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:

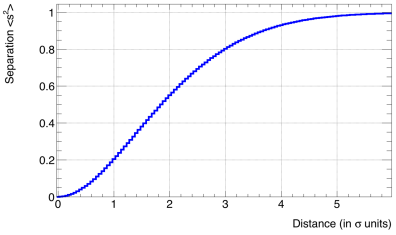
- en la *lógica de exclusión*, el umbral se sitúa a 95 % C.L. para declarar la exclusión de la hipótesis de *señal-más-fondo*;
- en la *lógica de descubrimiento*, el umbral se sitúa a tres sigma ($p < 1{,}35 \times 10^{-3}$) sobre la hipótesis *solamente-fondo* para afirmar que hay "evidencia";
- y un umbral a cinco sigma ($p < 2{,}87 \times 10^{-7}$) es requerido para afirmar que hay "observación".

Separación

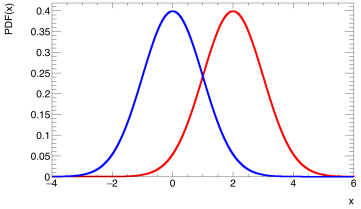
Las definiciones generales de *tamaño* y *potencia* pueden ser complementadas con otras definiciones más sencillas. Por ejemplo la "separación", que se determina a partir de las PDFs de las dos especies (señal $\mathcal{S}$ y fondo $\mathcal{B}$) para las cuales se quiere cuantificar el poder de discriminación:

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

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 nσ")

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:

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

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

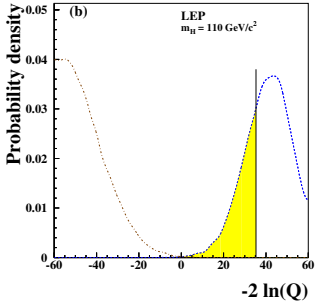
donde $N_{\rm ch}$ es el número de "canales de decaimiento" del Higgs estudiados, n_a es el número observado de candidatos en cada canal a, $\mathcal{S}_a$ y s_a ($\mathcal{B}_a$ y b_a) son las PDFs y los números de eventos para las especies de señal (fondo) de cada canal. El estatístico de prueba λ,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

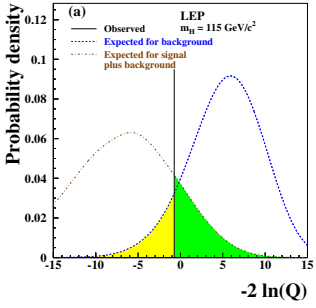
$$\lambda = -2 \ln Q \ , \ \text{con} \ Q = \frac{\mathcal{L}\left(H_1\right)}{\mathcal{L}\left(H_0\right)} \ ;$$

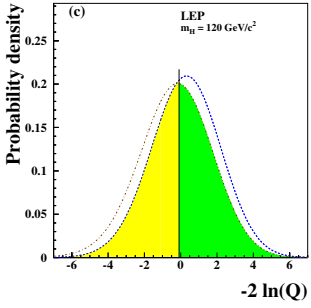
de manera que valores positivos de λ favorezcan un escenario "background-like", y valores negativos estén más tono con un escenario "señal-más-fondo"; valores cercanos a cero indicando una sensibilidad pobre para distinguir entre ambos.

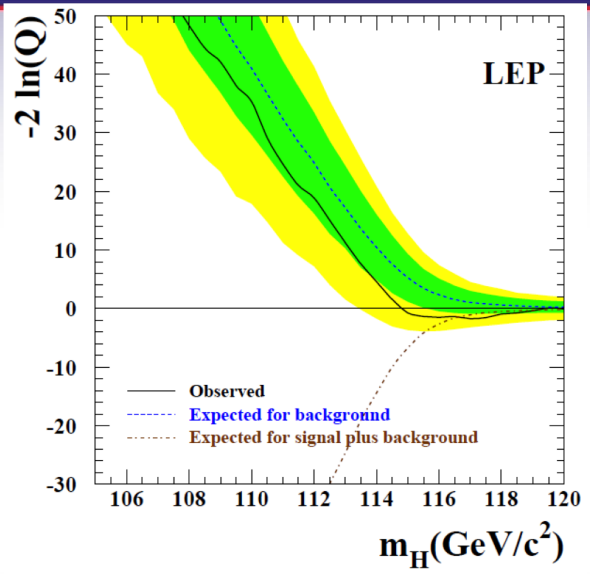
- Bajo la hipótesis "background-only", $\mathrm{CL}(b)$ es la probabilidad de tener $-2\ln Q$ más pequeño que el observado (amarillo);
- bajo la hipótesis "señal-más-fondo", $\mathrm{CL}(s+b)$ es la probabilidad de tener $-2\ln Q$ más grande que el observado (verde).

Las figuras bajo muestra, 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
$m_H(GeV/c^2)$
Observed
Expected for background
Expected for signal plus background

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

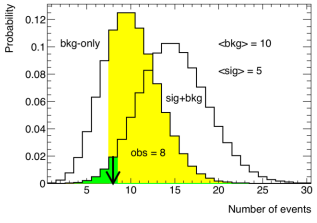
Tomemos un experimento de conteo de eventos: esperamos 10 eventos de tipo fondo, y 5 eventos de tipo señal

- 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:

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

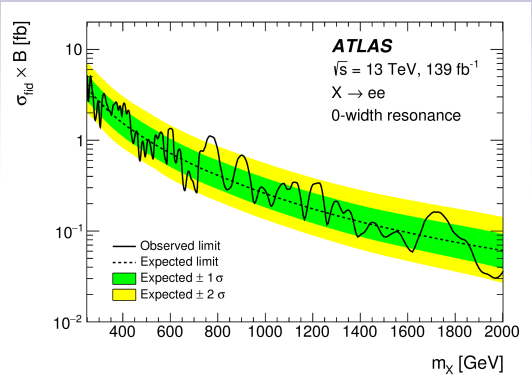
- el parámetro de interés es $\mu=\sigma/\sigma_{\rm SM}$, la "intensidad de señal": la tasa de conteo de candidatos de señal comparada a la predicción (p.e. la sección eficaz de producción del Higgs vs. la predicción teórica),
- $\hat{\hat{\theta}}$ son los valores de los NPs obtenidos en un ajuste 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).

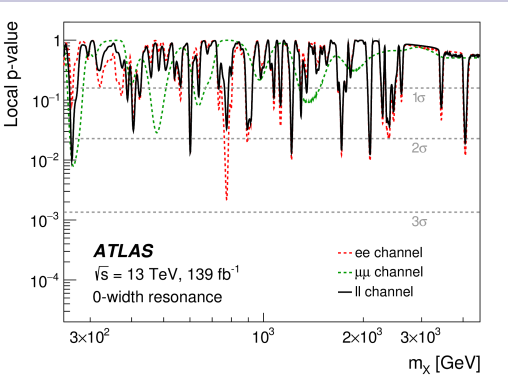
Para un valor observado del estadístico $\hat{\tilde{q}}_\mu$, las p-values asignadas a las hipótesis de signal-plus-background y background-only, $p(s+b)$ y $p(b)$, son

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

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

El "look-elsewhere-effect"





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 $X \to \mu^+\mu^-$...)