Асимптотические формулы для оценки статистической значимости в экспериментах на LHC

МГУ им. М. В. Ломоносова, физический факультет

Горин Даниил

Статистический анализ эксперимента

Discovery: $H_0 = H_b$, $H_1 = H_{s+b}$ **Upper limits:** $H_0 = H_{s+b}$, $H_1 = H_b$

Чтобы численно оценить результат эксперимента, рассчитывается **p-значение** и оценивается **значимость** $Z = \Phi^{-1}(1-p)$

Полезно также оценивать **чувствительность** эксперимента, рассчитывая **ожидаемую значимость**, которую можно было бы получить при данном измерении в предположении различных гипотез

На основании леммы Неймана-Пирсона удобно в качестве статистики для оценки значимости эксперимента взять отношение $\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$ **функций правдоподобия**

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

n и m - количество событий соответственно в сигнальной и контрольной областях,
s - количество сигнальных событий, b - количество фоновых событий
$\boldsymbol{\mu}$ - сила сигнала, $\boldsymbol{\theta}$ - дополнительные параметры
$\hat{\hat{\boldsymbol{\theta}}}$ - параметры, максимизирующие L для заданного μ
$\hat{\mu}$ и $\hat{\boldsymbol{\theta}}$ - параметры, совместно максимизирующие L

Статистики для discovery и upper limits

Discovery

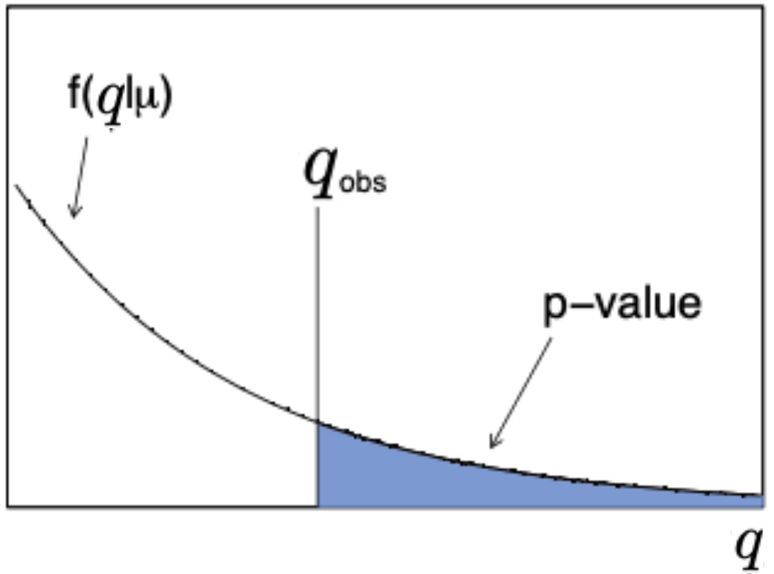
$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\, dq_0$$

Upper limits

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$



Помимо $f(q_0|0)$ и $f(q_\mu|\mu)$, для расчета ожидаемой значимости нужно знать $f(q_\mu|\mu')$, где $\mu \neq \mu'$

G. Cowan, K. Cranmer, E. Gross and O. Vitells, “Asymptotic formulae for likelihood-based tests of new physics,” Eur. Phys. J. C 71, 1554 (2011) [arXiv:1007.1727].

Аппроксимация распределения λ(μ)

Как показал Вальд*: $-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N})$, где $\hat{\mu}$ распределено по Гауссу со средним μ' и дисперсией σ^2

Пренебрегая асимптотическим членом, можно показать, что $t_\mu = -2\ln\lambda(\mu)$ подчиняется **нецентральному распределению хи-квадрат** с одной степенью свободы:

$$f(t_\mu;\Lambda) = \frac{1}{2\sqrt{t_\mu}}\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}+\sqrt{\Lambda}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}-\sqrt{\Lambda}\right)^2\right)\right] ,$$

где $\Lambda = \frac{(\mu-\mu')^2}{\sigma^2}$ - параметр нецентральности

В случае $\mu' = \mu$ распределение $-2\ln\lambda(\mu)$ стремится к (центральному) распределению хи-квадрат с одной степенью свободы, как было показано Вилксом**

*A. Wald, Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large, Transactions of the American Mathematical Society, Vol. 54, No. 3 (Nov., 1943), pp. 426-482.

**S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Statist. 9 (1938) 60-2.

Распределения q_0 и q_μ и значимость

$$q_0 = \begin{cases} \hat{\mu}^2/\sigma^2 & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

$$\downarrow$$

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

$$\downarrow$$

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

$$\downarrow$$

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right)$$

$$\downarrow$$

$$p_0 = 1 - F(q_0|0)$$

$$\downarrow$$

$$\boxed{Z_0 = \Phi^{-1}(1 - p_0) = \sqrt{q_0}}$$

$$q_\mu = \begin{cases} \frac{(\mu - \hat{\mu})^2}{\sigma^2} & \hat{\mu} < \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

$$\downarrow$$

$$f(q_\mu|\mu') = \Phi\left(\frac{\mu' - \mu}{\sigma}\right)\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{q_\mu} - \frac{\mu - \mu'}{\sigma}\right)^2\right]$$

$$\downarrow$$

$$F(q_\mu|\mu') = \Phi\left(\sqrt{q_\mu} - \frac{\mu - \mu'}{\sigma}\right)$$

$$\downarrow$$

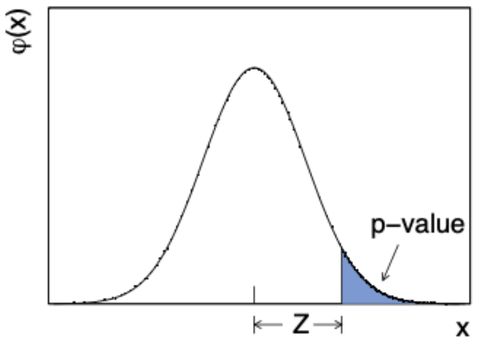
$$F(q_\mu|\mu) = \Phi\left(\sqrt{q_\mu}\right)$$

$$\downarrow$$

$$p_\mu = 1 - F(q_\mu|\mu)$$

$$\downarrow$$

$$\boxed{Z_\mu = \Phi^{-1}(1 - p_\mu) = \sqrt{q_\mu}}$$



G. Cowan, K. Cranmer, E. Gross and O. Vitells, “Asymptotic formulae for likelihood-based tests of new physics,” Eur. Phys. J. C 71, 1554 (2011) [arXiv:1007.1727].

Азимовский набор данных

Азимовский набор данных - это искусственный набор данных, в котором устранены статистические флуктуации, то есть:

$$n_{i,\mathrm{A}} = E[n_i] = \mu' s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta})$$
$$m_{i,\mathrm{A}} = E[m_i] = u_i(\boldsymbol{\theta})$$

Таким образом, для азимовского набора данных можно считать, что:

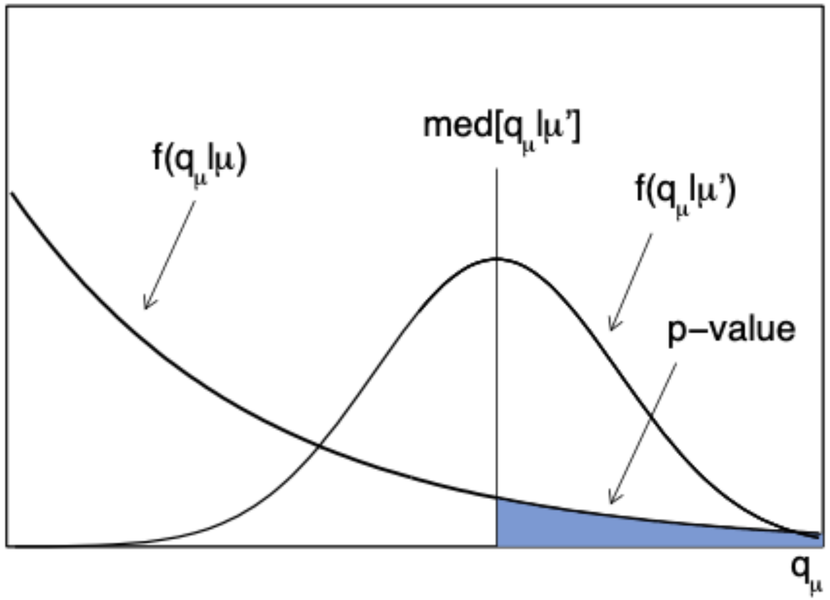
$$\hat{\mu} = \mu'$$
$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$$

Тогда в силу монотонности статистик q_0 и q_μ по $\hat{\mu}$:

$$med[q_0] = q_0(med[\hat{\mu}]) = q_0(\mu') = q_{0,A}$$
$$med[q_\mu] = q_\mu(med[\hat{\mu}]) = q_\mu(\mu') = q_{\mu,A}$$

Чувствительность эксперимента

Для вычисления чувствительности эксперимента интересна значимость не на одном наборе данных, а ожидаемая (медианная) значимость, с которой можно отвергать различные значения μ



Чувствительность эксперимента может быть оценена через p-значение, соответствующее медианному q_μ , в предположении силы сигнала μ'

В силу монотонности p-значения это соответствует медианному p-значению q_μ в предположении силы сигнала μ'

Так как Z монотонна по q, можно получить выражения для ожидаемой значимости с помощью азимовского набора данных:

$$\mathrm{med}[Z_0|\mu'] = \sqrt{med[q_0]} = \sqrt{q_{0,\mathrm{A}}}$$

$$\mathrm{med}[Z_\mu|0] = \sqrt{med[q_\mu]} = \sqrt{q_{\mu,\mathrm{A}}}$$

Постановка задачи

В случае одного сигнального и одного контрольного измерений:

$$L(\mu, b) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \frac{(\tau b)^m}{m!} e^{-\tau b}$$

Для расчета статистик необходимы следующие выражения:

$$\hat{\mu} = \frac{n - m/\tau}{s}, \quad \hat{b} = \frac{m}{\tau}$$

$$\hat{\hat{b}} = \frac{n + m - (1+\tau)\mu s}{2(1+\tau)} + \left[\frac{(n + m - (1+\tau)\mu s)^2 + 4(1+\tau) m \mu s}{4(1+\tau)^2} \right]^{1/2}$$

где $\hat{\hat{b}}$ – параметр, максимизирующий L для заданного μ , а $\hat{\mu}$ и $\hat{b}$ - параметры, совместно максимизирующие L

G. Cowan, K. Cranmer, E. Gross and O. Vitells, “Asymptotic formulae for likelihood-based tests of new physics,” Eur. Phys. J. C 71, 1554 (2011) [arXiv:1007.1727].

Вывод формул для Z_{disc}

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$, где $\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$, $\boldsymbol{\theta} = b$

Подставляя L и выражения для $\hat{\mu}$ и $\hat{b}$ и учитывая, что $\hat{\hat{b}}(0) = \frac{n+m}{1+\tau}$, получим:

$$q_0 = \begin{cases} -2\left[n\ln\left(\frac{n+m}{(1+\tau)n}\right) + m\ln\left(\frac{\tau(n+m)}{(1+\tau)m}\right)\right] & , n \geq \frac{m}{\tau} \\ 0 \ , n < \frac{m}{\tau} \end{cases}$$

Вывод формул для Z_{disc}

Для оценки ожидаемой значимости воспользуемся выражением $\mathrm{med}[Z_0|\mu'] = \sqrt{q_{0,A}}$ и азимовским набором данных с силой сигнала $\mu' = 1$, то есть сделаем замену $n = s + b$ и $m = \tau b$:

$$Z_{disc} = med[Z_0|1] = \sqrt{-2\left[(s+b)\ln\left(\frac{s+(1+\tau)b}{(1+\tau)(s+b)}\right) + \tau b \ln\left(1+\frac{s}{(1+\tau)b}\right)\right]}$$

Перепишем эту формулу в терминах систематической ошибки фона $\delta = \frac{\sigma_b}{b} = \frac{1}{\sqrt{\tau b}}$:

$$Z_{disc} = \sqrt{2\left[(s+b)\ln\left(\frac{(s+b)(1+\delta^2 b)}{b+\delta^2 b(s+b)}\right) - \frac{1}{\delta^2}\ln\left(1+\delta^2\frac{s}{1+\delta^2 b}\right)\right]}$$

В случае точного фона $\delta \to 0$:

$$Z_{disc} = \sqrt{2\left[(s+b)\ln\left(1+\frac{s}{b}\right) - s\right]}$$

A. Senol, H. Denizli, C. Helveci,"Sensitivity of anomalous quartic gauge couplings via tri-photon production at FCC-hh", Nucl.Phys.B 998 (2024) [arXiv:2303.14805]

Вывод формул для Z_{excl}

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$, где $\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$, $\boldsymbol{\theta} = b$

Для exclusion положим $\mu = 1$

Подставляя L и выражения для $\hat{\mu}$ и $\hat{b}$, получим:

$$q_1 = \begin{cases} -2\left[n\ln\left(\frac{s+\hat{\hat{b}}(1)}{n}\right) + m\ln\left(\frac{\tau\hat{\hat{b}}(1)}{m}\right) - (s+\hat{\hat{b}}(1)) + n - \tau\hat{\hat{b}}(1) + m\right] & , n - m\tau \le s \\ 0 \ , n - m\tau > s \end{cases}$$

Вывод формул для Z_{excl}

Для оценки ожидаемой значимости воспользуемся выражением $\mathrm{med}[Z_\mu|0] = \sqrt{q_{\mu,\mathrm{A}}}$ и азимовским набором данных с силой сигнала $\mu' = 0$, то есть сделаем замену $n = b$ и $m = \tau b$, и подставим выражение для $\hat{\hat{b}}(1)$

Записывая формулу в терминах систематической ошибки фона $\delta = \dfrac{1}{\sqrt{\tau b}}$:

$$Z_{excl} = \sqrt{2\left[s - b\ln\left(\frac{b+s+x}{2b}\right) - \frac{1}{\delta^2}\ln\left(\frac{b-s+x}{2b}\right)\right] - (b+s-x)\left(1+\frac{1}{\delta^2 b}\right)}$$

, где $x = \sqrt{(b+s)^2 - \dfrac{4\delta^2 b^2 s}{\delta^2 b + 1}}$

В случае точного фона $\delta \to 0$:

$$Z_{excl} = \sqrt{2\left[s - b\ln\left(1+\frac{s}{b}\right)\right]}$$

A. Senol, H. Denizli, C. Helveci,"Sensitivity of anomalous quartic gauge couplings via tri-photon production at FCC-hh", Nucl.Phys.B 998 (2024) [arXiv:2303.14805]

Обобщенные формулы

Представленные ранее выводы можно обобщить на случай N сигнальных и M контрольных измерений. В таком случае функция правдоподобия:

$$L(\mu, \boldsymbol{B}) = \prod_{i=1}^{N} \left\{ P\left(n_i \middle| \mu s_i + \sum_{j=1}^{M} b_{ji} \right) \right\} \cdot \prod_{j=1}^{M} \left\{ P\left(m_j \middle| \sum_{j'=1}^{M} \tau_{jj'}^{i'} b_{j'i'} \right) \right\}$$

Отсюда можно получить:

$$Z_{disc} = \sqrt{-2 \cdot \left(\sum_{i=1}^{N} \left\{ n_i \cdot \ln\left(\frac{\sum_{j=1}^{M} \hat{\hat{b}}_{ji}}{n_i} \right) + n_i - \sum_{j=1}^{M} \hat{\hat{b}}_{ji} \right\} + \sum_{j=1}^{M} \left\{ m_j \cdot \ln\left(\frac{[\boldsymbol{\tau}^1 \cdot \hat{\hat{\boldsymbol{B}}}]_{j1}}{[\boldsymbol{\tau}^1 \cdot \boldsymbol{B}]_{j1}} \right) + [\boldsymbol{\tau}^1 \cdot (\boldsymbol{B} - \hat{\hat{\boldsymbol{B}}})]_{j1} \right\} \right)},$$

где $\hat{\hat{\boldsymbol{B}}}$ удовлетворяет уравнению:

$$\sum_{i=1}^{N} \frac{\tau_{1\ell}^1}{\tau_{1\ell}^i} \cdot \left(\frac{n_i}{\sum_{j=1}^{M} \hat{\hat{b}}_{ji}} - 1 \right) + \sum_{j=1}^{M} \tau_{j\ell}^1 \cdot \left(\frac{m_j}{[\boldsymbol{\tau}^1 \cdot \hat{\hat{\boldsymbol{B}}}]_{j1}} - 1 \right) = 0\,;\ \ell = 1, \ldots, M$$

$\boldsymbol{\tau}^i$ - матрица, сопоставляющая фоны из i-го сигнального региона с контрольными регионами (τ_{jk}^i сопоставляет k-й фон в i-ом сигнальном регионе с j-м контрольным регионом)

$\boldsymbol{B}$ - матрица фонов в сигнальных регионах (b_{ji} соответствует j-му фону в i-ом сигнальном регионе)

M. J. Basso, “Generalized asymptotic formulae for estimating statistical significance in high energy physics analyses”, J. Phys. G 49 025001 (2022), [arXiv:2102.04275]

Выводы

Самостоятельно выведены и тем самым подтверждены асимптотические формулы для оценки статистической значимости **discovery** и **exclusion**

Эти формулы предоставляют простой и эффективный способ быстро делать оценки статистической значимости различных экспериментов, а также оценивать значения параметров интереса на заданном уровне уверенности

Вместо генерации событий методом Монте-Карло, что может отнимать много времени и ресурсов, и расчета значимости непосредственно из статистических распределений, можно воспользоваться полученными асимптотическими формулами