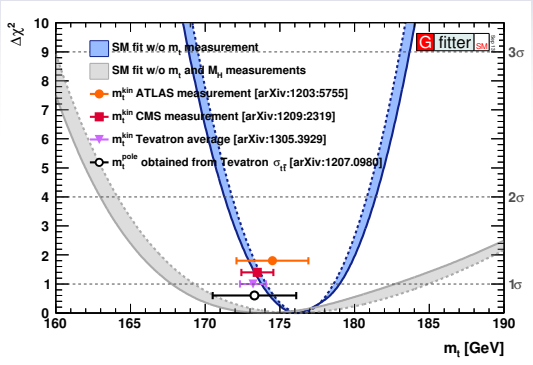
El ajuste de este histograma a una función Gaussiana nos da $\mu = (0.008 \pm 0.016)$ y $\sigma = (1.007 \pm 0.011)$.

Nota: si se hubiera usado otro "binning" ¡el resultado podría resultar un poco diferente!

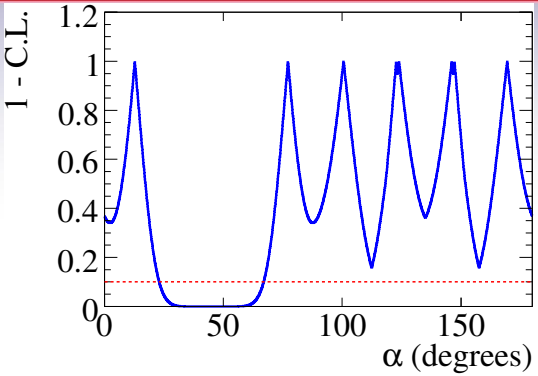
Nota: en cambio, el ajuste por verosimilitud máxima a esta misma muestra nos dará exactamente los resultados teóricos: la media empírica y el RMS empírico, con sus incertidumbres $\mathrm{RMS}/\sqrt{n}$ y $\mathrm{RMS}/\sqrt{2n}$.

Un ejemplo : los intervalos de confianza de Gfitter

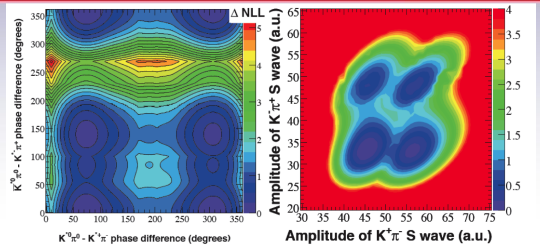




Otros ejemplos de funciones de verosimilitud complicadas



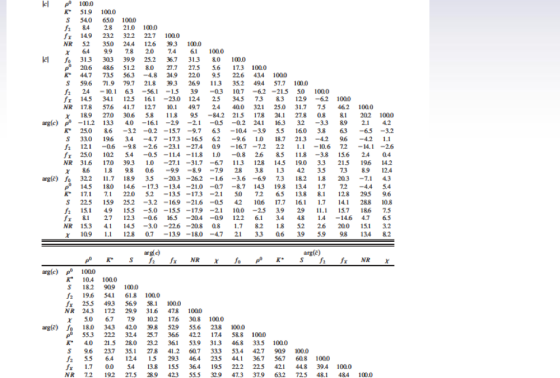
Un elemento importante del programa científico del experimento *BABAR* (y en general, de la *física de sabores*): la medida del ángulo α de la matriz CKM: $\alpha \neq 0 \longrightarrow$ violación de la simetría CP.
Problema : el observable físico es la asimetría dependiente del tiempo $B^0/\overline{B}^0 \to \pi^+\pi^-$, que es una función de $\sin 2(\alpha - \delta)$: ambigüedad octuple.



Situación más complicada: el perfil de interferencias entre las amplitudes de decaimiento $B^0/\overline{B}^0 \to K^+\pi^-\pi^0$;

- ambigüedades múltiples, sin forma analítica precisa...
- numerosas amplitudes resonantes intermedias: $K^{\star\pm}\pi^{\mp}$, $K^{\star 0}\pi^0$, $\rho^{\pm}\pi^{\mp}$, varios estados excitados $K^{\star}$, ρ, otras resonancias... todo para B^0 y para $\overline{B}^0$
- la función de verosimilitud contenía unos 60 parámetros de interés: amplitudes y fases...

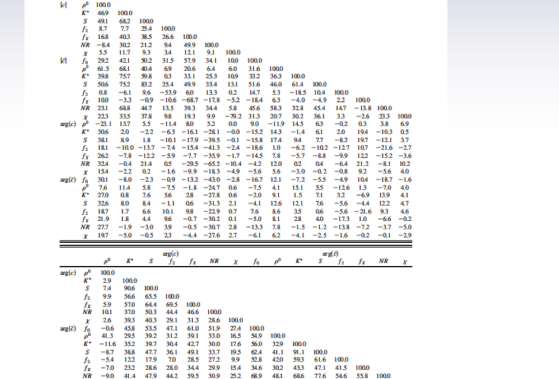
Otros ejemplos de funciones de verosimilitud complicadas



B. AUBERT *et al.* PHYSICAL REVIEW D **80,** 112001 (2009)

TABLE VIII. Full correlation matrix for the isobar parameters of solution I. The entries are given in percent. Since the matrix is symmetric, all elements above the diagonal are omitted.

		ρ^0	K^*	S	f_2	f_X	NR	χ	f_0	ρ^0	K^*	S	f_2	f_X	NR	χ
							$\|c\|$								$\|d\|$	
$\|c\|$	ρ^0	100.0														
	K^*	31.9	100.0													
	S	54.0	65.0	100.0												
	f_2	8.4	2.8	21.0	100.0											
	f_X	14.9	23.2	32.2	22.7	100.0										
	NR	5.2	35.0	24.4	12.6	39.3	100.0									
	χ	6.4	9.9	7.8	2.0	7.4	6.1	100.0								
$\|d\|$	f_0	31.3	30.3	39.9	25.2	36.7	31.3	8.0	100.0							
	ρ^0	20.6	48.6	51.2	8.0	27.7	27.5	5.6	17.3	100.0						
	K^*	44.7	73.5	56.3	−4.8	26.9	22.0	9.5	22.6	43.4	100.0					
	S	59.6	71.9	79.7	21.8	39.3	36.9	11.3	35.2	49.4	57.7	100.0				
	f_2	2.4	−10.1	6.3	−56.1	−1.5	3.9	−0.3	10.7	−6.2	−21.5	5.0	100.0			
	f_X	14.5	34.1	12.5	16.1	−23.0	12.4	2.5	34.5	7.3	8.3	12.9	−6.2	100.0		
	NR	17.8	57.6	41.7	12.7	10.1	49.7	2.4	40.0	32.1	25.0	31.7	7.5	46.2	100.0	
	χ	18.9	27.0	30.6	5.8	11.8	9.5	−84.2	21.5	17.8	28.1	27.8	0.8	8.1	10.2	100.0
$arg(c)$	ρ^0	−11.2	13.3	4.0	−16.1	−2.9	−2.1	−0.5	−0.2	24.1	16.3	3.2	−3.3	8.9	2.1	4.2
	K^*	25.0	8.6	−3.2	−0.2	−15.7	−9.7	6.3	−10.4	−3.9	5.5	16.0	3.8	6.3	−6.5	−3.2
	S	33.0	19.6	3.4	−4.7	−17.3	−16.5	6.2	−9.6	1.0	18.7	21.3	−4.2	9.6	−12.2	1.1
	f_2	12.1	−0.6	−9.8	−2.6	−23.1	−27.4	0.9	−16.7	−7.2	2.2	1.1	−10.6	7.2	−14.1	−2.6
	f_X	25.0	10.2	5.4	−6.5	−11.4	−11.8	1.0	−0.8	2.6	8.5	11.8	−3.8	15.6	2.4	0.4
	NR	31.6	17.0	39.3	1.0	−27.1	−31.7	−6.7	11.3	12.8	14.5	19.0	3.3	21.5	19.6	14.2
	χ	8.6	1.8	9.8	0.6	−9.9	−8.9	−7.9	2.8	3.8	1.3	4.2	3.5	7.3	8.9	12.4
$arg(d)$	f_0	32.2	11.7	18.9	3.5	−20.3	−26.2	−1.6	−3.6	−6.9	7.3	18.2	1.8	20.3	−7.1	4.3
	ρ^0	14.5	18.0	14.6	−17.3	−13.4	−21.0	−0.7	−8.7	14.3	19.8	13.4	1.7	7.2	−4.4	5.4
	K^*	17.1	7.1	2.0	5.2	−13.5	−17.3	−2.1	3.0	7.2	6.5	13.8	8.1	12.8	29.5	9.6
	S	22.5	15.9	25.2	−3.2	−16.9	−21.6	−0.5	4.2	10.6	17.7	16.1	1.7	14.1	28.8	10.8
	f_2	15.1	4.9	15.3	−5.0	−15.5	−17.9	−2.1	10.0	−2.3	3.9	2.9	11.1	15.7	18.6	7.5
	f_X	8.1	2.7	12.3	−0.6	16.5	−20.4	−0.9	12.2	6.1	3.4	4.8	1.4	−14.6	4.7	6.5
	NR	15.3	4.1	14.5	−3.0	−22.6	−20.8	0.8	1.7	8.2	1.8	5.2	2.6	20.0	15.1	3.2
	χ	10.9	1.1	12.8	0.7	−13.9	−18.0	−4.7	2.1	3.3	0.6	3.9	5.9	9.8	13.4	8.2



TIME-DEPENDENT AMPLITUDE ANALYSIS OF … PHYSICAL REVIEW D **80,** 112001 (2009)

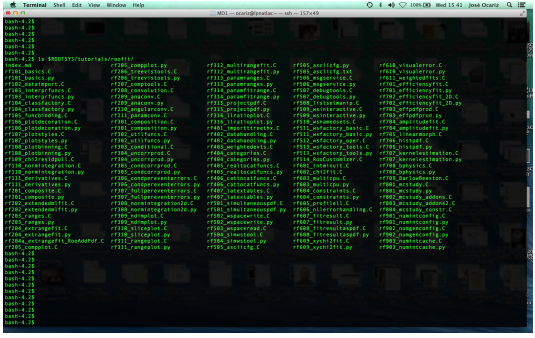
TABLE IX. Full correlation matrix for the isobar parameters of solution II. The entries are given in percent. Since the matrix is symmetric, all elements above the diagonal are omitted.

		ρ^0	K^*	S	f_2	f_X	NR	χ	f_0	ρ^0	K^*	S	f_2	f_X	NR	χ
							$\|c\|$								$\|d\|$	
$\|c\|$	ρ^0	100.0														
	K^*	66.9	100.0													
	S	69.1	68.2	100.0												
	f_2	8.7	7.7	25.4	100.0											
	f_X	16.8	40.3	38.5	26.6	100.0										
	NR	−8.4	30.2	21.2	9.4	49.9	100.0									
	χ	5.5	11.7	9.3	3.4	12.1	9.1	100.0								
$\|d\|$	f_0	29.2	42.1	50.2	31.5	57.9	34.1	10.0	100.0							
	ρ^0	61.5	68.1	40.4	6.9	20.6	6.4	6.0	31.6	100.0						
	K^*	59.8	75.7	59.8	0.3	33.1	25.3	10.9	33.2	36.3	100.0					
	S	50.6	75.2	85.2	25.4	49.9	33.4	13.1	51.6	46.0	61.4	100.0				
	f_2	0.8	−6.1	9.6	−53.9	6.0	13.3	0.2	14.7	5.3	−18.5	10.4	100.0			
	f_X	10.0	−3.3	−0.9	−10.6	−66.7	−17.8	−5.2	−18.4	6.3	−4.0	−4.9	2.2	100.0		
	NR	23.1	68.8	44.7	13.5	39.3	34.4	5.8	45.6	58.3	33.8	45.4	14.7	−13.8	100.0	
	χ	23.3	33.5	37.8	9.8	19.3	9.9	−79.2	31.3	20.7	30.2	36.1	3.3	−2.6	23.3	100.0
$arg(c)$	ρ^0	−23.1	13.7	5.5	−11.4	8.0	5.2	0.0	9.0	−11.9	14.5	6.3	−0.2	0.3	3.8	6.9
	K^*	30.6	2.0	−2.2	−6.3	−16.1	−28.1	−0.0	−15.2	14.3	−1.4	6.1	2.0	16.4	−10.3	0.5
	S	31.1	7.1	1.8	−10.1	−17.9	−39.5	−0.1	−15.8	17.4	9.4	7.7	−8.2	19.7	−12.1	3.7
	f_2	18.1	−10.0	−13.7	−7.4	−15.4	−41.3	−2.4	−18.6	1.0	−6.2	−10.2	−12.7	10.7	−21.6	−2.7
	f_X	26.2	−7.8	−12.2	−5.9	−7.7	−35.9	−1.7	−14.5	7.8	−5.7	−8.8	−9.9	12.2	−15.2	−3.6
	NR	32.4	−0.4	21.4	0.5	−29.5	−65.2	−10.4	−4.2	12.0	0.2	0.4	−6.4	21.2	−8.1	10.2
	χ	15.4	−2.2	0.2	−1.6	−9.9	−18.3	−4.9	−5.6	5.6	−3.0	−0.2	−0.8	9.2	−5.6	4.0
$arg(d)$	f_0	30.1	−8.0	−2.3	−0.9	−13.2	−43.0	−2.8	−16.7	12.1	−7.3	−5.5	−4.9	10.4	−18.7	−1.6
	ρ^0	7.6	11.4	5.8	−7.5	−1.8	−24.7	0.6	−7.5	4.1	15.1	5.5	−12.6	1.3	−7.0	4.0
	K^*	27.0	7.6	7.6	5.6	2.8	−27.4	0.6	−2.0	9.1	1.3	7.1	3.2	−6.9	13.9	4.1
	S	32.6	8.0	8.4	−1.1	0.6	−31.3	2.1	−4.1	12.6	12.1	7.6	−5.6	−4.4	12.2	4.7
	f_2	18.7	1.7	6.6	10.1	9.8	−22.9	0.7	7.6	8.6	3.5	0.6	−5.6	−21.6	9.3	4.6
	f_X	21.9	1.8	4.4	9.6	−0.7	−30.2	0.1	−5.0	8.1	2.8	4.0	−17.3	1.0	−6.6	−0.2
	NR	27.7	17.9	13.0	3.9	−6.5	−30.7	2.8	−13.3	7.8	−1.5	−1.2	−13.8	−7.2	−3.7	−5.0
	χ	19.7	−5.0	−0.5	2.3	−4.4	−27.6	2.7	−6.1	6.2	−4.1	−2.5	−1.6	−0.2	−0.1	−2.9

El paquete RooFit

- Desarrollado a partir del 2005 en el experimento *BABAR*
- Viene incluido en ROOT desde Root4
- Utilizado en muchos análisis experimentales de altas energías
- Ventajas :
  - Minimización por MINUIT
  - Normalización automática de las PDFs, analítica si la conoce, numérica de lo contrario
  - Numerosas clases y ejemplos

- Inconvenientes :
  - Documentación parcial y no actualizada
  - Normalización automática de las PDFs...



MLE en situaciones más elaboradas (I)

En un escenario típico, un proceso aleatorio puede tener contribuciones de origen diferente.
Para ser específicos, consideremos que los eventos que componen la muestra provienen de dos "especies", llamadas de manera genérica "señal" y "fondo" (la generalización a más de dos especies es sencilla).
Cada especie se realiza a partir de su propia densidad de probabilidad.
Si los rangos de las variables aleatorias no son totalmente disyuntos, es imposible saber evento a evento a cuál de las especies pertenece. Pero el MLE permite efectuar una *separación estadística*: la PDF subjacente es a combinación de mas PDFs de señal y fondo,

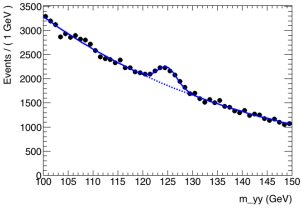
$$\mathcal{L}\left(f_{\rm sig}, \theta; \vec{x}\right) \;=\; \prod_{i=1}^{N} \left[f_{\rm sig} P_{\rm sig}(\vec{x};\theta) + (1 - f_{\rm sig}) P_{\rm bkg}(\vec{x};\theta) \right] \;,$$

donde $P_{\rm sig}$ y $P_{\rm bkg}$ son las PDFs de señal y fondo, respectivamente, y la fracción de señal $f_{\rm sig}$ es el parámetro que cuantifica la pureza de la muestra : $0 \leq f_{\rm sig} \leq 1$.

Ejemplo inspirado de la búsqueda del bosón de Higgs en el canal difotón :
Con ROOT instalado, la macro `H_yy.cc` debe correr sin problema, haciendo

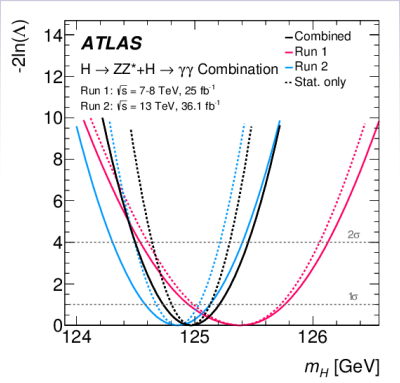
```
prompt> root -l H_yy.cc
```

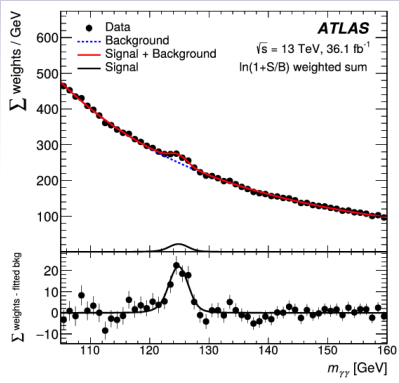
```
RooFitResult: minimized FCN value: 386054, estimated distance to minimum: 2.29788e-05
              covariance matrix quality: Full, accurate covariance matrix
              Status : MIGRAD=0 HESSE=0

  Constant Parameter    Value
------------------  ------------
             sig_s   2.0000e+00

  Floating Parameter  InitialValue    FinalValue +/-  Error     GblCorr.
------------------  --------------  --------------  --------  --------
              bkg_a     -2.2500e-02   -2.2788e-02 +/-  2.31e-04  <none>
              sig_f      2.0000e-02    1.8732e-02 +/-  1.33e-03  <none>
              sig_m      1.2500e+02    1.2494e+02 +/-  1.92e-01  <none>
```



MLE en situaciones más elaboradas (II)

Figuras tomadas de M. Aaboud *etal*, the ATLAS Collaboration, Phys.Lett.B 784 (2018) 345-366





La figura que representa $-2\Delta(\ln \mathcal{L})$ en función de m_H ilustra claramente la interpretación del MLE en términos de *intervalos de confianza*:

- el rango de m_H que corresponde a $-2\Delta(\ln \mathcal{L}) < 1$, "un sigma", cubre 68% de los resultados que se obtendrían repitiendo el experimento ATLAS numerosas veces;
- idem para el rango de m_H correspondiendo a $-2\Delta(\ln \mathcal{L}) < 4, 9, \cdots$ ("dos sigma", "tres sigma", etc...)

En experimentos de conteo de eventos, el número de eventos observados puede ser un parámetro de interés. Para el caso de una especie única, esto corresponde a "extender" la verosimilitud,

$$\mathcal{L}(\lambda, \theta; \vec{x}) \;=\; \frac{\lambda^N e^{-\lambda}}{N!} \prod_{i=1}^{N} P(\vec{x_i}; \theta) \;\; .$$

dónde el término multiplicativo adicional, corresponde a la distribución de Poisson (el término $N!$ en el denominador es irrelevante; es un factor global sin impacto sobre la forma de la verosimilitud)
Es fácil verificar que la verosimilitud es máxima cuand $\hat{\lambda} = N$, tal como se espera; ahora, si algunas de las PDFs dependen también de λ, el valor $\hat{\lambda}$ que maximiza $\mathcal{L}$ puede diferir.
La generalización a más de una especie es sencilla; para cada especie, un término multiplicativo de Poisson se incluye en la verosimilitud extendida, y las PDFs de cada especie son ponderadas por su fracción relativa de eventos.
Para el caso de dos especies, la versión extendida de la verosimilitud es

$$\mathcal{L}\left(N_{\rm sig}, N_{\rm bkg}, \theta\right) \;=\; (N_{\rm sig} + N_{\rm bkg})^N e^{-(N_{\rm sig}+N_{\rm bkg})} \prod_{i=1}^{N} \left[N_{\rm sig} P_{\rm sig}(\vec{x}; \theta) + N_{\rm bkg} P_{\rm bkg}(\vec{x}; \theta)\right] \;\; .$$

Estimar eficiencias a partir de ajustes MLE

Consideremos de nuevo el caso de un proceso aleatorio con dos resultados posibles: "yes" y "no". El estimador intuitivo de la eficiencia ε es el cociente entre el número de realizaciones de cada tipo, $n_{\rm yes}$ y $n_{\rm no}$, y su varianza $V\left[\hat{\varepsilon}\right]$ viene dada por :

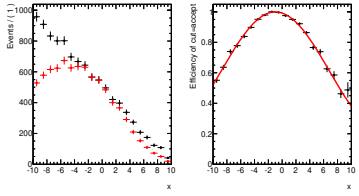
$$\hat{\varepsilon} \;=\; \frac{n_{\rm yes}}{n_{\rm yes}+n_{\rm no}} \;,\; V\left[\hat{\varepsilon}\right] \;=\; \frac{\hat{\varepsilon}(1-\hat{\varepsilon})}{n} \;,$$

donde $n = n_{\rm yes} + n_{\rm no}$ es el número total de realizaciones. (ejercicio: reproducir este resultado)
Este estimador ingenuo $\hat{\varepsilon}$ claramente falla para pequeños valores de n, y también en las situaciones de gran (in-)eficiencia. La técnica del MLE provee una solución robusta para la estimación de eficiencias: si la muestra contiene una variable x sensible a la eficiencia (es decir $\varepsilon(x)$), que sigue la PDF $P(x;\theta)$, entonces la inclusión de una nueva variable aleatoria discreta y bivariada $c = \{\rm yes, no\}$, nos da un modelo más elaborado:

$$P\left(x, c; \theta\right) \;=\; \delta(c-\text{yes})\varepsilon(x,\theta) + \delta(c-\text{no})\left[1-\varepsilon(x,\theta)\right] \;.$$

- la función $\varepsilon(x)$ ha de ser correctamente normalizada, para ser también una PDF
- la eficiencia ya no es un valor único, sino una función de x
- (y de otros parámetros θ necesarios para caracterizar su forma)

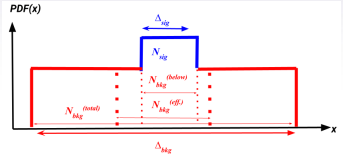
(`$ROOTSYS/tutorials/roofit/rf701_efficiencyfit.C`)



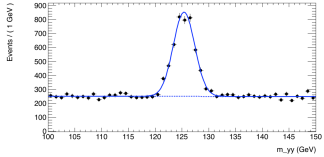
Sobre el fondo efectivo

De manera general, la zona que contiene la señal también está contaminada por fondo(s). El impacto de esos fondos se traduce en una *dilución* o disminución de la precisión en la medida de los parámetros de interés.
Ejercicio : Consideremos un escenario compuesto por dos especies: una señal, distribuida de manera uniforme en un intervalo $\Delta(\mathrm{sig})$, y un fondo distribuido de manera uniforme en un intervalo $\Delta(\mathrm{bkg})$ más amplio que cubre ambos lados del intervalo de señal ("sidebands").

- Generar una realización aleatoria, eligiendo valores particulares de los intervalos $\Delta(\mathrm{sig})$ y $\Delta(\mathrm{bkg})$, y del número de eventos de señal y fondo N_{sig} y N_{bkg}.
- El parámetro de interés es el número de eventos de señal N_{sig}. Estimar su valor y error $\hat{N}_s \pm \hat{\sigma_s}$ por verosimilitud máxima.
- Si el fondo fuera nulo, se tendría $\hat{\sigma_s} = \sqrt{N_s}$.
- Definir el "fondo efectivo" como el causante del aumento en el error, $\hat{\sigma_s} = \sqrt{N_s + N_{\mathrm{bkg}}^{\mathrm{eff}}}$.
- Comparar $N_{\mathrm{bkg}}^{\mathrm{eff}}$ al fondo "debajo" de la señal, $N_{\mathrm{bkg}}^{\mathrm{below}}$.
- Repitiendo el ejercicio para diferentes valores de N_{sig}, verificar que el fondo efectivo es siempre el mismo.
- Repitiendo el ejercicio para varios valores crecientes de $\Delta(\mathrm{bkg})$, verificar que el fondo efectivo tiende a $N_{\mathrm{bkg}}^{\mathrm{below}}$.
- Interpretar.

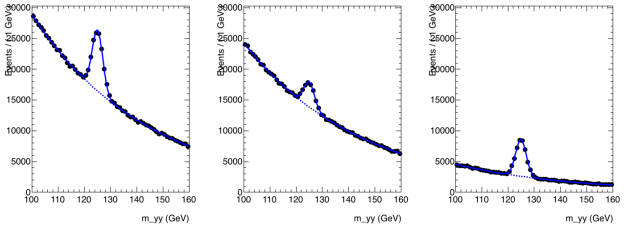


Ejercicio: Para una señal Gaussiana, realizar un ejercicio similar, con dos parámetros de interés adicionales: la posición del pico y su anchura.



Categorización

En ocasiones, la muestra de análisis puede descomponerse en dos o más submuestras ("categorías"), cada una de ellas con sus propias PDFs y purezas.
Cuando las características de cada especie son razonablemente diferentes, puede ser de interés descomponer la función de verosimilitud de tal manera que cada categoría utilice sus propias PDFs y purezas. El resultado combinado sobre un parámetro común de interés tendrá una significación superior a la que se obtendría de un análisis "inclusivo", es decir usando PDFs y purezas promedio sobre la muestra completa.
Ejemplo sencillo (y ejercicio): supongamos que la muestra $H \to \gamma\gamma$ se descompone en dos categorías: una "limpia" con excelente cociente señal/fondo, y una "sucia" en la que el fondo es ampliamente dominante. Si el parámetro de interés es la masa del Higgs, la ventaja de realizar un análisis en categorías puede ser significativo.

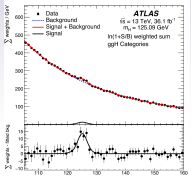


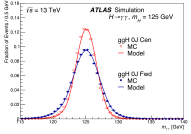
Aquí las dos categorías difieren en el cociente señal/fondo, grande para la "limpia", y pequeño para la "sucia". Otra posibilidad es aprovechar diferencias en resolución.

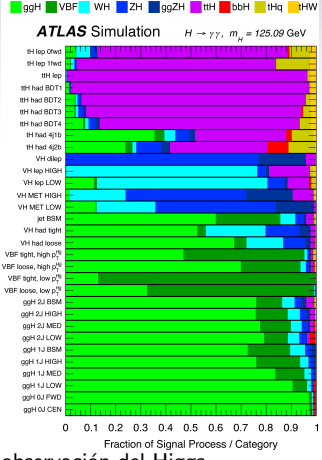
Error en el análisis inclusivo: $\sigma(m_H) = \pm 2.25\%$
Errores en las categorías sucia y limpia : $\pm 3.70\%$ y $\pm 1.85\%$, respectivamente.
Error en la combinación de ambas categorías: $\pm 1.72\%$! Equivale a aumentar en 70% la estadística inclusiva !

El uso de categorías en el análisis $H \to \gamma\gamma$

ATLAS separa su muestra de candidatos difotón en 31 categorías, definidas en función de varios criterios: el modo de producción del Higgs, diferencias en la resolución experimental en masa, diferencias en la pureza.







En el 2012, la categorización fue crucial para alcanzar los 5 σ de significancia en la observación del Higgs...

CAPITULO VI

INCERTIDUMBRES SISTEMATICAS

Incertidumbres sistemáticas (I)

En el MLE, la matriz de covariancia es el estimador de las incertidumbres estadísticas.
Pero otras fuentes de incertidumbre contribuyen también a diluir la precisión de una medida. En lenguaje de física, éstas a menudo se llaman "errores sistemáticos". Para discutir de estas incertidumbres en el contexto MLE, modificamos ligeramente la notación, y reescribimos la función de verosimilitud como sigue:

$$\mathcal{L}\left(\mu_1, \ldots, \mu_p \ , \ \theta_1, \ldots, \theta_k; \vec{x}\right) \ ,$$

donde hemos explícitamente separado el conjunto de parámetros de $\mathcal{L}$ en dos subconjuntos:

- los *parámetros de interés* $\mu_1, \ldots, \mu_p$, (POI en inglés), son las cantidades que deseamos estimar;
- los *parámetros de molestia* $\theta_1, \ldots, \theta_k$ (NP en inglés, por nuisance parameters) que representan fuentes potenciales de sesgos sistemáticos;

La lógica es la siguiente: si le asignásemos valores imprecisos o errados a algunos NPs, la formas de las PDFs resultantes pueden distorsionarse, y los estimadores de los POIs resultarán sesgados.
El error sistemático cuantifica la magnitud de esos sesgos.

- A menudo, la estimación de los sistemáticos es parte la dominante del trabajo de análisis de datos
- La calidad de un resultado científico releva en gran parte de la calidad de la estimación de la compomente sistemática de su incertidumbre
- Las medidas experimentales son a menudo clasificados en dos categorías:
  - "*dominadas por la estadística*", $\sigma_{\mathrm{STAT}} > \sigma_{\mathrm{SYST}}$
  - "*dominadas por los sistemáticos*", $\sigma_{\mathrm{STAT}} < \sigma_{\mathrm{SYST}}$

Las incertidumbres sistemáticas debidas a los NPs se clasifican a menudo en dos categorías:

- Errores de "Tipo-I" : nuestra muestra (u otras muestras de control) pueden en principio proveer información sobre el NP considerado, y la incertidumbre proveniente de este NP debe en principio disminuir con el tamaño de las muestras utilizadas.
- Errores de "Tipo-II", que provienen de suposiciones incorrectas del modelo (p.e. el uso de funciones inadecuadas para las PDFs), u aspectos mal controlados en los datos, como cambios en las condiciones de adquisición de las muestras, o la presencia de especies no tomadas en cuenta.
  - Algunas personas lo separan en dos categorías: los Tipo-II "de modelo" y los Tipo-III "de teoría" ... lo cual da "The Good, the Bad, the Ugly".

Algunos comentarios:

- Ciertos sistemáticos son una mezcla de ambos tipos: por ejemplo, si las muestras de control para controlar un NP no son completamente representativas de las propiedades de ese NP.
- Un NP enteramente de Tipo-I es una variable aleatoria, y por tanto se pueden estimar intervalos de confianza usando los métodos antes descritos.
- Los errores de Tipo-II son difíciles de manejar correctamente, y no siempre hay un procedimiento bien definido y consensual para estimar el impacto de esos errores. En los peores casos, no siempre se puede saber de qué manera un error de Tipo-II impacta los intervalos de confianza sobre los POIs.
- Frente a casos ambiguos, hay cierto consenso (parcial) en que es mejor tener estimaciones "conservadoras" de los sistemáticos, en particular cuando contribuyen de manera subdominante al "error budget". Pero en esas situaciones, las discusiones suelen ser controversiales, incluso amargas...

El método de *profile-likelihood*, (¿perfilar la verosimilitud?) permite manejar de manera elegante los sistemáticos de Tipo-I.
Consiste en asignar una verosimilitud específica a los NPs "perfileables", de tal manera que la $\mathcal{L}$ original se modifica en dos componentes:

$$\mathcal{L}(\mu,\theta) \;=\; \mathcal{L}_{\mu}(\mu,\theta)\,\mathcal{L}_{\theta}(\theta) \;\;.$$

Entonces, para un valor fijo de μ, $\mathcal{L}$ es maximizada con respecto al NP θ. Si se recorre una secuencia de valores de μ, $\mathcal{L}$ es una función que solamente depende de μ; si dice que la molestia ha sido "profiled-out".
Supongamos que el valor del NP es conocido: $\theta = (\theta_0 \pm \Delta_0)$.

- en ese caso la penalidad $\mathcal{L}_{\theta}(\theta)$ corresponde a una Gaussiana,

$$\mathcal{L}_{\theta}(\theta) = P_{\mathrm{Gauss}(\theta;\theta_0,\Delta_0)} = \frac{1}{\sqrt{2\pi}\Delta_0} e^{-(\theta-\theta_0)^2/2\Delta_0^2} \;,$$

- la maximización de $\mathcal{L}(\mu,\theta)$ "empujará" el estimador $\hat{\theta}$ del NP hacia su valor nominal θ_0
- el "pull' $(\hat{\theta}-\theta_0)/\Delta_0$ es un indicador de la coherencia del modelo; grandes valores (positivos o negativos) del *pull* indican una "tensión" proveniente de ese NP.

Como ejemplo en física, consideremos la medida de una sección eficaz de un proceso $\sigma\,(\text{inicial} \to \text{final})$.

- El número de eventos observados de ese tipo permite acceder a la sección eficaz:

$$N = \sigma \int \mathcal{L} dt \ .$$

- Si solamente una fracción de los procesos de ese tipo son detectados (p.e. debido a efectos de *aceptancia geométrica* del detector, u otras fuentes de ineficiencia), se requiere conocer la eficiencia de reconstrucción ε para convertir el número observados de procesos $\hat{N}$ en una medida de $\hat{\sigma}$:

$$\frac{\hat{N}}{\varepsilon} = \hat{\sigma} \int \mathcal{L} dt \ .$$

- La eficiencia ε es claramente un NP: independientemente de la precisión con la cual se haya medido $\hat{N}$,
  - un valor incorrecto de ε sesga directamente la medida de $\hat{\sigma}$
  - la incertidumbre sobre ε se propaga directamente sobre la incertidumbre sobre $\hat{\sigma}$
- Si $\hat{\varepsilon}$ puede estimarse sobre una muestra de control de calidad (por ejemplo, una simulación con gran estadística, o una muestra de control de alta pureza), el impacto de la molestia será controlado de manera coherente y su propagará correctamente sobre la medida del PIO.
- Un análisis *elegante* efectuará un *ajuste simultáneo a las muestras de señal y control*, de tal manera que
  - los valores y las incertidumbres de los NPs se extraen directamente de la covariancia del ajuste ML
  - lo cual asegura que las correlaciones con los POIs se propagan correctamente
  - y que los intervalos de confianza heredan de todas esas relaciones entre POIs y NPs.

Otro ejemplo de "profile-likelihood" : la búsqueda de una resonancia (un "bump") sobre un fondo uniforme.

- Si la fracción de señal (pureza) es muy pequeña, la anchura Γ del bump no puede estimarse directamente sobre la muestra (no hay manera de distinguir entre el fondo y una señal muy ancha),

$\rightarrow$ el valor de anchura a utilizar en la PDF de señal debe provenir de fuentes externas (p.e. una simulación detallada de la función de resolución del detector).

- Esa anchura es claramente un parámetro de molestia:

$\rightarrow$ si su valor se sobreestima, eso se traduciría en una subestimación del cociente señal/fondo, y por tanto en un aumento de la varianza de los POIs de la señal, así como tal vez posibles sesgos en sus valores centrales (p.e. si el fondo es asimétrico, el impacto será diferente a ambos lados del pico de la señal)

- Si las muestras de control permiten acceder a una estimación independiente de la anchura $\hat{\Gamma} \pm \hat{\sigma}_\Gamma$ de la anchura del pico de señal, esta información puede implementarse utilizando una PDF Gaussiana, de media $\hat{\Gamma}$ y de anchura $\hat{\sigma}_\Gamma$, en la componente $\mathcal{L}_\Gamma$ de la verosimilitud.

$\rightarrow$ Este término adicional cumple el rol de una penalidad en el MLE, y por lo tanto "empuja" los valores de Γ adentro del intervalo dado por $\pm\hat{\sigma}_\Gamma$.

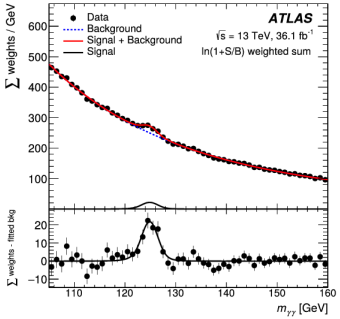
Incertidumbres sistemáticas (VI)

Un ejemplo práctico del "profile-likelihood" : el análisis del bosón de Higgs en el canal difotón en el LHC. Este canal se caracteriza por un cociente S/B extremadamente bajo.
Los parámetros de interés son :

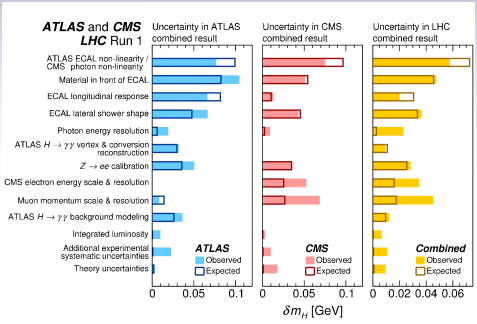
- la masa del Higgs,
- el producto $\sigma(pp \to H + X) \times \mathcal{BR}(H \to \gamma\gamma)$ de la sección eficaz de producción del Higgs en colisiones protón-proton, multiplicada por la tasa de decaimiento del Higgs en el canal difotón.

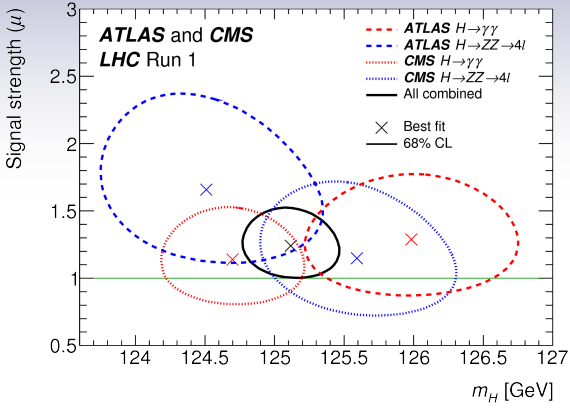
Mientras que la (larga) lista de parámetros de molestia incluye, entre otros:

- La eficiencia $\varepsilon(H \to \gamma\gamma)$ de reconstrucción de la señal, es decir el cociente entre el número de eventos $H(\to \gamma\gamma)_{\mathrm{RECO}}$ registrados, reconstruidos, identificados y seleccionados, y el número de eventos $(H \to \gamma\gamma)_{\mathrm{TRUE}}$ realmente producidos,
- Los parámetros necesarios para caracterizar la resolución en masa invariante difotón, es decir la distribución de la diferencia entre la masa reconstruida $m(\gamma\gamma)_{\mathrm{RECO}}$ y la verdadera masa $m(\gamma\gamma)_{\mathrm{TRUE}}$,
- Los parámetros necesarios para caracterizar la forma del fondo. La figura sugiere que una función uni-paramétrica es suficiente (p.e. una función exponencial decreciente), pero en práctica el asunto es más complicado...



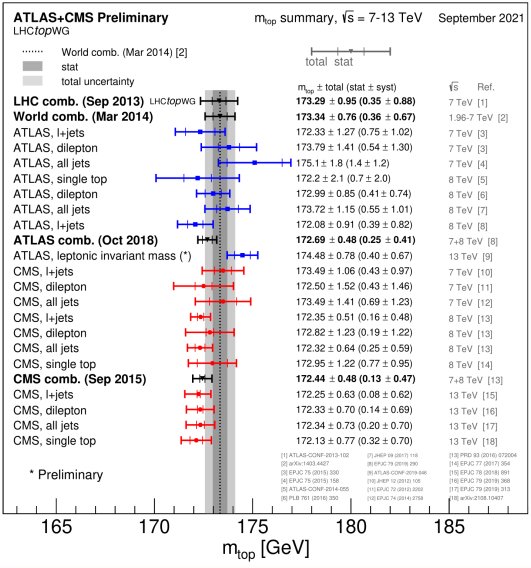
Incertidumbres sistemáticas (VI)





En la práctica, la estimación detallada de las incertidumbres sistemáticas concentra una gran parte del esfuerzo humano para producir un resultado de precisión.
Más aún en el caso de un análisis que combina informaciones provenientes de varias fuentes (p.e. aquí la combinación de ATLAS y CMS) : hay que identificar las fuentes comunes de sistemáticos, las correlaciones, etc...

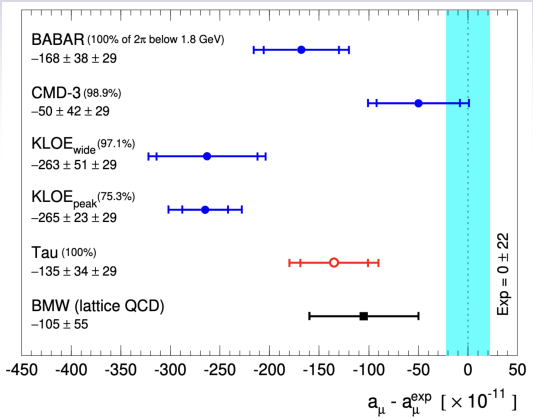
Reportar las incertidumbres sistemáticas: ejemplos

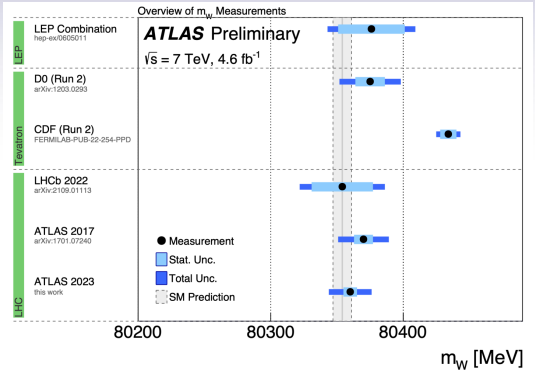


Source	Uncertainty [%]
Statistical uncertainty	7.5
Systematic uncertainties	6.4
Background modelling (spurious signal)	3.8
Photon energy scale & resolution	3.6
Photon selection efficiency	2.6
Luminosity	1.8
Pile-up modelling	1.4
Trigger efficiency	1.0
Theoretical modelling	0.4
Total	9.8

Objetivo :
Dar tanta información detallada como sea posible...

Reportar las incertidumbres sistemáticas: ejemplos





CAPITULO VII

CONTRASTE DE HIPOTESIS

Contraste de hipótesis

Las discusiones pasadas se centraron mayoritariamente en extraer información numérica a partir de muestras de datos: efectuar mediciones de parámetros, y reportar el resultado de esas mediciones bajo forma de

- valores centrales e incertidumbres, en particular cuando se trata de un sólo parámetro de interés;
- matrices de covarianza, en particular para medidas simultáneas de varios parámetros, y si las correlaciones no pueden ser despreciadas;
- perfiles completos de verosimilitud, cuando la aproximación "parabólica" no es suficiente;

La etapa siguiente en un análisis es producir información cualitativa a partir de los datos disponibles: se habla entonces de efectuar un *contraste estadístico de hipótesis* (hypothesis testing en inglés).
La herramiento cuantitativa para declarar el acuerdo entre una hipótesis y las observaciones (los datos) se llama un *estadístico de prueba*. El resultado de una prueba se da en términos de una "p-*value*": la probabilidad, bajo la hipótesis en consideración, de observar un estadístico de prueba similar o "peor" que el observado en la muestra.
En términos intuitivos: suponemos que los datos son una realización aleatoria de la hipótesis sometida a prueba, y comparamos cuantitativamente el estadístico observado sobre los datos con el ensamble de realizaciones aleatorias de estadísticos. Es lo que se llama la *interpretación frecuentista* de la estadística:

- "dada mi hipótesis, ¿cuál es la probabilidad de observar una muestra de datos menos representativa de mi hipótesis que la que realmente observé?"
- equivale a cuantificar la probabilidad matemática

$$\mathcal{P}\left(\text{datos} \mid \vec{\mu}\right) \ ,$$

donde $\vec{\mu}$ son nuestros parámetros de interés POI.

Existe una interpretación diferente de la estadística, llamada *bayesiana*, que busca resolver el *problema inverso*:

- "dada mi muestra de datos, ¿cuál es la probabilidad que mi hipótesis sea verdadera?"
- equivale a cuantificar una probabilidad matemática diferente a la frecuentista:

$$\mathcal{P}\left(\vec{\mu} \mid \text{datos}\right) = \mathcal{P}\left(\vec{\mu}\right) \times \frac{\mathcal{P}\left(\text{datos} \mid \vec{\mu}\right)}{\mathcal{P}\left(\text{datos}\right)} \ ,$$

donde

- $\mathcal{P}\left(\text{datos} \mid \vec{\mu}\right)$ es la probabilidad frecuentista (p.e. extraída de un análisis de verosimilitud máxima),
- $\mathcal{P}\left(\vec{\mu}\right)$ es la *probabilidad previa*, o *prior* bayesiano,
- $\mathcal{P}\left(\text{datos}\right)$ es un coeficiente de normalización sin importancia.
- $\mathcal{P}\left(\vec{\mu} \mid \text{datos}\right)$ es la *probabilidad subjetiva* o "*grado de creencia*" (¡!)

Las diferencias entre ambas interpretaciones son profundas, y tocan a la esencia del proyecto de inferencia científica. Ahora, los debates entre adeptos de una u otra interpretación son en ocasiones amargos y poco estimulantes...

> "Bayesians address the questions everyone is interested in by using assumptions that no one believes. Frequentists use impeccable logic to deal with an issue that is of no interest to anyone."
> (L. Lyons)

El test del χ^2

Dado un conjunto de n medidas independientes x_i con varianzas σ_i^2, y un conjunto de predicciones μ_i, se define un test estadístico llamado χ^2 de la manera siguiente:

$$\chi^2 \;=\; \sum_{\mathrm{i}=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \;\; .$$

El test de χ^2 es una variable aleatoria que, como vimos previamente, sigue la distribución de $P_{\chi^2}(x;n)$ para n *grados de libertad*. Su valor de expectación es n y su varianza $2n$.
Por ello uno espera que el valor de χ^2 observado sobre una muestra no debe alejarse mucho del número de grados de libertad, y por tanto ese valor permite sondear el acuerdo entre la observación y la predicción.
Para ser más preciso, se espera que 68% de los tests se encuentren contenidos dentro de un intervalo $n \pm \sqrt{2n}$.
La p-value, o probabilidad de observar un test con valores mayores viene dada por

$$p \;=\; \int_{\chi^2}^{+\infty} dx \; P_{\chi^2}\,(x;n) \;\; .$$

Intuitivamente hablando, uno debe sospechar de pruebas que arrojen pequeñas p-values, dado que esas podrían indicar un problema. Este puede provenir del lado de las predicciones, o reflejar la calidad de los datos registrados, o ser solamente resultado de la "mala suerte".
La interpretación de las p-values (i.e. para decidir qué valores son demasiado pequeños o suficientemente grandes) es un tópico importante, que requiere un marco de análisis que apunte a separar las partes objetivas y subjetivas.

Propiedades generales del contraste de hipótesis (I)

Consideremos dos hipótesis mutuamente excluyentes, $\mathcal{H}_0$ y $\mathcal{H}_1$, que proveen ambas una descripción de un proceso aleatorio, y del cual extraemos una muestra de datos. El contraste de hipótesis apunta a evaluar:

- cuán robusta es la *hipótesis nula* $\mathcal{H}_0$, en describir los datos, y
- cuán incompatible con esos mismos datos es la *hipótesis alternativa* $\mathcal{H}_1$.

La dinámica del contraste de hipótesis se puede resumir así:

- construir un estadístico de prueba q, una función que reduce una muestra a un valor numérico único;
- definir un intervalo de confianza $W \to [q_{\rm lo} : q_{\rm hi}]$;
- medir $\hat{q}$ sobre la muestra en estudio;
- si $\hat{q}$ está contenido en el intervalo W, se declara que la hipótesis nula es aceptada, y rechazada en caso contrario.

Caso particular : física de altas energías

Allí, un ejemplo común concierne la búsqueda de una señal (aún) desconocida, que implica dos casos:

- en la *lógica de descubrimiento*, la hipótesis nula corresponde al escenario *background-only*, mientras que la hipótesis alternativa sería *signal-plus-background*: "¿cuán probable es que una fluctuación del fondo explique el exceso que estoy observando?";
- en la *lógica de exclusión*, las dos hipótesis se invierten: "¿cuál es la cantidad máxima de señal compatible con mi observación?".

Para caracterizar el resultado de la secuencia antes descrita, se definen dos criterios:

- se incurre en un "Error de Tipo-I" si $\mathcal{H}_0$ es rechazada aún siendo cierta ("falso negativo");
- se incurre en un "Error de Tipo-II" si $\mathcal{H}_0$ es aceptada aún siendo falsa ("falso positivo").

Las tasa de errores de Tipo-I y Tipo-II se llaman usualmente α y β respectivamente, y se determinan por integración de las densidades de probabilidad asociadas a las hipótesis $\mathcal{H}_0$ y $\mathcal{H}_1$ sobre el intervalo W:

$$\begin{aligned} 1-\alpha &= \int_W dq\mathcal{P}\left(q|H_0\right) \ , \\ \beta &= \int_W dq\mathcal{P}\left(q|H_1\right) \ . \end{aligned}$$

La tasa α es llamada *tamaño del contraste* (*size of the test* en inglés), puesto que fijar α determina el tamaño del intervalo W. De manera análoga, $1-\beta$ es llamado *potencia del contraste* (*power*).
Juntos, tamaño y potencia caracterizan el performance de un estadístico: el lema de Neyman-Pearson afirma que *a tamaño fijo*, el estadístico óptimo viene dado por el cociente de verosimilitudes q_λ:

$$q_\lambda\left(\text{data}\right) = \frac{\mathcal{L}\left(\text{data}|H_0\right)}{\mathcal{L}\left(\text{data}|H_1\right)} \ .$$

En la práctica, las distribuciones de $\mathcal{H}_0$ y $\mathcal{H}_1$ son obtenidas a partir de simulaciones o muestras de control. De esas distribuciones se se obtiene la distribución esperada del estadístico q_λ, y la p-value observada se obtiene al integrarla con respecto al valor observado de q_λ.

La significación del estadístico viene dada por su p-value,

$$p \;=\; \int_{\hat{q}}^{+\infty} dq \mathcal{P}\left(q|H_0\right) \;.$$

que es a menudo reportado en unidades de "sigmas",

$$p \;=\; \int_{n\sigma}^{+\infty} dz \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \;=\; 1 - \frac{1}{2}\mathrm{erf}\left(\frac{n}{\sqrt{2}}\right) \;,$$

de manera que por ejemplo una p-value $p < 0.0228$ se reporta como un "efecto a dos sigma".
Igualmente común es reportar la p-value bajo forma de un intervalo de confianza (C.L.).
Esta definición de la p-value es clara y sin ambiguedades. Pero su interpretación es parcialmente subjetiva: la conveniencia de un *umbral de tolerancia* puede depender del tipo de hipótesis bajo prueba, o de hábitos en cada disciplina.
En física de altas energías:

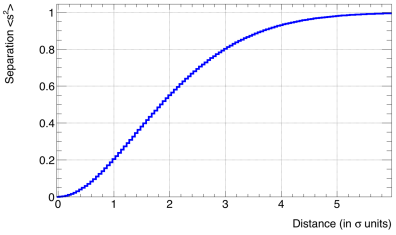
- en la *lógica de exclusión*, el umbral se sitúa a 95% C.L. para declarar la exclusión de la hipótesis de *señal-más-fondo*;
- en la *lógica de descubrimiento*, el umbral se sitúa a tres sigma ($p < 1.35 \times 10^{-3}$) sobre la hipótesis *solamente-fondo* para afirmar que hay "evidencia";
- y un umbral a cinco sigma ($p < 2.87 \times 10^{-7}$) es requerido para afirmar que hay "observación".

Separación

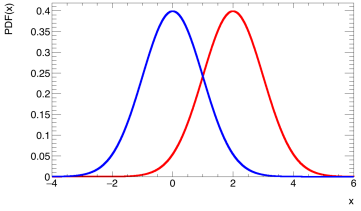
Las definiciones generales de *tamaño* y *potencia* pueden ser complementadas con otras definiciones más sencillas. Por ejemplo la "separación", que se determina a partir de las PDFs de las dos especies (señal $\mathcal{S}$ y fondo $\mathcal{B}$) para las cuales se quiere cuantificar el poder de discriminación:

$$\langle s^2 \rangle = \frac{1}{2} \int d\vec{x} \frac{[\mathcal{S}(\vec{x}) - \mathcal{B}(\vec{x})]^2}{\mathcal{S}(\vec{x}) + \mathcal{B}(\vec{x})} \ .$$

Por construcción, $0 \leq \langle s^2 \rangle \leq 1$, valores límites que corresponden a los casos extremos en que el poder de discriminación es nulo (señal y fondo son imposibles de distinguir) o total (no hay ningún solapamiento entre las dos especies).



La gráfica a la izquierda muestra la separación entre dos Gaussianas de misma anchura σ, en función de la distancia entre sus picos (en unidades de σ). Ejemplo para una distancia de 2σ:



(a menudo se habla de "separación a $n\sigma$")

En la HEP experimental, hay tradición de definir por consenso la elección de los estadísticos de prueba, para simplificar las combinaciones de resultados de diferentes experimentos, de manera que las componentes relacionadas con los detectores (específicas a cada experimento) se factoricen con respecto a los observables físicos (que son en principio universales).
Ejemplo: en el contexto de la búsqueda del Bosón de Higgs del Modelo Estándar, los cuatro experimentos en el LEP (ALEPH, DELPHI, OPAL, L3) decideron describir sus datos utilizando las siguientes verosimilitudes:

$$\mathcal{L}\left(H_1\right) = \prod_{a=1}^{N_{\rm ch}} \mathcal{P}_{\rm Poisson}\left(n_a, s_a + b_a\right) \prod_{j=1}^{n_a} \frac{s_a \mathcal{S}_a(\vec{x_j}) + b_a \mathcal{B}_a(\vec{x_j})}{s_a + b_a} \ ,$$

$$\mathcal{L}\left(H_0\right) = \prod_{a=1}^{N_{\rm ch}} \mathcal{P}_{\rm Poisson}\left(n_a, b_a\right) \prod_{j=1}^{n_a} \mathcal{B}_a(\vec{x_j}) \ .$$

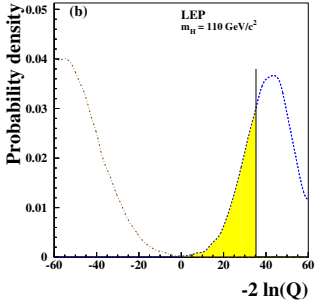
donde $N_{\rm ch}$ es el número de "canales de decaimiento" del Higgs estudiados, n_a es el número observado de candidatos en cada canal a, $\mathcal{S}_a$ y s_a ($\mathcal{B}_a$ y b_a) son las PDFs y los números de eventos para las especies de señal (fondo) de cada canal. El estatístico de prueba λ, derivado de un cociente de verosimilitudes, es

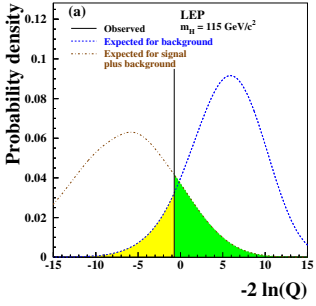
$$\lambda = -2 \ln Q \ , \ \text{con} \ Q = \frac{\mathcal{L}\left(H_1\right)}{\mathcal{L}\left(H_0\right)} \ ;$$

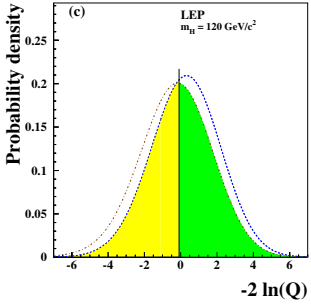
de manera que valores positivos de λ favorezcan un escenario "background-like", y valores negativos estén más tono con un escenario "señal-más-fondo"; valores cercanos a cero indicando una sensibilidad pobre para distinguir entre ambos.

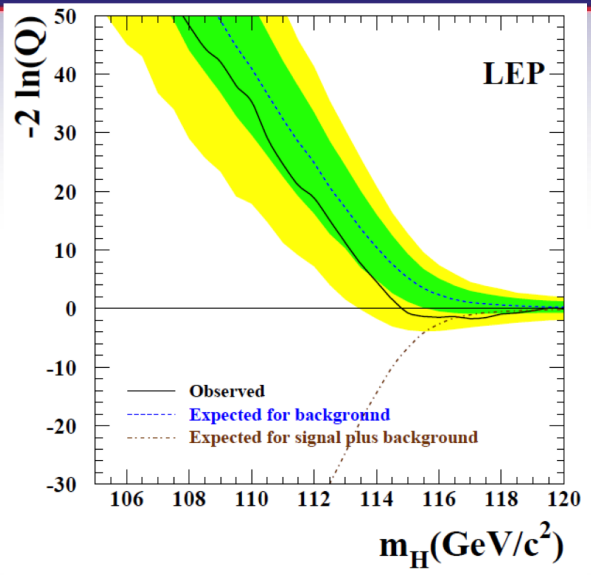
- Bajo la hipótesis "background-only", $\mathrm{CL}(b)$ es la probabilidad de tener $-2\ln Q$ más pequeño que el observado (amarillo);
- bajo la hipótesis "señal-más-fondo", $\mathrm{CL}(s+b)$ es la probabilidad de tener $-2\ln Q$ más grande que el observado (verde).

Las figuras abajo muestran, para tres hipótesis diferentes de la masa del Higgs, los valores de $-2\ln Q$ obtenidos al combinar los resultados de los cuatro experimentos LEP. También se muestran las distribuciones de $\mathrm{CL}(s+b)$ y $1-\mathrm{CL}(b)$.








LEP
-2 ln(Q)
50
40
30
20
10
0
-10
-20
-30
106
108
110
112
114
116
118
120
Observed
Expected for backgrounnd
Expected for signal plus background
$m_H(GeV/c^2)$

El estimador modificado $\mathrm{CL}(s)$

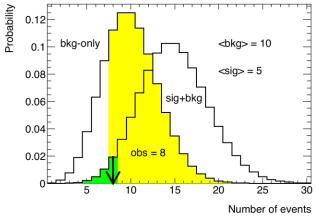
Tomemos un experimento de conteo de eventos: esperamos 10 eventos de tipo fondo, y 5 eventos de tipo señal

- pero... observamos 8 eventos en total...
- lo razonable es suponer es que tanto señal como fondo experimentaron una fluctuación negativa...
- pero en la interpretación estándar, ¡ se hubiera asignado una exclusión a 95% C.L. !
- (ello incluso si el experimento no tiene sensibilidad alguna a la señal)

Para evitar esta situación, se define un nivel de confianza modificado, $\mathrm{CL}(s)$, definido como

$$\mathrm{CL}(s) \;=\; \frac{\mathrm{CL}(s+b)}{1-\mathrm{CL}(b)} \;.$$

que si bien *stricto sensu* no es una p-value (un cociente de probabilidades no es una probabilidad) por lo menos tiene la propiedad de proteger contra fluctuaciones negativas del fondo.



(cuidado con las convenciones de color amarillo/verde...)

- $\mathrm{CL}(s+b) = 3.7\%$
- $1-\mathrm{CL}(b) = 33\%$
- $\mathrm{CL}(s) = 11\%$

No sin controversia, el estimador $\mathrm{CL}(s)$ ha sido sin embargo adoptado por varias colaboraciones internacionales, incluyendo los experimentos del Tevatron (Fermilab) y ATLAS y CMS en el LHC.

Los cocientes de perfil de verosimilitud

Los experimentos ATLAS y CMS usan como estadístico de prueba el llamado *cociente de perfil de verosimilitud* (profiled likelihood ratio) definido así:

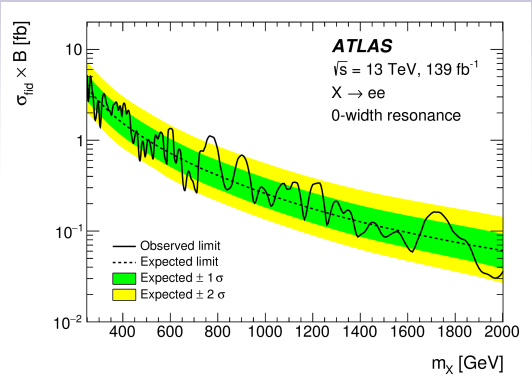
$$\tilde{q}_\mu(\mu) \;=\; -2\ln\frac{\mathcal{L}\left(\mu,\hat{\hat{\theta}}\right)}{\mathcal{L}\left(\hat{\mu},\hat{\theta}\right)}\;,\;\text{con}\;0\le\hat{\mu}\le\mu\;,$$

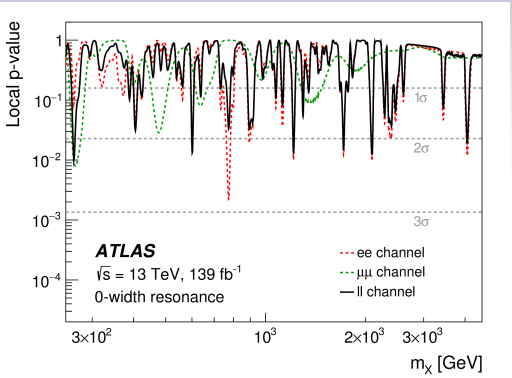
- el parámetro de interés es $\mu=\sigma/\sigma_{\rm SM}$, la "intensidad de señal": la tasa de conteo de candidatos de señal comparada a la predicción (p.e. la sección eficaz de producción del Higgs vs. la predicción teórica),
- $\hat{\hat{\theta}}$ son los valores de los NPs obtenidos en un ajuste a intensidad de señal μ fija,
- $\hat{\mu}$ y $\hat{\theta}$ son los valores ajustados cuando tanto μ como los NPs son libres en el ajuste,
- (el límite inferior en $0\le\hat{\mu}\le\mu$ es para asegurar tener una intensidad de señal positiva, y el límite superior es para evitar que una fluctuación hacia arriba no desfavorezca la hipótesis de señal).

Para un valor observado del estadístico $\hat{\tilde{q}}_\mu$, las p-values asignadas a las hipótesis de signal-plus-background y background-only, $p(s+b)$ y $p(b)$, son

$$p(s+b)=\int_{\hat{\tilde{q}}_\mu}^{\infty}dqP\left(q;\mu=\hat{\mu},\hat{\theta}\right)\;,\;1-p(b)=\int_{\hat{\tilde{q}}_\mu}^{\infty}dqP\left(q;\mu=0,\hat{\hat{\theta}}\right)\;.$$

- los resultados de exclusión se muestran bajo forma de un "Brazil-plot",
- los resultados de observación bajo forma de un "local-p_0-plot".





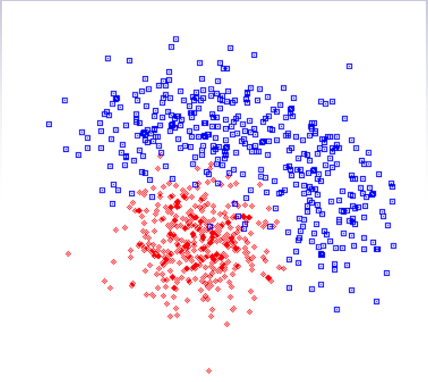
Cuando varias “regiones” estadísticamente independientes son probadas en un mismo análisis (aquí la búsqueda de otras posibles resonancias en dileptón), hay que tomar en cuenta el Look-Elsewhere-Effect (o trials factors en la literatura) para evaluar la p-value...

(Nota: a alta energía la resolución en impulso es superior para los electrones que para los muones, por eso hay más “regiones” en el canal $X \to e^+e^-$ que para el canal $X \to \mu^+\mu^-$...)

CAPITULO VIII

ANALISIS MULTIDIMENSIONAL

Análisis multidimensional (I)

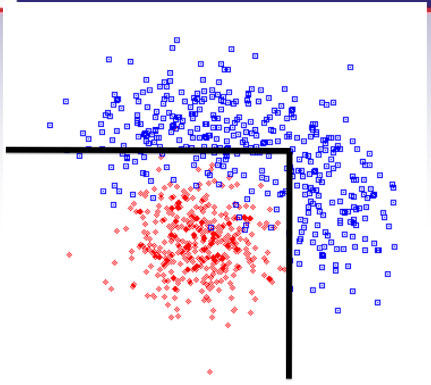


Puntos azules : muestra de control finita, distribuída como el fondo
Puntos rojos : muestra de control finita, distribuída como la señal

A menudo, hay grandes regiones del espacio de la muestra en las cuales los fondos son ampliamente dominantes, o donde la densidad de la señal es nula o despeciable.
Si se reduce la muestra a subconjuntos "enriquecidos en señal" del espacio completo, la pérdida de información puede ser mínima, y otras ventajas pueden compensar esas posibles pérdidas:

- para muestras multidimensionales, puede ser difícil caracterizar las formas de las PDFs en las regiones de baja densidad de eventos
- el reducir el tamaño de la muestra puede aliviar el consumo de memoria y CPU en las partes numéricas del análisis (p.e. la minimización)

Análisis multidimensional (II)



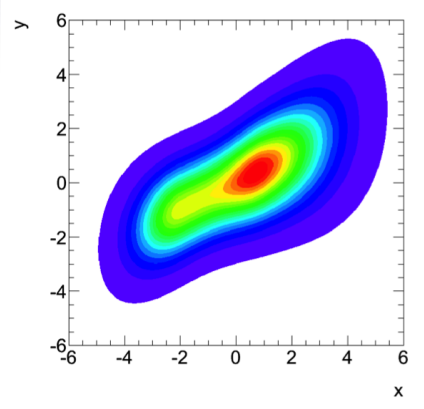
El método más sencillo de reducción de una muestra es restringiendo las variables, una a una, a intervalos finitos.
En la práctica, esas selecciones "cut-based" aparecen a varios niveles de la definición del espacio de muestreo: umbrales en las decisiones en línea (triggers), filtros a varios niveles posteriores del proceso de adquisición, eliminación de datos a partir de criterios de calidad...
Pero ya en etapas más avanzadas del análisis de datos, esas selecciones "accept-reject" convienen ser reemplazadas por procedimientos más sofisticados. Estos son llamados de manera genérica *técnicas multivariadas*.

Las líneas negras indican una selección "cut-based", definida de manera de conservar cerca de 100% de la señal.

(para ser más precisos, la selección es 100% eficaz sobre la muestra de control de señal, pero la caracterización precisa de la eficiencia de selección require de estimar la densidad de probabilidad de la señal fuera de la región seleccionada (y para caracterizar con precisión el nivel de fondo se require estimar la densidad de probabilidad del fondo dentro de ella)

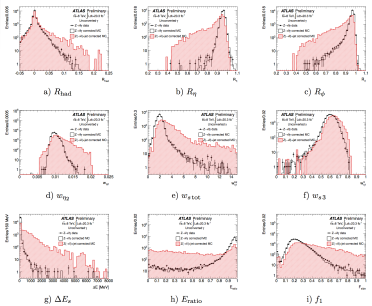
Análisis multidimensional (III)

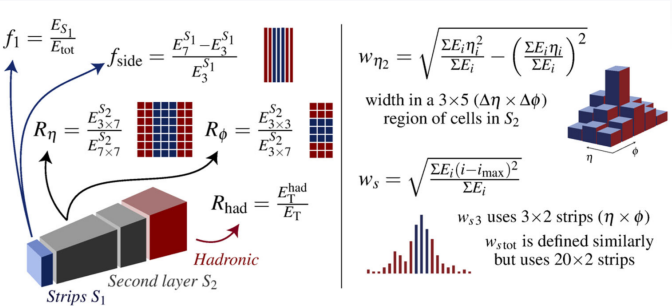
Consideremos un conjunto de n variables aleatorias $\vec{x} = \{x_1, x_2 \ldots, x_n\}$. Si todas las variables son no-correlacionadas, Las PDFs n-dimensionales están completamente determinadas por el producto directo de sus n PDFs uni-dimensionales.
si algunas de las variables están correlacionadas, y si sus patrones de correlacíon son completamente lineales, es posible definir un nuevo conjunto de variables $\vec{y}$, que son combinaciones lineales de $\vec{x}$, obtenidas diagonalizando el inverso de la matriz de covarianza.



En ocasiones, cuando los patrones de correlación son no-lineales, es tal vez posible en algunos casos definir una descripción analítica: por ejemplo, el perfil de correlación que incluye una (ligera) componente no lineal representado aquí, fue producido con el paquete `RooFit` aplicando la opción `Conditional` en `RooProdPdf` para producir un producto de PDFs: la anchura de y varía de manera no-lineal con x.
En la práctica, ésta solución elegante no puede extenderse fácilmente a más de dos dimensiones, y no hay garantía que se puedan reproducir patrones no lineales complicados.
Frente a tales escenarios, un protocolo de *reducción dimensional* puede a menudo producir resultados más efectivos que un intento de descripción analítica de correlaciones.

Un escenario típico para considerar la reducción dimensional es cuando varias variables arrastran en gran parte información común (y por tanto exhiben correlaciones fuertes), pero contienen también algumos elementos (diluídos pero importantes) de información independiente.
Un ejemplo: la caracterización de cascadas en calorímetros segmentados. Las señales depositadas en celdas vecinas están fuertemente correlacionadas (tienen un origen común), pero permiten reconstruir detalles precisos del desarrollo de la cascada. Esas correlaciones son utilizadas para caracterizar las "formas de cascadas":





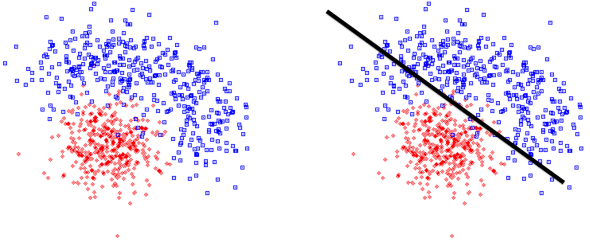
El resultado de combinar las informaciones de todas las "shower shape variables" es la *clasificación* del candidato en dos especies: cascadas electromagnéticas (fotones, electrones) vs. cascadas hadrónicas ("jets").

Discriminantes lineales

El algoritmo más sencillo de reducción dimensional es el discriminante de Fisher: es una función lineal de las variables, con coeficientes definidos a partir de un criterio de optimización de la separación entre especies (c.f. próxima lámina).

- Fisher es un caso particular de los llamados PCA (principal component analysis);
- un análisis MLE es también un PCA: reduce un problema multidimensional a un problema en 1 dimensión: el comportamiento del estadístico de test $\lambda = \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)}$.

Por construcción, los discriminantes lineales son óptimos para variables multinormales, por tanto con correlaciones perfectamente lineales.



Ejercicio:

- dos variables aleatorias, x_1, x_2;
- dos especies: señal y fondo;
- la PDF de la señal es una bigaussiana, con $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $c_{12} = 0$;
- la PDF del fondo es una bigaussiana, con $\mu_1 = \mu_2 = 1$, $\sigma_1 = \sigma_2 = 1$, $c_{12} = 0.5$;

→ Determinar los coeficientes de Fisher

→ Representar gráficamente las distribuciones de $\mathcal{F}$ para las dos especies

→ Evaluar la separación $\langle s^2 \rangle$ de $\mathcal{F}$

Tenemos dos especies, "señal" y "fondo":

- distinguimos las especies con un índice $c = 1, 2$.

Tenemos n variables discriminantes $\vec{x} = x_1, x_2, \cdots x_n$:

- caracterizamos las PDFs de cada especie usando muestras de control compuestas por N_c eventos cada una, formando las llamadas "n-tuplas" $\mathbf{x}^c$.

Algunas definiciones necesarias:

- la "tupla-media" $\overline{\mathbf{x}}^c$ de cada especie, y la tupla especie-promediada $\overline{\mathbf{x}}$:

$$\overline{\mathbf{x}}^c = \frac{1}{N_c} \sum_{k=1}^{N_c} \mathbf{x}_k^c \quad , \quad \overline{\mathbf{x}} = \frac{1}{N_1 + N_2} \sum_{c=1}^{2} \sum_{k=1}^{N_c} \mathbf{x}_k^c \ ,$$

- las matrices de covarianza *intra-especie* W_{ij} ("within") e *inter-especie* B_{ij} ("between"):

$$W_{ij} = \frac{1}{N_1 + N_2} \sum_{c=1}^{2} \sum_{k=1}^{N_c} \left(x_i^{ck} - \overline{x}_i^c \right) \left(x_j^{ck} - \overline{x}_j^c \right) \quad , \quad B_{ij} = \frac{1}{N_1 + N_2} \sum_{c=1}^{2} N_c \left(\overline{x}_i^c - \overline{x}_i \right) \left(\overline{x}_j^c - \overline{x}_j \right) \ ,$$

- la matriz de covariancia especie-promediada $T_{ij} = B_{ij} + W_{ij}$. Los índices $i, j = 1, 2, \cdots, n$.

El discriminante de Fisher (II)

El *discriminante de Fisher* $\mathcal{F}$, es una combinación lineal de las variables aleatorias $\vec{x} = x_1, x_2, \cdots x_n$ dada por

$$\mathcal{F} = f_0 + \sum_{i=1}^{n} f_i x_i \ .$$

donde los f_i son los llamados coeficientes de Fisher-Mahalanobis, asignados a cada variable $i = 1, 2, \cdots, n$:

$$f_i = \frac{\sqrt{N_1 N_2}}{N_1 + N_2} \sum_{i=1}^{n} C_{ij}^{-1} \left(\overline{x}_i^1 - \overline{x}_i^2 \right) \quad , \quad f_0 = \sum_{i=1}^{n} f_i \left(\overline{x}_i^1 + \overline{x}_i^2 \right) \ ,$$

(aquí la matriz C_{ij}^{-1} coresponde a W_{ij}^{-1} para Fisher, y a T_{ij}^{-1} para Mahalanobis)
El discriminante lineal de Fisher *reduce* la dimensionalidad de nuestro problema:

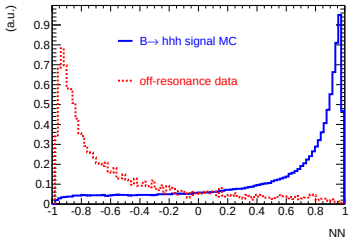
- teníamos n variables aleatorias discriminantes $\vec{x} = x_1, x_2, \cdots x_n$,
- terminamos con una única variable aleatoria discriminante $\mathcal{F}$,
- $\mathcal{F}$ es una combinación lineal de las variables iniciales.

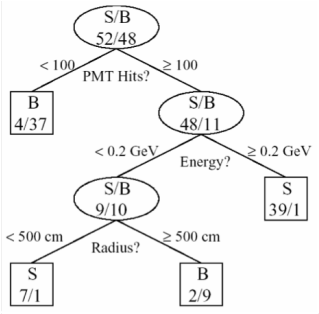
De manera general, el discriminante de Fisher

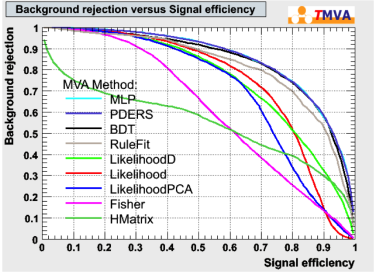
- produce una discriminación *óptima* para variables multinormales
  - (que comportan por tanto únicamente correlaciones perfectamente lineales)
- una discriminación subóptima pero elevada para distribuciones con correlaciones únicamente lineales
- puede ser claramente subóptimo, e incluso fallar totalmente, en caso de correlaciones altamente no-lineales.

Discriminantes no lineales (I)

Para tratar correlaciones no lineales y con perfiles complejos, existe una variedad de técnicas y herramientas.
El paquete TMVA es una implementación popular en HEP de varios algoritmos de reducción dimensional: además de una biblioteca de discriminantes lineales y basados en likelihood, incluye métodos de entrenamiento y prueba con redes de neuronas artificiales y árboles de decisión (boosted decision trees), que forman parte de los más utilizados en HEP.
Idea general: un análisis multivariado utiliza una colección de variables de entrada, que se combinan de manera optimizada, a partir de un algoritmo "entrenado" sobre dos muestras independientes (correspondientes a señal y fondo), con dos protocolos: entrenamiento (training) y prueba (test). El performance del algoritmo entrenado se evalúa sobre las muestras de prueba, para evitar el efecto de "over-training".



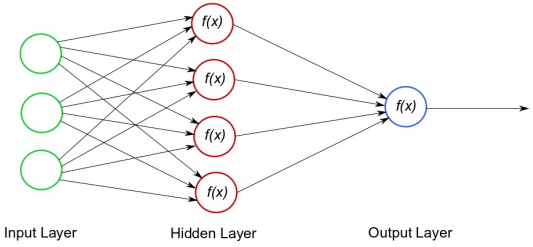


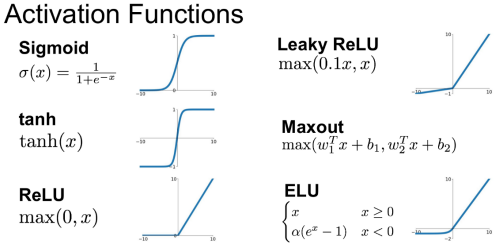


"ROC-curve":
Receiver Operating Characteristic

Funcionamiento genérico de una red neuronal :

- Una "capa" de celdas (o neuronas) de entrada (en nuestro caso serían las variables aleatorias que queremos combinar) ;
- una secuencia de una o varias capas escondidas de celdas, que combinan linearmente capas anteriores y aplican funciones de activación ;

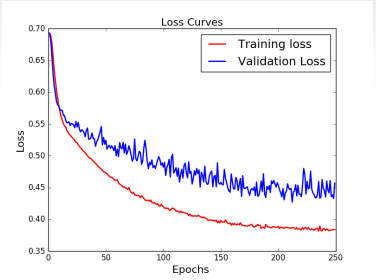




Discriminantes no lineales (III)

Funcionamiento genérico de una red neuronal :

- los pesos de las combinaciones lineales son optimizados utilizando una "función de pérdida" (loss function, que sería en nuestro caso el equivalente de la función a minimizar, i.e. χ^2 o $-\ln \mathcal{L}$) ;
- una capa de salida (en nuestro caso sería la variable que reduce la dimensionalidad de nuestro sistema).



Ejemplos de Loss functions :

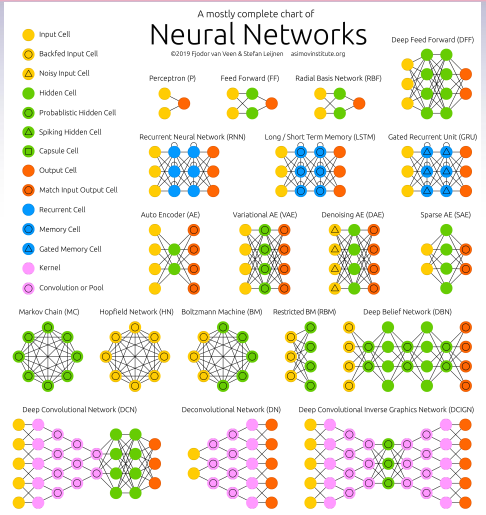
$$\mathcal{L} = \sum_{i=1}^{N} (\mathrm{score}_i - \mathrm{truth}_i)^2 \ ,$$

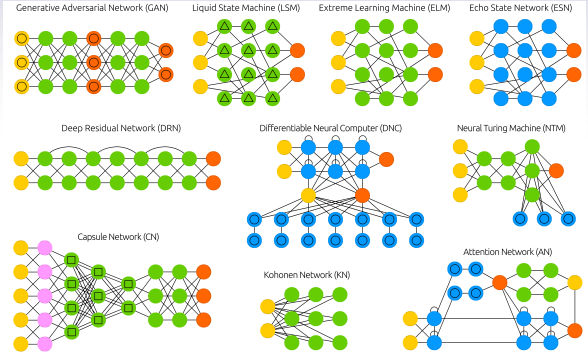
$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} q_i \log\left(p(q_i)\right) + (1 - q_i) \log\left(1 - p(q_i)\right) \ ,$$

la primera se usa para el proceso de *regresión*, y consiste en asegurar que el score reproduce la información verdadera truth y la segunda se llama *entropía cruzada* y se usa para el proceso de *clasificación*.

- asegurar la calidad de las muestras de entrenamiento y prueba ;
- ¡ evitar el "over-fitting" !

Discriminantes no lineales (IV)





(tomado de `https://www.asimovinstitute.org/neural-network-zoo/`)

En resumen, un *análisis multivariado* produce una reducción dimensional, proyectando un espacio de n variables aleatorias $\vec{x} = \{x_1, x_2 \ldots, x_n\}$, sobre una variable final $\mathcal{Z}$. Esta variable puede ser

- una combinación lineal de las $\vec{x}$ (discriminante de Fisher) ;
- un estadístico de prueba más elaborado (por ejemplo el cociente de verosimilitudes $\lambda = \mathcal{L}(H_1)/\mathcal{L}(H_0)$) ;
- la salida de un algoritmo de *Machine Learning*. Ejemplos: *Multilayer Perceptron* (red de neuronas artificiales en sus múltiples apelaciones, tipo NN, CNN, ANN, GNN...), *Boosted Decision Tree* (árbol de decisiones), *algoritmos genéticos*, y otros más.

De manera general, los algoritmos de ML obtienen *performances* superiores que los algoritmos más sencillos, pero ... performances con respecto a la calidad de las muestras de control, o con respecto al algoritmo de entrenamiento. del análisis se podrá preferir la robustez al performance, o se privilegiará una combinación entre ambas... Dos citas sacadas de WikiPedia:

- Trees can be very non-robust. A small change in the training data can result in a large change in the tree and consequently the final predictions.
- Decision-tree learners can create over-complex trees that do not generalize well from the training data. (This is known as overfitting.)

Una frase de conclusión : este curso trató sobre la estadística para la inferencia científica. Otros usos muy importantes (de formalismo similar) son la *teoría de la decisión* y la *clasificación*
Estos tópicos serán tratados desde varias perspectivas complementarias en el curso siguiente: “Tópicos avanzados en ciencia de datos”.
¡ Feliz continuación !