Albert-Ludwigs-Universität Freiburg Wintersemester 2009

Statistische Methoden der Datenanalyse

Markus Schumacher

Übung XI

Matthew Beckingham und Henrik Nilsen

21 .01.2010

Computerübung

Aufgabe 43 *Entdeckung eines neuen Teilchens*

Betrachtet wird folgendes Szenario: Eine Theorie sagt die Existenz eines neuen Teilchens mit einer Masse von $m_0 = 8\,\mathrm{GeV}$ vorher, welches im Experiment als eine resonante Überhöhung über einem exponentiell verteilten Untergrund ($\tau = 10\,\mathrm{GeV}$) beobachtet werden könnte. Die Wahrscheinlichkeitsdichtefunktion für den Untergrund sei also eine Exponentialverteilung, und die für das Signal eine Gaussfunktion mit Mittelwert 8 GeV und Standardabweichung 0,5 GeV, da wir weiterhin annehmen, dass die durch die Detektorauflösung beobachtete Breite der Resonanz – sofern sie existiert – 0,5 GeV betrage. Des weiteren sagt unsere bisherige Standardtheorie eine Gesamtanzahl von Untergrundereignissen von $N_{\mathrm{UG}} = 10000$ voraus, sowie unsere neue Theorie $N_{\mathrm{Sig}} = 175$ Signalereignisse. Im folgenden soll mittels der Neymann-Pearson-Lemma, die in der Vorlesung besprochen wurde, die Sensitivität des Experiments auf eine eventuelle Entdeckung untersucht werden. Die Werte Q ist definiert über das Verhältnis

$$Q = \frac{L(\vec{x}|H_1)}{L(\vec{x}|H_0)}, \tag{1}$$

wobei $\vec{x}$ die beobachteten Daten, L die unter der betreffenden Hypothese Likelihoodfunktion, H_0 die „nur-Untergrund“ Hypothese und H_1 die „Signal und Untergund“ Hypothese sind. Zumeist wird dann die Größe

$$q = -2 \ln Q \tag{2}$$

betrachtet.

Im folgenden soll die Monte-Carlo-Methode benutzt werden, um das Pseudoexperiment einerseits nur mit Untergrund, andererseits auch mit Signal- und Untergrund durchzuführen. Mittels dieses Pseudoexperimentes können dann die Verteilungen von $q(\vec{x}_{\mathrm{UG}}) \equiv q_0$ und von $q(\vec{x}_{\mathrm{Sig.+UG}}) \equiv q_1$ erzeugt werden, um festzustellen, wie sensitiv das Experiment auf das vorhergesagte neue Teilchen ist.

In dem Makro `SigPlusBg_i.C` finden Sie den Anfang des Skriptes, welches Sie benötigen, um die Verteilungen für q_0 und q_1 zu berechnen. Gehen Sie wie folgt vor, um es zu vervollständigen:

(i) Erstellen Sie zur Durchführung von 1000 Pseudoexperimenten eine Schleife.

(ii) Berechnen Sie für jedes Pseudoexperiment eine Anzahl an Signal- und Untergrundereignissen, verteilt nach einer Poisson-Statistik mit Erwartungswerten von $N_{\mathrm{Sig}} = 175$, beziehungsweise $N_{\mathrm{UG}} = 10000$. Verwenden Sie zum Beispiel die Funktion

```
gRandom->Poisson(nBg);
```

um eine poissonverteilte Variable zu erhalten.

(iii) Erstellen Sie eine Schleife über alle Signalereignisse, in welcher Sie einen Satz von Variablen der Signal WDF $\vec{x}_{\mathrm{Sig}}$ erstellen. Füllen Sie diese einerseits in ein Histogramm der Massenverteilung der Signalereignisse und gleichsam in eines für die Massenverteilung der Ereignisse von Signal + Untergund.

(iv) Erstellen Sie eine zweite Schleife über die Anzahl der Untergrundereignisse und erstellen Sie ebenfalls einen Satz von Variablen der Untergrund WDF $\vec{x}_{\mathrm{UG}}$. Diese füllen Sie nun analog in ein Histogramm der Massenverteilung der Untergrundereignisse und in das Histogramm für Signal + Untergrund aus dem letzten Teilabschnitt.

(v) Berechnen Sie den Wert von q jeweils für die reine Signal- und Untergrundverteilung. Hierzu müssen Sie eine Schleife über jedes Bin des Histogramms in dem Bereich $m_0 - 2\sigma < m < m_0 - 2\sigma$ erstellen und

$$q = \sum_{Bin_{min}}^{Bin_{max}} -s_i + n_i \ln\left(1 + \frac{s_i}{b_i}\right) \tag{3}$$

berechnen, wobei n_i für die Zahl an Ereignissen in Bin i des Histogramms des Pseudoexperimentes steht, und s_i und b_i die theoretischen Vorhersagen bezeichnen, wie viele Signal- und Untergrundereignisse in Bin i zu erwarten wären. Die Gesamtzahl an Ereignissen im Bereich $m_0 - 2\sigma < m < m_0 - 2\sigma$ der theoretischen Annahme ist gegeben durch s.

Sie können die Nummer des kleinsten Bins des Schleifenbereichs finden mit der Funktion

```
dataSigHist->FindBin(peakPos - 2.0* peakWidth)
```

Die Zahl an Einträgen in Bin i des Massenhistogramms erhalten Sie durch

```
dataBgHist->GetBinContent(i);
```

Die Zahl an theoretisch erwarteten Einträgen, beispielsweise an Signalereignissen, kann bestimmt werden durch

```
nSig * funkSig->Integral(dataSigHist->GetBinLowEdge(i),
dataSigHist->GetBinLowEdge(i+1))
```

(vi) Füllen Sie die Werte von q_0 und q_1 für jedes Pseudoexperiment in Histogramme. Den Code um diese Histogramme zeichnen zu lassen finden Sie am Ende des Makros.

(vii) Welchen Wert würden Sie als kritischen Wert von q setzen?

Aufgabe 44 *Vergleich von Messungen einer gaussverteilter Variablen*

Betrachten Sie den Fall, Sie hätten einen Satz von N Messungen einer gaussverteilten Variablen $\vec{x} = (x_1, x_2, \ldots \ldots, x_N)$ aufgenommen, wobei $\vec{x}$ gemäß $f_G(x; \mu_0, \sigma_0)$ verteilt sei. In dem vorliegenden Beispiel sollen Sie zwei verschiedene Hypothesentests betrachten, um sowohl Mittelwert als auch Varianz Ihrer Messungen mit der erwarteten Verteilung $f_G(x; \mu_0, \sigma_0)$ zu vergleichen.

Das Makro `HypoTest_i.C` beinhaltet Code, welcher einen Satz von M Experimenten generiert, jeweils mit N Messungen einer gaussverteilten Variablen. Jede Messung wird anschließend in ein Histogramm gefüllt, welches am Ende angezeigt wird.

(i) Vergleichen Sie zuerst den Mittelwert der generierten Messdaten mit der Gaussverteilung, welche Sie dazu verwendeten, die Messungen zu erstellen.

   a) Nehmen Sie an, Sie kennen Mittelwert μ wie auch σ der Gausskurve, welche sie zur Generierung der Daten benutzten. Um zu prüfen, ob Ihre Daten den Mittelwert $\mu = \mu_0$ besitzen, berechnen Sie für jedes Experiment die Teststatistik

$$t = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{N}} \tag{4}$$

   und füllen Sie dies in ein Histogramm. Diese Variable sollte nach der gaussischen WDF $f_G(t; 0,1)$ verteilt sein. Überzeugen Sie sich davon, indem Sie die Methode:

```
hist->Fit("gaus");
```

   verwenden, um eine Gaussverteilung in Ihr Histogramm von t zu fitten.

   b) Nehmen Sie nun an, Sie würden lediglich den Mittelwert μ, jedoch nicht die Breite der den Messungen zugrunde liegenden Gaussverteilung kennen. Folglich prüfen Sie, ob Ihre Daten den Mittelwert $\mu = \mu_0$ besitzen, indem Sie für jedes Experiment die Teststatistik

$$t' = \frac{\overline{x} - \mu_0}{s/\sqrt{N}} \tag{5}$$

berechnen, wobei die Standardabweichung s gegeben ist durch

$$s^2 = \frac{1}{N-1}\sum(x_i - \overline{x})^2 = \frac{1}{N-1}(\overline{x^2} - N\overline{x}^2) \tag{6}$$

Füllen Sie Ihre Werte von t' in ein Histogramm. Die Variable t' sollte entsprechend einer Studentschen t-Verteilung $f_t(t; N-1)$ mit $N-1$ Freiheitsgraden verteilt sein. Überzeugen Sie sich analog davon, dass dies der Fall ist, indem Sie eine Studentschen t-Verteilung in Ihr Histogramm fitten. Verwenden Sie

```
TF1* tFit = new TF1("tFit", "[1]*TMath::Student(x, [0])", -50., 50.);
```

um einen Fit in Form einer studentschen t-Verteilung zu definieren. Der [0]te Parameter steht für die Zahl an Freiheitsgraden. Der Methodenaufruf

```
hist->Fit("tFit");
```

fittet oben genannte Verteilung in das t'-Histogramm.

(ii) Als nächstes vergleichen Sie die Breite der generierten Daten mit der der Gaussverteilung, welche Sie verwendet haben, um die Messungen zu erstellen.

a) Gehen Sie davon aus, Sie würden einzig den Mittelwert μ, jedoch nicht die Breite der den Daten zugrundeliegenden Gaussverteilung kennen. Um nun zu prüfen, ob Ihre Daten die Breite $\sigma = \sigma_0$ aufweisen, berechnen Sie für jedes Experiment die Teststatistik

$$t'' = \frac{(n-1)s^2}{\sigma_0^2} \tag{7}$$

und füllen Sie dies in ein Histogramm. Die Variable t'' sollte nach einer χ^2 WDF $f_{\chi^2}(t; N-1)$ mit $N-1$ Freiheitsgraden verteilt sein. Überzeugen Sie sich davon, dass dies der Fall ist, indem Sie eine χ^2-Verteilung in Ihr Histogramm fitten. Verwenden Sie

```
TF1* tChi2Fit = new TF1("tChi2Fit", "[0]*(1.0/(TMath::Power(2,[1]/2.0)
      *TMath::Gamma([1]/2.0)))*TMath::Power(x,([1]/2.0)-1.0)
      *TMath::Exp(-x/2.0)", 0., 50.);
```

um die Fitfunktion einer χ^2-Verteilung zu definieren. Der [0]te Parameter steht für den Normierungsfaktor und der [1]te für die Anzahl an Freiheitsgraden. Mit

```
hist->Fit("tChi2Fit");
```

fitten Sie anschließend eine χ^2-Verteilung in Ihr Histogramm von t''.

(iii) Wie verändern sich die Verteilungen von t, t' und t'', wenn Sie einen systematischen Fehler hinzufügen, welcher alle Messungen um einen Wert von 1 erhöht, in der Art
$\vec{x} \rightarrow \vec{x'} = (x_1 + 1, x_2 + 1, \ldots \; \ldots, x_N + 1)$?

(iv) Wie verändern sich die Verteilungen von t, t' und t'', wenn Sie einen gaussischen Fehler mit Standardabweichung $\sigma = 0.1$ zu allen Messungen hinzufügen?