Albert-Ludwigs-Universität Freiburg Wintersemester 2011/2012

Statistische Methoden der Datenanalyse

Markus Schumacher

Übung X

Markus Warsinsky

16.1.2012

Anwesenheitsaufgaben

Aufgabe 59 *Vergleich von Messungen einer gaussverteilten Variablen*

Betrachten Sie den Fall, Sie hätten einen Satz von N Messungen einer gaussverteilten Variablen $\vec{x} = (x_1,x_2,\ldots,x_N)$ 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 den Mittlwert als auch Varianz Ihrer Messungen mit der erwarteten Verteilung $f_G(x;\mu_0,\sigma_0)$ zuvergleichen.

Das Makro `/home/warsinsk/sd_ws1112/ueb10/aufgabe_59_anfang.C` beinhaltet Code, welcher einen Satz von M Experimenten generiert, jeweils mit N Messungen einer gaussverteilten Variablen. Jede Messung wird 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 Gaussfunktion, die zum Generieren der Daten benutzt wurde. Um zu prüfen, ob Ihre Daten den Mittelwert $\mu = \mu_0$ besitzen, berechnen Sie für jedes Experiment die Teststatistik

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

und füllen Sie diese in ein bereitgestelltes 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 an Ihr Histogramm von t anzupassen. Stimmen die angepassten Parameter mit der Erwartung überein?

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

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

berechnen, wobei die Standardabweichung s gegeben ist durch

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

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 davon, dass dies der Fall ist, indem Sie eine Studentsche t-Verteilung an Ihr Histogramm anpassen. Verwenden Sie

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

um eine Anpassungsfunktion in Form einer Studentsches t-Verteilung bereitzustellen. Der `[0]`-te Prameter steht für die Anzahl der Freiheitsgrade. Passen Sie diese Funktion an Ihr Histogram von t' an (`thist.Fit("tFit");`).

(ii) Als nächstes vergleichen Sie die Breite der generierten Daten mit der der Gaussverteilung, die zum Generieren der Messungen benutzt wurde.

Gehen Sie davon aus, dass Ihnen der Mittelwert μ, jedoch nicht die Breite der Gussverteilung bekannt ist. Um nun zu prüfen, ob die Daten die Breite $\sigma = \sigma_0$ aufweisen, berechnen Sie für jedes Experiment die Teststatistik

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

und füllen Sie diese in ein weiteres 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 an Ihr Histogramm anpassen. Verwenden Sie

```
TF1 tChi2Fit = 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 eine χ^2-Verteilung zur Anpassung bereitzustellen. Der `[0]`-te Parameter steht für den Normierungsfaktor und der `[1]`-te für die Anzahl an Freiheitsgraden. Passen Sie diese Funktion mittels `tthist.Fit("tChi2Fit");` an Ihr Histogramm von t'' an.

(iii) Wie verändern sich die Verteilungen von t, t' und t'', wenn Sie einen systematischen Fehler hinzufügen, der alle Messungen um einen Wert von 1 erhöht, in der Art

$$\vec{x} \to \vec{x'} = (x_1 + 1, x_2 + 1, \ldots, x_N + 1)?$$

(iv) Wie verändern sich die Verteilungen von t, t' und t'', wenn Sie einen gaussischen Fehler mit Standardabweichung $\sigma = 0{,}1$ als zusätzliche Verschmierung zu allen Messungen hinzufügen?

Hausaufgaben

Aufgabe 60 *Studentsche t-Verteilung* **6 Punkte**

Betrachten Sie zwei Variablen: die erste, x, ist eine Standard-Normalverteilung $N(0,1)$ und die zweite, u, ist eine Chi-Quadrat verteilte Variable mit ν Freiheitsgraden, $\chi^2(\nu)$. x und ν seien unabhängig. Wenn die Variable t definiert ist als

$$t \equiv \frac{x}{\sqrt{u/v}} \qquad -\infty \leq t \leq \infty; \nu > 0 \tag{1}$$

dann ist diese gemäß der WDF

$$f(t;\nu) = \frac{\Gamma(\frac{1}{2}(\nu+1))}{\sqrt{\pi\nu}\,\Gamma(\frac{1}{2}\nu)} \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}} \tag{2}$$

verteilt, welche auch 'Studentsche t-Verteilung mit ν Freiheitsgraden' genannt wird (siehe Abb. 1). Die Studentsche t-Verteilung kann dazu benutzt werden, um auf einem Datensatz eine Nullhypothese H_0 zu testen.

Gegeben sei eine Stichprobe vom Umfang n aus einer Gaussverteilung $N(\mu,\sigma^2)$. Falls σ bekannt ist, ist die Verteilung für

$$t = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \tag{3}$$

eine Gaussverteilung $N(0,1)$. Wenn σ^2 jedoch nicht bekannt ist, dann ist t gegeben durch:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \tag{4}$$

mit der Stichprobenvarianz $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$. In diesem Fall ist t nach der Studentschen t-Verteilung mit $n-1$ Freiheitsgraden verteilt.

Betrachten Sie als Beispiel die Messung eines monoenergetischen Strahls von Teilchen mit Impuls $P_0 = 24.90\ \mathrm{GeV}/c$. Dieser trifft auf eine Blasenkammer und durch Messung der Krümmung entlang der Teilchen spur wird der inverse Impuls $1/P_i$ bestimmt. Nehmen Sie an, dass $1/P$ für 20 Teilchen durch zwei verschiedene Detektoren A and B mit den Ergebnissen $1/P_A = (40.12 \pm 0.46) \times 10^{-3}(\mathrm{GeV}/c)^{-1}$ und $1/P_B = (40.25 \pm 0.25) \times 10^{-3}(\mathrm{GeV}/c)^{-1}$ gemessen wurde.

Um zu testen, ob beide Messungen mit der Bestimmung des inversen Impulses der einfallenden Teilchen, $1/P_0$, konsistent sind, sollten Sie diese beiden Hypothesen betrachten:

$$H_0 : \frac{1}{P_i} = \frac{1}{P_0}$$
$$H_1 : \frac{1}{P_i} \neq \frac{1}{P_0}$$

(i) Was sind, unter Hinzunahme von Gleichung 4, die Werte von t für beide Messungen?

(ii) Wie viele Freiheitsgrade hat jede Messung?

(iii) Nutzen Sie die zur Verfügung gestellte Tabelle, um die Grenze der kritischen Region mit einer Signifikanz von $\alpha = 0.05$ zu finden. Bedenken Sie hierbei, dass Sie einen beidseitigen Test durchführen. Wieso muss dieser Test auf zwei Seiten durchgeführt werden?

(iv) In Bezug auf den inversen Impuls der einfallenden Teilchen: Sind beide Messungen damit konsistent?

Aufgabe 61 *F-Test für zwei verschiedene Messungen* **6 Punkte**

Betrachten Sie zwei unabhängige Chi-Quadrat verteilte Variablen, u_1 und u_2, mit ν_1 und ν_2 Freiheitsgraden, d.h. u_1 ist gemaß $\chi^2(\nu_1)$ verteilt und u_2 nach gemaß $\chi^2(\nu_2)$. Dann ist die Variable F, definiert durch

$$F \equiv \frac{u_1/\nu_1}{u_2/\nu_2} \qquad 0 \leq F \leq \infty;\ \nu_1,\nu_2 > 0 \tag{5}$$

verteilt nach der WDF

$$f(F;\nu_1,\nu_2) = \frac{\Gamma\left(\frac{1}{2}(\nu_1+\nu_2)\right)}{\Gamma\left(\frac{1}{2}\nu_1\right)\Gamma\left(\frac{1}{2}\nu_2\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{1}{2}\nu_1} \frac{F^{\frac{1}{2}\nu_1-1}}{\left(1+\frac{\nu_1 F}{\nu_2}\right)^{\frac{1}{2}(\nu_1+\nu_2)}} \tag{6}$$

welche 'F-Verteilung für (ν_1,ν_2) Freiheitsgrade' genannt wird (siehe Abb. 2).

Für zwei Datensätze $x_1,x_2,\ldots\ldots,x_n$, gaussverteilt nach $N(\mu_1,\sigma_1^2)$, und $y_1,y_2,\ldots\ldots,y_n$, gaussverteilt nach $N(\mu_2,\sigma_2^2)$, wobei die Mittelwerte μ_1 und μ_2 beider Verteilungen bekannt sind, ist die Größe

$$F = \frac{s_1}{s_2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_1)^2}{\frac{1}{m-1}\sum_{i=1}^{m}(y_i-\mu_2)^2} \tag{7}$$

durch die F-Verteilung mit (n,m) Freiheitsgraden verteilt. Daher kann das Verhältnis s_1/s_2 benutzt werden, um die Hypothese, dass beide Verteilungen die selbe Varianz ($H_0 : \sigma_1^2 = \sigma_2^2$) aufweisen, gegen die Hypothese, dass beide Varianzen verschieden sind ($H_1 : \sigma_1^2 > \sigma_2^2$), zu testen.

Kehren wir noch einmal zur Messung des Teilchenimpulses aus Aufgabe 60 zurück. Beide Messungen haben den selben Mittelwert $\mu_1 = \mu_2 = \mu_0$. Betrachten Sie hier nun folgende beiden Hypothesen für die Varianzen des inversen Impulses:

$$H_0 : \frac{1}{\sigma_1^2} = \frac{1}{\sigma_2^2}$$
$$H_1 : \frac{1}{\sigma_1^2} < \frac{1}{\sigma_2^2}$$

(i) Berechnen Sie den Wert von F für beide Messungen.

(ii) Wie viele Freiheitsgrade haben die Messungen?

(iii) Was ist der kritische Wert von F bei einer Signifikanz von 5%? Sollte der Test auf einer oder auf zwei Seiten durchgeführt werden?

(iv) Sind daher die Präzisionen der beiden Messungen miteinander konsistent?

Aufgabe 62 *Zählexperiment für eine Signal- und Untergrundmessung* **8 Punkte**

Betrachtet wird ein Experiment, dessen Ziel die Entdeckung eines neuen Teilchens oberhalb eines von momentanen Theorien vorhergesagten Untergrundes ist. Dabei könnte es sich beispielsweise um das Higgs-Boson oder auch um supersymmetrische Teilchen handeln, wobei die Untergrundvorhersage durch das Standardmodell der Teilchenphysik erfolgt.

Die zu betrachtenden Hypothesen sind also die Nullhypothese H_0, dass nur Ereignisse aus Untergrundprozessen gemessen wurden, sowie die Alternativhypothese H_1, dass sowohl Signal- als auch Untergrundereignisse beobachtet wurden.

Im Experiment wurde eine Gesamtanzahl x von Ereignissen aufgezeichnet. Weiterhin wurde eine andere, signalfreie kinematische Region definiert, aus der man die Normierung des Untergrundes bestimmen kann. In dieser Region wurden y Ereignisse gefunden. Das Verhältnis der Untergrundereignisse in der signalfreien Region zu denjenigen in der Signalregion sei τ. Die mittlere Anzahl der Untergrundereignisse in der Signalregion sei b und die mittlere Anzahl der Signalereignisse im Falle von Hypothese H_1 sei s.

(i) Stellen Sie die Likelihoodfunktionen für die Hypothesen H_0 und H_1 auf. Nehmen Sie dabei an, dass die Gesamtanzahlen von Ereignissen in Signal- und Kontrollregion jeweils Poissonverteilt sind.

(ii) Betrachten Sie jetzt die Schätzer für s und b unter der Hypothese H_1 ($\hat{s}$ bzw. $\hat{b}$) sowie den Schätzer für b unter der Nullhypothese, $\hat{\hat{b}}$.

a) Stellen Sie die Profile Likelihood λ auf.

b) Bestimmen Sie Ausdrücke für $\hat{s}$, $\hat{b}$ und $\hat{\hat{b}}$. Betrachten Sie dazu

$$\left.\frac{\partial L(H_0)}{\partial b}\right|_{\hat{\hat{b}}}, \tag{8}$$

sowie die die beiden gleichzeitigen Einschränkungen

$$\left.\frac{\partial L(H_1)}{\partial s}\right|_{\hat{s}} \quad \text{und} \quad \left.\frac{\partial L(H_1)}{\partial b}\right|_{\hat{b}}. \tag{9}$$

c) Berechnen Sie $q = -2\ln\lambda$.

(iii) Monte-Carlo-Studien von Proton-Proton-Kollisionen im ATLAS-Detektor haben gezeigt, dass der Wirkungsquerschnitt für $pp \to H + X \to \gamma\gamma + X$-Ereignisse, die die Ereignisselektion passieren, gegeben ist durch $\sigma_S = 25{,}4\,\text{fb}$. Der Wirkungsquerschnitt für Untergrundereignisse, die dieselbe Ereignisselektion passieren, beträgt $\sigma_B = 947\,\text{fb}$. In einer weiteren Analyse kann ein reiner Untergrunddatensatz mit einem Wirkungsquerschnitt von $\sigma_T = 10300\,\text{fb}$ ausgewählt werden.

a) Benutzen Sie die Relation

$$N = \mathcal{L}\sigma, \tag{10}$$

um die Anzahlen von Signal- (x), Untergrundereignissen (y) sowie die Anzahl von Ereignissen in der Seitenbandregion (τb) für eine integrierte Luminosität von $\mathcal{L} = 10\,\text{fb}^{-1}$ auszurechnen.

b) Berechnen Sie die Schätzer für die Anzahl der Signalereignisse $\hat{s}$, der Untergrundereignisse $\hat{b}$ unter der Hypothese von Signal plus Untergrund , sowie den Schätzer $\hat{\hat{b}}$ auf die Anzahl der Untergrundereignisse in der Nur-Untergrund Hypothese.

c) Berechnen Sie die Größe

$$q = -2\ln\lambda. \tag{11}$$

d) Berechnen Sie daraus die Signifikanz des vorhergesagten Signals.

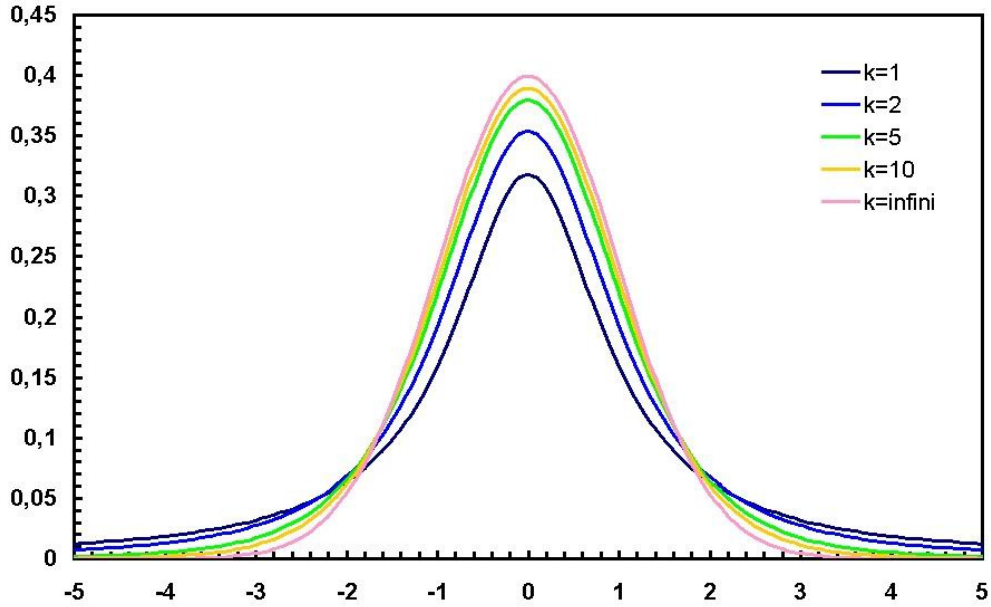

Abbildung 1: Die Studentsche t-Verteilung

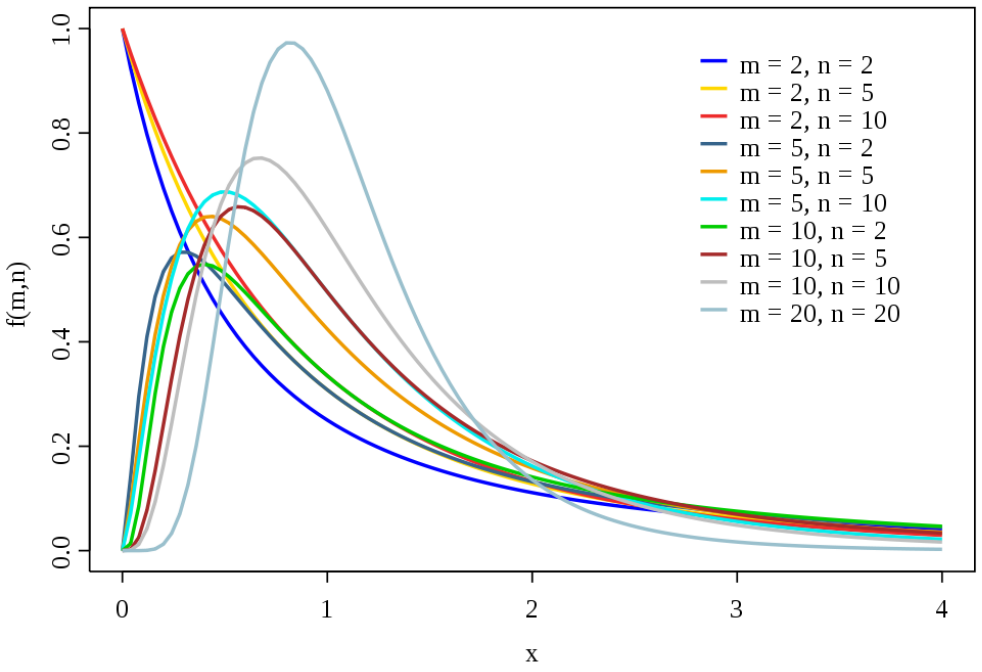

Abbildung 2: Die F-Verteilung