FH FFM / FB 2
Prof. Dr. D. Hackenbracht

Computeralgebra 2
Übung Fit_1

Literatur zum Thema Datenanalyse, genauer: zur (mathematischen) Statistik:

Das Thema ist sehr umfangreich. Allein im Papula wird es auf ca. 250 Seiten behandelt (ohne die vorgeschaltete Wahrscheinlichkeitsrechnung !!).

Papula, Band 3
Bleymüller et al., Statistik für Wirtschaftswissenschaftler, Verlag Vahlen
Henze, Stochastik für Einsteiger, Vieweg
Fischer, Stochastik einmal anders, Vieweg

Teil I: Einfache statistische Kennzahlen für eine Mess-Größe

Teil II: Lineare Regression: Zusammenhang zwischen zwei Größen

Nicht Bestandteil dieses Kurses sind:

Teil III: Interpolation durch Polynome (Splines)

Teil IV: Interpolation durch trigonometrische Polynome (Fast Fourier Transform)

Hinweis:

1)
Evaluieren Sie das notebook ...**testdaten**.nb, bevor Sie Ihre Lösung beginnen!
Sie können dann den jeweils benötigten Datensatz in Ihrer Lösung verwenden
(die Variable ist dann bekannt!).
Sie können den Datensatz auch aus ...**testdaten**.nb kopieren und ihn einer Variablen zuweisen.

2)
Erzeugen Sie stets geeignete graphische Darstellungen.

Teil I: Einfache statistische Kennzahlen für eine Mess-Größe

Wir simulieren die mehrfache Messung (Mess-Reihe) der Erdbeschleunigung g .
n ist die Anzahl der Messwerte, g_i ist der i - te Messwert.

Vorbemerkungen:
Der **Mittelwert** $\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g_i$ ist der beste Schätzwert für den unbekannten „wahren" Wert g (Papula 661). Er minimiert die Größe $\tilde{S} = \sum_{i=1}^{n}(g_i - a)^2$, d.h. bei a = $\bar{g}$ ist deren Minimum.

Die einzelnen Messwerte streuen um den Mittelwert. Dieser reicht daher als Angabe nicht aus. Man ist aber nicht an der Abweichung des einzelnen Messwertes vom Mittelwert interessiert, sondern möchte die Streuung des ganzen Datensatzes charakterisieren. Ein Maß für diese Streuung ist die (empirische) **Varianz** s^2 (auch **Stichprobenvarianz** genannt):
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(g_i - \bar{g})^2$; beachten Sie den Nenner (n - **1**) !

Die Wurzel daraus ist die (empirische) **Standardabweichung** s . Auch sie ist ein Maß für die Streuung (s.u.: A1). Der Zusatz „empirisch" wird oft weggelassen (z.B. in Mathematica und im Papula). Früher wurde s auch als „mittlerer Fehler der Einzelmessung" bezeichnet (Pap 662).

Das Ergebnis einer Messreihe wird mit der **Standardabweichung des Mittelwertes** $s/\sqrt{n}$ angegeben, hier also: g = $\bar{g} \pm s_{\bar{g}}$, $s_{\bar{g}} = s/\sqrt{n}$ (s.u.: A2). Ohne Angabe dieser Standard-abweichung ist ein solches Ergebnis wertlos!
Die Standardabweichung des Mittelwertes wurde früher auch als „mittlerer Fehler des Mittelwertes" bezeichnet (Pap 662). Sie beschreibt auch die Streuung der aus verschiedenen Messreihen erhaltenen Mittelwerte um den wahren Wert (Papula 661; s.u.: A3).

Die Standardabweichung s ändert sich nur wenig, wenn man die Zahl der Messwerte (also n) erhöht, denn die Ungenauigkeit einer Einzelmessung $g_i - \bar{g}$ schwankt zwar von Messung zu Messung, aber sie bleibt doch in der gleichen Größenordnung, wenn jede Messung mit der gleichen Präzision durchgeführt wird. Die Standardabweichung des Mittelwertes wird dagegen kleiner, wenn man die Zahl der Messungen (also n) erhöht.

Wir gehen davon aus, dass die Messwerte normalverteilt sind (also dass ihre Verteilung der Gauß-Verteilung folgt). Dann gilt:

(A1) Es liegen ca. 68,3 % aller Einzelmesswerte im Intervall $\bar{g} \pm s$.

(A2) Mit einer Wahrscheinlichkeit von ca. 68,3% liegt der wahre Wert g im Intervall $\bar{g} \pm t\, s/\sqrt{n}$.). t ist eine Zahl, deren Wert von der Anzahl der Messwerte n abhängt (Pap 666). Für n > 30 ist t ≈ 1.
Die Wahrscheinlichkeit wird jetzt auch als Vertrauensniveau oder statistische Sicherheit bezeichnet (Pap 666). Man spricht auch von einem **Konfidenz-** oder **Vertrauensintervall** zum vorgegebenen Vertrauensniveau (Papula 666, Pitka-Skript). Der Wert von t hängt außer von n auch noch von dem gewählten Vertrauensniveau γ ab (Pap 666). Man findet die Werte von t in entsprechenden Tabellen (Stichworte: t-Verteilung, Student-Vert.). Für γ = 68,3% und n > 30 ist t ≈ 1 (Pap 667).

(A3) Wenn man eine Messreihe (mit n Messungen) ν mal wiederholt, erhält man die Mittelwerte $\bar{g}_1$,…, $\bar{g}_\nu$. Es liegen ca. 68,3 % dieser Mittelwerte im Intervall $\bar{g} \pm s_{\bar{g}}$.

1. [cas2_fit0]

Lesen Sie zunächst das notebook 0_MessreiheAuswertung !

Die „Mess“ergebnisse (Messung der Erdbeschleunigung) stehen in der Liste **data0** im notebook ...testdaten.

a)
Berechnen Sie den **Mittelwert** $\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g_i$ mit dem Befehl *Sum* und mit dem Befehl *Mean*.

b)
Berechnen Sie s^2 sowohl über den Befehl *Sum* als auch über den Mathematica-Befehl *Variance*. Berechnen Sie s sowohl als Wurzel aus s^2 als auch über *StandardDeviation* . Berechnen Sie die Standardabweichung des Mittelwertes.

c)
Überprüfen Sie die Aussage (A1) für unsere Messreihe, indem Sie mit dem Befehl *Count* die Anzahl derjenigen Werte berechnen, welche im Intervall $\bar{g} \pm s$ liegen.

Zeigen Sie, dass für eine (exakt) normalverteilte Größe g die Aussage (A1) gilt. Integrieren Sie dazu die zugehörige Wahrscheinlichkeitsdichte (-Funktion) in dem Intervall $\bar{g} \pm s$.

Teil II: Lineare Regression: Zusammenhang zwischen zwei Größen

Lesen Sie zunächst das notebook 1_fit

2. [cas2_fit1]
Es soll eine **Ausgleichskurve für eine „Messreihe“** bestimmt werden.
Die „Mess“ergebnisse stehen in der Liste **data1** .

a)
Man vermutet, dass die Abhängigkeit einer physikalischen Größe Y von einer zweiten Größe X durch eine Funktion $Y = y(x) = a\,x^2 + b\,x + c$ beschrieben wird.
Diese Funktion ist unsere Modellfunktion. Die Parameter (hier: a, b, c) müssen noch bestimmt werden: an die Messdaten muss die theoretische Kurve, der Graph der Funktion y(x), **optimal angepasst** („gefittet“) werden. Aus dieser Prozedur, die in Mathematica in der einfachsten Ausführung als Befehl *Fit* vorliegt, ergibt sich dann die beste Schätzung für die Parameter.

b)
Zur Sicherheit wird auch noch ein anderer Zusammenhang geprüft: $y = y(x) = a\,x^3 + c$.
Welches Modell soll man bevorzugen? Führen Sie eine Bewertung durch

i) über die **Summe der Schwankungsquadrate** = Summe der quadrierten Residuen, **SQR** (Tipp: notebook / Ergänzung, Bemerkung 2),
ii) über die **Varianz der Residuen** und **die Standardabweichung der Residuen** (Tipp: notebook / Ergänzung, Bemerkung 2),
iii) über das **Bestimmtheitsmaß** (Tipp: notebook / Ergänzung, Bemerkung 3).

c)
Führen Sie die Bestimmung der Ausgleichskurven auch mit dem Befehl *FindFit* durch.

Sie können sich zunächst auf das Thema „Tayloremtwicklung“ einstimmen, indem Sie bei ComputerAlgebraSysteme 1 (CAS1/ Einführung) das entsprechende Kapitel studieren.

Vorbemerkung:

Wenn man den tatsächlichen Zusammenhang zwischen den Größen Y und X nicht kennt, versucht man es oft mit einem Ansatz in Form eines Polynoms. **Benutzt man als Polynom ein Taylor-Polynom** (Entwicklungspunkt x_0), gibt man eine bestimmte Struktur vor, wobei sich der Entwicklungspunkt x_0 aus der Aufgabenstellung ergibt. Die **Koeffizienten** des Taylor-Polynoms, die sich bei einem bekannten funktionalen Zusammenhang als Ableitungen der Funktion ergeben, **sind jetzt die gesuchten Parameter**.

Diesen Ansatz kann man auch für eine mehrdimensionale Abhängigkeit wählen, also einer Abhängigkeit zwischen einer Größe Y und den (Einstell-) Größen X_1, X_2, … .

Lesen Sie das notebook 2_taylor .

3. [cas2_fit2]

a)
Fitten Sie an die „Mess“-Daten **data2a** ein Taylorpolynom 4. Grades um $x_0 = 2$.
Vergleichen Sie das Ergebnis mit dem Fit an das allgemeine Polynom 4. Grades.

b)
In einer Diplomarbeit für die Firma Braun (Wittekind, 2006) sollte der Zusammenhang zwischen der Zug-Reißfestigkeit von Chip-Bonds mit den Einflussgrößen Zugkraft F , Temperatur T und Belastungsdauer B untersucht werden: Z = f (B, T, F) .

Die Belastungsdauer soll jetzt bei allen Versuchen gleich sein. Die zugehörigen „Mess“daten stehen in **data2b** .

Mit B = const muss man jetzt die Funktion Z = g (T, F) untersuchen.
Entwickeln Sie g um $(T_0, F_0) = (0, 0)$ in ein **Taylorpolynom,** wobei die **Variablen T und F beide** bis zur **2. Ordnung gehen**. Reduzieren Sie diese Entwicklung auf ein Polynom 2. Grades. Dieses Polynom soll dann an die Messdaten gefittet werden. Lesen Sie aus dem Polynom die Basisfunktionen für den Fit ab. Der *Fit* – Befehl muss jetzt für ZWEI Variablen ausgeführt werden. Er hat die Struktur Fit [daten, basisfunktionen, { T , F }] . Lassen Sie sich die Fit-Funktion auch plotten: *Plot3D* , ebenso die Messdaten über den Befehl *ListPointPlot3D* .