Albert-Ludwigs-Universität Freiburg Sommersemester 2009

Statistische Methoden der Datenanalyse

Markus Schumacher

Übung VII

Matthew Beckingham und Markus Warsinsky

17.6.2009

Anwesenheitsaufgaben

Aufgabe 33 *Studentsche t-Verteilung*

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

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

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

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

mit der Stichprobevarianz $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. 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 34 *Teilchenidentifikation 1*

Durchquert ein geladenes Teilchen ein Gasvolumen, erzeugt dieses in dem Medium Ionisation. Die mittlere Menge hängt dabei von der Masse und Impuls des Teilchens ab. Daher können durch Ausarbeitung einer Testhypothese, basierend auf der erfassten Ionisation im Gasvolumen bei bekannten Impuls verschiedene, Teilchen identifiziert werden.
Betrachten Sie einen Strahl von Teilchen, welcher entweder Pionen oder Elektronen enthält. So kann man, als eine Funktion der Ionisation t, die WDF der Hypothese $g(t|e)$, dass das Teilchen ein Elektron ist, und der Hypothese $g(t|\pi)$, dass das Teilchen ein Pion ist, aufstellen. Hierzu wählt man eine Menge von Elektronen aus, indem gefordert wird, dass $t \leq t_{cut}$:

(i) Bestimmen Sie Ausdrücke für die Effizienzen Elektronen oder Pionen zu selektieren. Wie stehen diese Werte in Beziehung zu der Signifikanz der Elektronenhypothese und zur Mächtigkeit diese von der Pionhypothese zu unterscheiden?

(ii) Wenn der relative Anteil von Elektronen, a_e, und Pionen, $1 - a_e$, im Strahl den ausgewählten Teilchen bekannt ist, dann ist die Teststatistik t verteilt gemäß :

$$f(t; a_e) = a_e g(t|e) + (1 - a_e) g(t|\pi) \tag{5}$$

Bestimmen Sie dann einen Ausdruck für die Gesamtzahl an Teilchen, welche nach dem Schnitt auf t übrig bleiben, N_{acc}, als eine Funktion der Gesamtzahl an Teilchen, der Zahl an Elektronen und der Effizienzen, Elektronen und Pionen zu selektieren.

(iii) Stellen sie danach die Gleichung so um, dass Sie die Zahl an Elektronen in der Auswahl angeben können. Was sagt Ihnen dies über die Elektron- und Pioneffizienzen, wenn Sie die Zahl von Elektronen berechnen wollen?

Hausaufgaben

Aufgabe 35 *Kombination von Messungen mit der Methode der kleinsten Quadrate* **8 Punkte**

Es ist möglich, einen Spezialfall der Methode der kleinsten Quadrate zu benutzen, um Messungen mit der selben Qualität zu kombinieren. Betrachten Sie N Messungen, y_i, welche den wahren, aber unbekannten Wert λ bestimmen sollen. Jede Messung y_i hat einen geschätzten Fehler von σ_i. Da λ für alle Ereignisse konstant ist, ergibt sich die χ^2-Variable zu:

$$\chi^2(\lambda) = \sum_{i=1}^{N} \frac{(y_i - \lambda)^2}{\sigma_i^2} \tag{6}$$

(i) Wenn jedoch die Messungen von y_i nicht unabhängig sind, sondern eine Korrelation, gegeben durch die Kovarianzmatrix V, besitzen, ergibt sich:

$$\chi^2(\lambda) = \sum_{i,j=1}^{N} (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda) \tag{7}$$

Zeigen Sie, dass in diesem Fall der Schätzer der Methode der kleinsten Quadrate für λ gegeben ist durch

$$\hat{\lambda} = \sum_{i=1}^{N} w_i y_i \tag{8}$$

wobei die Gewichtungen w_i gegeben sind durch

$$w_i = \frac{\sum_{j=1}^{N}(V^{-1})_{ij}}{\sum_{k,l=1}^{N}(V^{-1})_{kl}}, \tag{9}$$

und dass die Varianz gegeben ist durch

$$V[\hat{\lambda}] = \sum_{i,j=1}^{N} w_i V_{ij} w_j. \tag{10}$$

(ii) Betrachten Sie jetzt zwei Messungen, y_1 und y_2, mit einer zugehörigen Kovarianzmatrix

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{11}$$

mit dem Korrelationskoeffizienten $\rho = V_{12}/(\sigma_1\sigma_2)$. Zeigen Sie durch Berechnung des Inversen der Kovarianzmatrix V^{-1}, dass der Schätzer für λ gegeben ist durch

$$\hat{\lambda} = wy_1 + (1 - w)y_2 \tag{12}$$

mit

$$w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \tag{13}$$

(iii) Zeigen Sie, dass die Varianz von $\hat{\lambda}$ gegeb ist durch

$$\frac{1}{V[\hat{\lambda}]} = \frac{1}{1-\rho^2}\left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2}\right] = \frac{1}{\sigma^2} \tag{14}$$

und zeigen Sie dass folglich

$$\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} \geq 0 \tag{15}$$

was bedeutet, dass die Kombination von zwei Messungen immer zu einer Verbesserung der Varianz des Schätzers führt.

Aufgabe 36 *Teilchenidentifikation 2* **6 Punkte**

Nehmen Sie jetzt an, dass beide Hypothesen der vorigen Aufgabe durch eine um $t = 0$ zentrierte Gaussverteilung für die Elektronen und eine um $t = 2$ zentrierte Gaussverteilung für die Background-Pionen konstruiert sind. Beide Gaussschen Hypothesen haben eine Standardabweichung von eins. Ein Test zur Elektronenselektion wird konstruiert, indem eine Ionisation $t < 1$ gefordert wird.

(i) Was ist der Signifikanzlevel des Tests (d.h. die Wahrscheinlichkeit Elektronen zu akzeptieren)?

(ii) Wie groß ist die Mächtigkeit des Tests gegen die Hypothese, dass das Teilchen ein Pion ist? Wie groß ist die Wahrscheinlichkeit, dass ein Pion als Elektron akzeptiert wird?

(iii) Betrachten Sie eine Verhältnis von 99% Pionen and 1% Elektronen im Strahl. Wie groß ist die Reinheit der durch $t < 1$ selektierten Auswahl?

Aufgabe 37 *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

$$F \equiv \frac{u_1/\nu_1}{u_2/\nu_2} \qquad 0 \leq F \leq \infty;\ \nu_1,\nu_2 > 0 \tag{16}$$

verteilt nach der WDF

$$f(F;\nu_1,\nu_2) = \frac{\Gamma\left(\frac{1}{2}(\nu_1+\nu_2)\right)}{\Gamma\left(\frac{1}{2}\nu_1\right)\Gamma\left(\frac{1}{2}\nu_2\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{1}{2}\nu_1} \frac{F^{\frac{1}{2}\nu_1-1}}{\left(1+\frac{\nu_1 F}{\nu_2}\right)^{\frac{1}{2}(\nu_1+\nu_2)}} \tag{17}$$

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

$$F = \frac{s_1}{s_2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_1)^2}{\frac{1}{m-1}\sum_{i=1}^{m}(y_i-\mu_2)^2} \tag{18}$$

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 33 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?

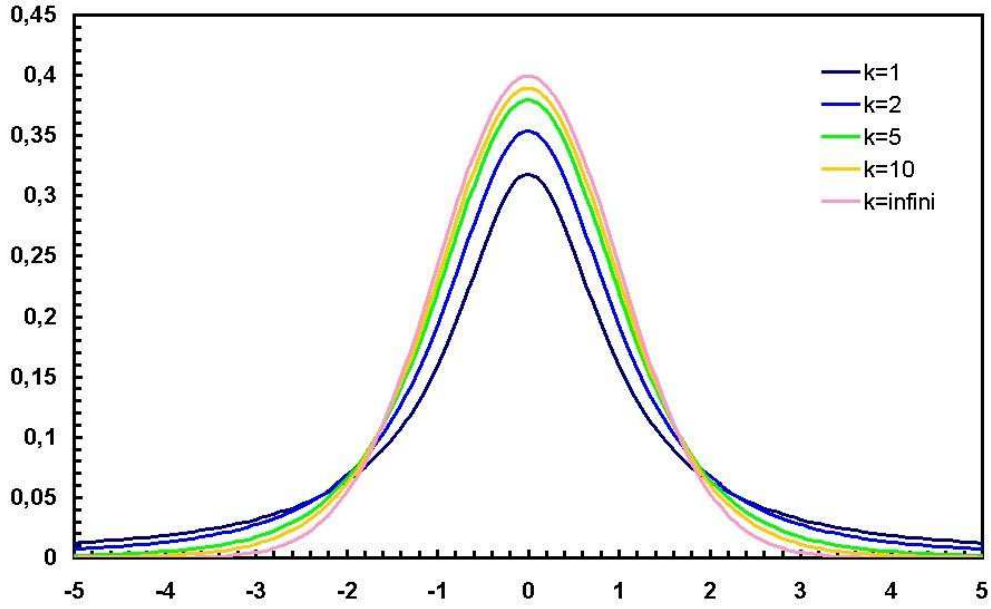

Abbildung 1: Die Studentsche t-Verteilung

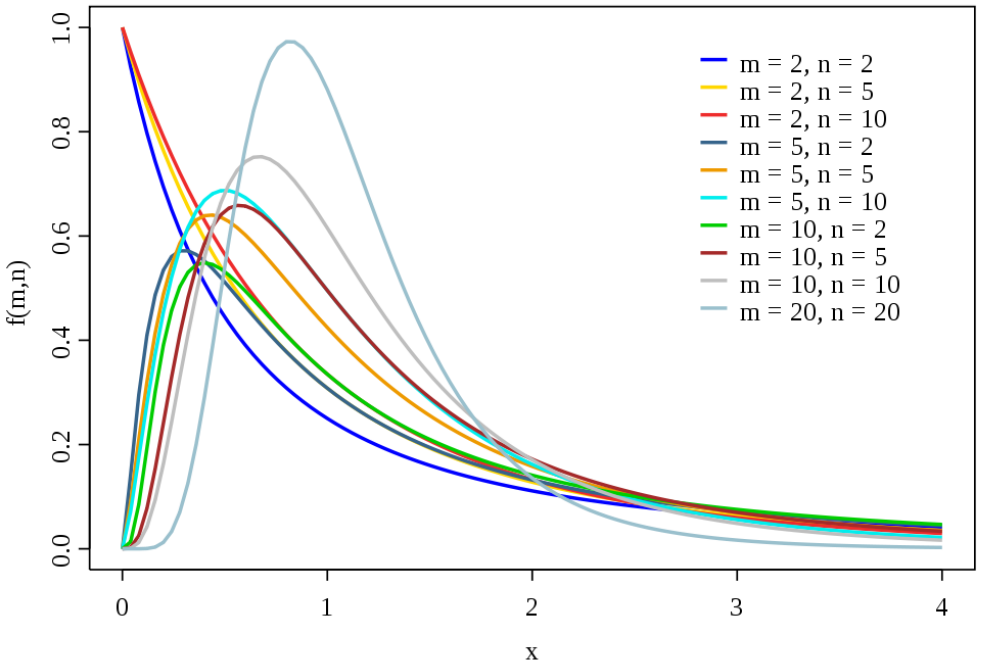

Abbildung 2: Die F-Verteilung