Projekt 2A / Metody analizy danych doświadczalnych

Wszystkie pliki potrzebne do wykonania tego projektu znajdują się na serwerze `fatcat` w kartotece `~adamczyk/stat/projekt_2`

1 Plik DaneA1.txt zawiera pewne dane pomiarowe. Zakładamy, że dane te podlegają rozkładowi wykładniczemu:

$$f(x;\tau_1) = \frac{1}{\tau_1} e^{-\tau_1 x}$$

gdzie τ_1 jest nieznanym parametrem tego rozkładu. Na podstawie znanych wyrażeń na estymatory największej wiarygodności parametru τ_1 oraz jego wariancji oblicz estymator $\hat{\tau}_1$ oraz jego niepewność.

2 Znajdź estymator $\hat{\tau}_1$ oraz jego niepewność metodą graficzną rysując zależność logarytmu funkcji wiarygodności od parametry τ_1. Określ dolne ograniczenie na wariancje z twierdzenia Cramera-Rao. Narysuj drugą pochodną logarytmu funkcji wiarygodności oraz samą funkcję wiarygodności.

3 Oszacuj niepewność estymatora $\hat{\tau}_1$ metodą Monte Carlo.

4 Znajdź estymator $\hat{\tau}_1$ oraz jego niepewność używając pakiet MINUIT. Plik Fit4.cc zawiera przykład użycia programu MINUIT do znajdowania minimum funkcji $-2ln(L)$ dla rozkładu logistycznego. Zmodyfikuj ten przykład dla swoich potrzeb.

5 Znajdź bayesowski estymator parametru τ_1 (ze stałym zaczątkiem) interpretując funkcję wiarygodności z dokładnością do czynnika normalizującego jako rozład gęstość prawdopodobieństwa parametru τ_1. Oszacuj jego niepewność jako: a) dyspersję τ_1, b) podając przedział wiarygodności na poziomie wiarygodności 68%.

6,7,8,9 Powtórz analizę z punktów 2,3,4 i 5 ograniczając liczbę danych do np. 10.

10 Często nie dysponujemy poszczególnymi danymi, do dyspozycji mamy tylko histogram. Plik DaneA.root zawiera te same dane co plik DaneA1.txt ale w postaci histogramu (`h1`). Dla takich danych nadal możemy utworzyć funkcje wiarygodności w postaci:

$$\ln L(\tau_1) = \sum_{i=1}^{B} n_i \ln \mu_i(\tau_1),$$

gdzie n_i jest liczbą danych w przedziale i, $i = 1, ..., B$ a $\mu_i(\tau_1)$ jest spodziewaną liczbą przypadków w tym przedziale daną przez

$$\mu_i(\tau_1) = N \int_{x_{min,i}}^{x_{max,i}} f(x;\tau_1) dx$$

gdzie N jest całkowitą liczbą danych $N = \sum_{i=1}^{B} n_i$ a $x_{min,i}$ i $x_{max,i}$ jest dolną i górną granicą przedziału i. Zmodyfikuj program z punktu 4 tak aby operować danymi w postaci histogramu.

11 W praktyce obliczanie całek w wyrażeniu na spodziewaną liczbę przypadków w poprzednim punkcie może okazać się czasochłonne. Jeśli tylko funkcja $f(x;\tau_1)$ wewnątrz każdego binu jest w przybliżeniu liniowa to możemy użyć przybliżenia

$$\mu_i(\tau_1) \approx N f(x_i;\tau_1)\Delta x_i$$

gdzie $x_i = (x_{min,i} + x_{max,i})/2$ a $\Delta x_i = x_{max,i} - x_{min,i}$. Powtórz analizę z poprzedniego punktu korzystając z powyższego przybliżenia.

12 Aby sprawdzić, czy przybliżenie z poprzedniego punktu zbytnio nie zaburza wyników. Powtórz analizę z poprzedniego punktu używając zmodyfikowanego wyrażenia na $\mu_i(\tau_1)$

$$\hat{\mu}_i(\tau_1) = N f(\bar{x}_i;\tau_1)\Delta x_i$$

gdzie $\bar{x}_i$ jest średnią wartością zmiennej x w przedziale i dla $\tau_1 = \hat{\tau}_1$ gdzie $\hat{\tau}_1$ jest estymatorem parametru τ_1 uzyskanym w poprzednim punkcie. Powtórz tę iteracyjną procedurę aż wartości estymatora parametru τ_1 przestaną się zmieniać.

13 Aby sprawdzić czy wyniki nie są czułe na sposób dzielenia danych powtórz analizę z punktu 10 lub 11 dla histogramu (`h2`), który zawiera te same dane co histogram `h1` ale podzielone w inny sposób.

14 Zmodyfikuj program z punktu 10 tak, aby program MINUIT estymował nieznany parametr metodą najmniejszych kwadratów.

15 Zmodyfikuj program z punktu 14 tak, aby program MINUIT estymował nieznany parametr zmodyfikowaną metodą najmniejszych kwadratów.

16,17 Aby sprawdzić czy wyniki nie są czułe na sposób dzielenia danych powtórz analizę z punktu 14 i 15 dla histogramu (`h2`), który zawiera te same dane co histogram `h1` ale podzielone w inny sposób.

Zaliczenie tego projektu powinno zawierać makra w ROOT w osobnych plikach o nazwach `nazwa_n.C` gdzie n oznacza podpunkt projektu którego dotyczy dane makro, oraz pisemną analizę wyników ze szczególnym uwzględnieniem różnic w wynikach między poszczególnymi metodami oraz ich interpretacją. Prosze pamiętać, że analizujecie Państwo te same dane więc różnice w wynikach większe niż 20% oszacowanych błęd ów są już istotne.