Projekt 2D / Metody analizy danych doświadczalnych

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

1 Plik DaneD.txt zawiera pewne dane pomiarowe. Zakładamy, że dane te podlegają dwuwymiarowemu rozkładowi normalnemu $\mathcal{N}(x, y; \mu_x, \sigma_x, \mu_y, \sigma_y, \rho)$, Załóżmy, że wiemy że $\mu_x = \mu_y = \mu$, $\sigma_x = \sigma_y = \sigma$. Załóżmy dodatkowo (na chwilę) , że $\mu = 0$ a $\sigma = 1$ i jedynem nieznanym parametrem pozostaje współczynnik korelacji ρ. Na podstawie znanych wyrażeń na estymatory największej wiarygodności parametru ρ oraz jego wariancji oblicz estymator $\hat{\rho}$ oraz jego niepewność.

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

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

4 Znajdź estymator $\hat{\rho}$ 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 ρ (ze stałym zaczątkiem) interpretując funkcję wiarygodności z dokładnością do czynnika normalizującego jako rozład gęstość prawdopodobieństwa parametru ρ. Oszacuj jego niepewność jako: a) dyspersję ρ, 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 DaneB.root zawiera te same dane co plik DaneB1.txt ale w postaci histogramu (`h1`). Dla takich danych nadal możemy utworzyć funkcje wiarygodności w postaci:

$$\ln L(\rho) = \sum_{i,j=1}^{B} n_{i,j} \ln \nu_{i,j}(\rho),$$

gdzie $n_{i,j}$ jest liczbą danych w przedziale i, j, $i, j = 1, ..., B$ a $\nu_{i,j}(\rho)$ jest spodziewaną liczbą przypadków w tym przedziale daną przez

$$\nu_{i,j}(\rho) = N \int_{x_{min,i}}^{x_{max,i}} \int_{y_{min,j}}^{y_{max,j}} \mathcal{N}(x, y; \rho) dy dx$$

gdzie N jest całkowitą liczbą danych $N = \sum_{i,j=1}^{B} n_{i,j}$ a $x_{min,i}$, $x_{max,i}$ oraz $y_{min,j}$, $y_{max,j}$ są dolną i górną granicą przedziału i, j. 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 $\mathcal{N}(x, y; \mu$ wewnątrz każdego przedziału jest w przybliżeniu płaska, to możemy użyć przybliżenia

$$\nu_{i,j}(\rho) \approx N \mathcal{N}(x_i, y_i; \rho) \Delta y_j \Delta x_i$$

gdzie $x_i = (x_{min,i} + x_{max,i})/2$ a $\Delta x_i = x_{max,i} - x_{min,i}$ $y_j = (y_{min,j} + y_{max,j})/2$ a $\Delta y_j = y_{max,j} - y_{min,j}$. 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 $\nu_{i,j}(\rho)$

$$\hat{\nu}_{i,j}(\rho) = N\mathcal{N}(\bar{x}_{i,j}, \bar{y}_{i,j}; \rho)\Delta y_j \Delta x_i$$

gdzie $\bar{x}_{i,j}$ jest średnią wartością zmiennej x w przedziale i, j a $\bar{y}_{i,j}$ jest średnią wartością zmiennej y w przedziale i, j dla $\rho = \hat{\rho}$ gdzie $\hat{\rho}$ jest estymatorem parametru ρ uzyskanym w poprzednim punkcie. Powtórz tę iteracyjną procedurę aż wartości estymatora parametru ρ 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.