Statistische Methoden der Datenanalyse WS 2017/18

Prof. Dr. Ulrich Landgraf

Aufgabenblatt 5 vom 29.11.2010

Aufgabe 1 (3 Punkte)

Generieren Sie analog zu Aufgabe 3 des vierten Übungsblattes in ROOT Verteilungen der Summe aus 3,4,5 und 6 Zufallsvariablen.

a) Zeigen Sie durch die Simulation einer hinreichenden Anzahl von Zufallsereignissen, dass der Mittelwert einer Summe von n Zufallszahlen, die jeweils zwischen Null und 1 verteilt sind, durch die Formel

$$E(x_1 + \ldots + x_n) = \frac{n}{2}$$

und die Varianz durch die Formel

$$V(x_1 + \ldots + x_n) = \frac{n}{12}$$

beschrieben wird. Normieren Sie Ihr Histogramm nach dem Füllen, so dass das Integral Eins ergibt (Methode `TH1:Scale(Double_t f)`). Überlegen Sie sich anhand des Falles $n = 1$, wie Sie den Faktor f wählen müssen.

b) Damit sie die Grenzen des Histogramms nicht jedes Mal anpassen müssen, dividieren Sie jetzt die Summe der Zufallszahlen jeweils durch n. Wie ändert sich dadurch die Varianz? Stimmt Ihre Normierung noch? Zeichnen Sie jetzt dazu in dasselbe Diagramm eine entsprechende Gaußfunktion mit gleichem Mittelwert und Varianz mit `TF1::Draw("Same");`.

c) Mit der Methode `TH1::Add(TF1* func,Double_t f)` können Sie die Summe (oder mit $f = -1$. die Differenz) mit einer Funktion (`TF1::`) oder mit einem anderen Histogramm bilden. Nutzen Sie diese Methode, um die Verteilung des Durchschnitts von n Zufallszahlen mit der Gaußverteilung zu vergleichen. Beobachten Sie, wie sich die Differenz mit wachsendem n ändert.

Aufgabe 2 (4 Punkte)

In der Datei
`http://hep.uni-freiburg.de/tl_files/home/wwwherten/statistik/Gauss2D.h`
wird eine Klasse `Gauss2D` erklärt. Diese Klasse besitzt außer dem Konstruktor und Destruktor nur eine einzige Methode:

```
void GetXY(Double_t sigma_x, Double_t sigma_y, Double_t rho,
                                                Double_t &x, Double_t &y).
```

Damit erhält man zwei Zufallszahlen, x und y, die aus einer zweidimensionalen Gaußfunktion gezogen werden, die in x-Richtung die Breite `sigma_x` und in y-Richtung die Breite `sigma_y` besitzt. Der Mittelwert in beiden Richtungen liegt bei Null. Ist der dritte Eingangsparameter (*rho*) von Null verschieden, werden die Wahrscheinlichkeiten, die Variablen x und y zu erhalten, voneinander abhängig (korreliert).

Bitte wenden!

a) Benutzen Sie diese Klasse, um 10 000 Zufallszahlenpaare (x, y) in ein zweidimensionales Histogramm (Klasse `TH2F`) einzusortieren. Verwenden Sie dazu die Werte `sigma_x`=5. und `sigma_y`=3. Generieren Sie drei verschiedene Histogramme: eines mit `rho`=0, eines mit `rho`= -0.7 und ein drittes mit `rho`=+0.5.

b) Erstellen Sie mit geeigneten Methoden der Klasse `TH2` (von der die Klasse `TH2F` abgeleitet ist, die Projektionen der zweidimensionalen Histogramms auf die x-Achse bzw. y-Achse. Verifizieren Sie, dass die Projektionen gleich für alle drei Histogramme gleich aussehen, d.h. Gaußverteilungen mit dem gleichen Mittelwert und der gleichen Standardabweichung darstellen.

c) Berechnen Sie mit geeignetem Code einen Schätzwert für die nichtdiagonalen Elemente der Kovarianzmatrix, d.h. bilden Sie die Summe

$$\frac{1}{n} \sum_i (x_i - \mu_x)(y_i - \mu_y) \,,$$

wobei in diesem Falle die Mittelwerte $\mu_x = \mu_y = 0$ sind. Berechnen Sie mit dem Ergebnis für die drei Histogramme den Korrelationskoeffizienten und zeigen Sie, dass das Ergebnis mit den Werten des Parameters `rho` der Funktion `GetXY` übereinstimmt.

d) Das Histogramm wird durch die Koordinatenachsen in 4 Teile geteilt, die wie folgt mit den Buchstaben A, B, C, D bezeichnet werden:

C	D
A	B

In Region A und C bzw. B und D hat die Zufallsvariable x jeweils denselben Wertebereich, während in A und B (bzw. C und D) jeweils y gleich und x verschieden sind (nämlich in A ist $x < 0$ und in B ist $x > 0$, aber in beiden Regionen ist $y < 0$).

Demonstrieren Sie mit Ihren drei Histogrammen: Wenn x und y unabhängige Variablen sind (aber nur dann!) gilt für die Anzahl der Ereignisse in den jeweiligen Regionen:

$$\frac{B}{D} = \frac{A}{C} \,.$$

Diese Relation kann man verwenden, um eine Vorhersage des Ergebnisses in einer der vier Regionen zu machen, wenn man die beiden Parameter in den anderen drei Regionen gut bestimmen kann. Das wird häufig angewendet, wenn die Messung in einer Region durch einen unbekannten Störfaktor (Untergrund) schlecht möglich ist. Allerdings stellt sich dabei immer die Frage, wie sicher man sich ist, dass die beiden Variablen wirklich unabhängig sind, denn man kann ja in diesem Falle die Gültigkeit der Relation nicht überprüfen!