De ontdekking van het Higgsdeeltje: versimpelde statistiek

In de ontdekking van het Higgsdeeltje is zo goed mogelijk gebruik gemaakt van alle data, en is het model om het signaal en de achtergronden te beschrijven zeer gecompliceerd (met honderden "nuisance parameters"). Dit maakt het onmogelijk om een enigszins realistische weergave te maken van deze statistische analyse.

Niettemin is het vrij eenvoudig om de essentie van het probleem uit te leggen aan de hand van een zogenaamd "counting experiment", waar alleen het totale aantal interacties beschouwd wordt. Na zekere selectie resulteert dan een verwacht signaal s alsmede een schatting van het verwachte aantal achtergrond-interacties b (met een geassocieerde onzekerheid σ; we zullen aannemen dat deze onzekerheid voldoende klein is dat met de mogelijkheid dat de werkelijke achtergrond $b' < 0$ geen rekening gehouden hoeft te worden). Het waargenomen aantal interacties n volgt een Poissonverdeling,

$$P(n|\mu) = \frac{e^{-\mu}\mu^n}{n!}$$

voor gegeven verwachting μ.

Bij wijze van voorbeeld kan een van de vervalskanalen van het Higgsdeeltje beschouwd worden ($H \to ZZ$), waar $n = 13$, $s = 5,3$, $b = 4,85$, en $\sigma = 0,29$. (Dit is al een benadering van het combineren van drie vervalskanalen; meer details kunnen gevonden worden in de eigenlijke – en lange – publicatie: http://arxiv.org/pdf/1207.7214v2 .)

1. Stel dat $\sigma = 0$. In dit geval is het eenvoudig een *test statistic* op te stellen die optimaal onderscheidt tussen de hypotheses van enerzijds alleen de achtergrond en anderzijds de som van achtergrond en signaal. Met enkele sets (niet al te hoge) waarden van n, s en b kan dan de p-waarde voor iedere set vastgesteld worden. Hiervoor is het nodig om in ieder geval vele *toy experiments* te genereren, waarin steeds andere waarden van n gevonden worden. (Ook bovenstaande getallen kunnen als voorbeeld genomen worden.)

2. Stel dat $\sigma = b/10$. In dit geval zijn de hypotheses die getest worden niet meer enkelvoudig, en wordt de statistische behandeling ingewikkelder. De methode die in de voordracht is behandeld kan dan nog steeds gebruikt worden, met ofwel *toy experiments* of de asymptotische benadering; de geldigheid van de asymptotische benadering kan dan geverifieerd worden.

   In het verleden is, in plaats van de volledig frequentistische benadering die ik in mijn voordracht bespreek, voor dit soort problemen ook vaak een hybride benadering gebruikt, waar op een Bayesiaanse manier gemarginaliseerd wordt over de werkelijke achtergrond b'. In dit geval is de "prior" verdeling van b' nodig; vaak wordt een uniforme verdeling genomen, met $0 < b' < b_{\max}$, voor zekere (maar tamelijk willekeurige!) $b_{\max}$. Ook deze benadering kan expliciet uitgewerkt worden.