Projekty z analizy danych
Leszek Adamczyk (leszek.adamczyk@agh.edu.pl)

Proszę przesłać swoje programy (Jupyter notebook z kodem, opisem, rysunkami i wynikami) na adres mailowy: leszek.adamczyk@agh.edu.pl najpóźniej we wtorek o godzinie 8:00. Temat maila ma mieć format: Projekt 1: Imię Nazwisko. Nie spodziewam aby aby projekt był w pełni skończony.
Przydatne informacje:

- serwer jupyter: telesfor.fis.agh.edu.pl:8000
- opis danych ATLAS'a : http://opendata.atlas.cern/release/2020/documentation
- przykładowy program: ATLAS_OpenData_13TeV__analysis_examplecpp_Hyy_channel.ipynb

Projekt 1: Pomiar sygnału w obecności tła

Tematem projektu jest analiza potencjalnej produkcji nowego neutralnego bozonu pośredniczącego Z' poprzez jego poszukiwanie w danych w postaci rezonansu $Z' \to \mu^+\mu^-(e^+e^-)$ pojawiającego się w obecności tła przewidywanego przez Model Standardowy (SM). Dane rzeczywiste oraz symulacyjne znajdują się w /mnt/sdb1/opendata/2lep. Każdy z Państwa zajmie się analizą produkcji bozonu o jednej z potencjalnych mas $M_{Z'} = 2, 3, 4, 5$ TeV w jednym z kanałów $\mu^+\mu^-(e^+e^-)$. Selekcja przypadków jak w dokumentacji "Example of physics analysis: the case of the SM Z-boson production in the two-lepton final state" ale w obszarze ± 200 GeV wokół hipotetycznej masy $M_{Z'}$. Próbki MC sygnałowego to 301220-301223 (301215-301218). Próbki tła SM znajdują się w tabeli (top-quark, W/Z (+jets), dibosons)

- Proszę narysować rozkłady masy niezmienniczej pary leptonów w analizowanym kanale (dane + MC (tło+sygnał)).
- Na podstawie MC sygnałowego proszę wyznaczyć przedział (symetryczny) wokół $M_{Z'}$ zawierający 90% sygnału (zakres sygnałowy).
- Jeśli bozon o masie $M_{Z'}$ jest obecny w danych, to spodziewamy się nadmiaru przypadków w zakresie sygnałowym w porównaniu do spodziewanej ilości zakładającej tylko tło SM.
- Miarą nadmiaru (odchylenia) obserwowanej liczby przypadków od spodziewanej przy hipotezie tylko tła jest prawdopodobieństwo ($p-$value) obserwacji takiej samej lub większej ilości przypadków niż obserwowanej w danych (n_{obs}) zakładając tylko tło SM:
$$p = \sum_{n=n_{obs}}^{\infty} f(n;\mu_b)$$
gdzie $f(n;\mu_b)$ jest rozkładem Poissona o idealnie znanej wartości oczekiwanej tła μ_b
- Powszechną praktyką jest konwersja $p-$value na obserwowane znaczenie nadmiaru (significance Z_{obs}) odpowiadające standardowemu rozkładowi normalnemu:
$$\int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt = p$$
$$1 - \Phi(Z) = p$$
$$Z = \Phi^{-1}(1-p)$$
gdzie Φ jest dystrybuantą standardowego rozkładu normalnego. Z jest wyrażone w jednostkach odchylenia standardowego (σ).

- Z_{exp} lub spodziewane odchylenie sygnału od hipotezy jedynie tła jest zdefiniowane jako znaczenie związanie z wartością oczekiwaną ($n_{exp} = \mu_b + \mu_s$) ilości przypadków przy hipotezie tło+sygnał. Taka wielkość jest miarą mocy analizy do separacji tych dwóch hipotez.

- Proszę wyznaczyć: μ_s, μ_b oraz n_{obs} dla wyznaczonego wcześniej zakresu sygnałowego.

- Kolejnym krokiem będzie wyznaczenie spodziewanych wartości $p-$value i Z_{exp}. W przypadku idealnie znanej wartości μ_b ($\Delta\mu_b = 0$) można to zrobić na dwa sposoby:

  – bezpośrednio z definicji
  $$p = \sum_{n=\mu_b+\mu_s}^{\infty} f(n; \mu_b)$$
  – za pomocą metody MC poprzez generację dużej liczby (np. miliona) pseudo-eksperymentów dla hipotez tylko tła oraz tło+sygnał. W praktyce oznacza to wyznaczenie rozkładu ilości przypadków generowanych z rozkładów Poissona o wartości oczekiwanej μ_b oraz $\mu_b + \mu_s$

  Proszę użyć obu metod. W przypadku metody MC proszę narysować rozkłady ilości przypadków dla obu hipotez z zaznaczeniem obszaru odpowiadającego $p-$value.

- Jeśli sygnał jest obecny to wartości $p-$value oraz Z obserwowane w rzeczywistym eksperymencie będą fluktuować wokół wyznaczonych wcześniej wartości spodziewanych. Proszę wyznaczyć wartości obserwowane zastępując wartość spodziewaną $n_{exp} = \mu_b + \mu_s$ wartością obserwowaną w danych n_{obs}.

- Otrzymane wartości obserwowaną i spodziewane Z proszę porównać z estymatą
  $$\hat{Z}_{obs} = \sqrt{2n_{obs}\ln(1 - \mu_s/\mu_b) - 2\mu_s}$$
  gdzie n_{obs} jest obserwowaną ilością przypadków w eksperymencie w którym spodziewamy się μ_b ilości tła i μ_s sygnału. Z_{exp} otrzymamy podstawiając $n_{obs} = \mu_b + \mu_s$

- Wybór zakresu sygnałowego był przypadkowy. Proszę zoptymalizować wybór zakresu sygnałowego poprzez maksymalizację Z_{exp} powtarzając analizę dla np. 10 wartości szerokości zakresu sygnałowego tak aby zaobserwować maksimum. Ta wartość optymalna jest kompromisem pomiędzy wydajnością selekcji przypadków oraz odrzucaniem tła. Taka optymalizacja powinna być przeprowadzona tylko na podstawie spodziewanych rozkładów przed analizą samych danych.

- Proszę zoptymalizować wybór zakresu sygnałowego dla 10-krotnie większej świetlności.

- Proszę zoptymalizować wybór zakresu sygnałowego na podstawie Z_{obs} (w rzeczywistości procedura zabroniona bo obciąża wynik aktualną fluktuacja w danych).

- Narysuj Z_{exp} w funkcji świetlności (od 1, 5, 10, 50, 100, 500, 1000 1/fb)

- Dla jakiej wartości świetlności spodziewamy się odkrycia ($Z_{exp} = 5$)?

- Z optymalizuj zakres sygnałowy aby odkrycie ($Z_{exp} = 5$) pojawiło się dla najmniejszej świetlności.

- Jak zmienią się powyższe odpowiedzi jeśli przyjmiemy że oszacowane tło jest obarczone błędem systematycznym $\Delta\mu_b = 0.1\mu_b$? W tym przypadku proszę posłużyć się pseudo-eksperymentami w których wartość μ_b nie jest ustalona ale pochodzi z rozkładu normalnego o wartości oczekiwanej $\mu = \mu_b$ i odchylenie standardowemu $\sigma = \Delta\mu_b$.