Physikalisch-chemisches Grundpraktikum Gruppe 16
Philipp von den Hoff
Andreas J. Wagner

Versuch 30:
Statistik
durchgeführt am 24.06.2004

Zielsetzung:
Mit Hilfe von in BASIC geschriebenen Programmen soll ein Verfahren zur linearen Regression entwickelt werden sowie über einen Zufallsgenerator der Unterschied zwischen einer Bernoulli- und einer Gaussverteilung demonstriert werden; desweiteren sind die bei einer Gaussverteilung erhaltenen Mittelwerte und Standardabweichungen in Abhängigkeit von der Anzahl der Stichproben zu diskutieren.

Theoretischer Hintergrund:
a) Lineare Regression:
Werden bei einer Messung verschiedene Werte erhalten, die von einer idealen Geraden abweichen, so ist es praktisch, diese mit Hilfe einer „Ausgleichsgeraden“ zu mitteln, was über das Verfahren der sogenannten *Linearen Regression* geschieht.
Wie jede Gerade besitzt auch eine solche Ausgleichsgerade die mathematische Form

$$\mathbf{y = a + bx}$$

mit der Steigung b und dem Ordinatenabschnitt a.
Jeder Punkt mit den Koordinaten x_2/y_2 , der nicht auf der Ausgleichsgeraden liegt, hat einen „Fehler“ (y-Differenz) von

$$\mathbf{\Delta y = a + bx - y_2}$$

Damit die Ausgleichsgerade optimal wird, muss die Summe der Fehlerquadrate über alle Messpunkte möglichst klein werden, d.h. die partiellen Ableitungen des Fehlerquadrats S nach a und b müssen minimal werden, S entspricht dem Ausdruck

$$S=\sum_{i}^{N}(a\ +b\ \cdot x_i-y_i)^2$$

Es ergibt sich:

$$a\ =\frac{(\sum_{i}^{N} y_i-b\ \cdot\sum_{i}^{N} x_i)}{N}$$

und

$$b\ =\frac{Q_{xy}}{Q_x}$$

mit

$$Q_x = \sum_i^N x_i^2 - \frac{(\sum_i^N x_i)^2}{N} \qquad Q_y = \sum_i^N y_i^2 - \frac{(\sum_i^N y_i)^2}{N}$$

$$Q_{xy} = \sum_i^N x_i \cdot y_i - \frac{(\sum_i^N x_i) \cdot (\sum_i^N y_i)}{N}$$

Der Quellcode für das BASIC-Programm, welches die lineare Regression durchführt, befindet sich im Anhang.

b) Zufallszahlenverteilung
Durch ein BASIC-Programm sollen Zufallszahlen zwischen 0 und 1 erzeugt werden. Da alle Zahlen in diesem Intervall mit gleicher Wahrscheinlichkeit auftreten sollten, ist bei genügend hoher Stichprobenzahl eine sogenannte *Rechteckverteilung* der Häufigkeit dieser Zahlen zu erwarten.

c) Verteilungen nach Bernoulli und Gauss
Es wird der Fall betrachtet, dass mit einer Anzahl von Würfeln N eine Anzahl von Würfen K durchgeführt wird. Untersucht werden soll, wie die Häufigkeitsverteilung des Ereignisses, „eine 6“ zu würfeln, von N und K abhängt.
Der Parameter K wird als der **Stichprobenumfang** bezeichnet, wobei für K gegen unendlich von der sogenannten **Grundgesamtheit** gesprochen wird. Der Parameter N bestimmt die Häufigkeitsverteilung, die erhalten wird, wobei für kleine Werte von N eine Bernoulli-Verteilung und für größere eine Gauss-Verteilung erhalten wird.

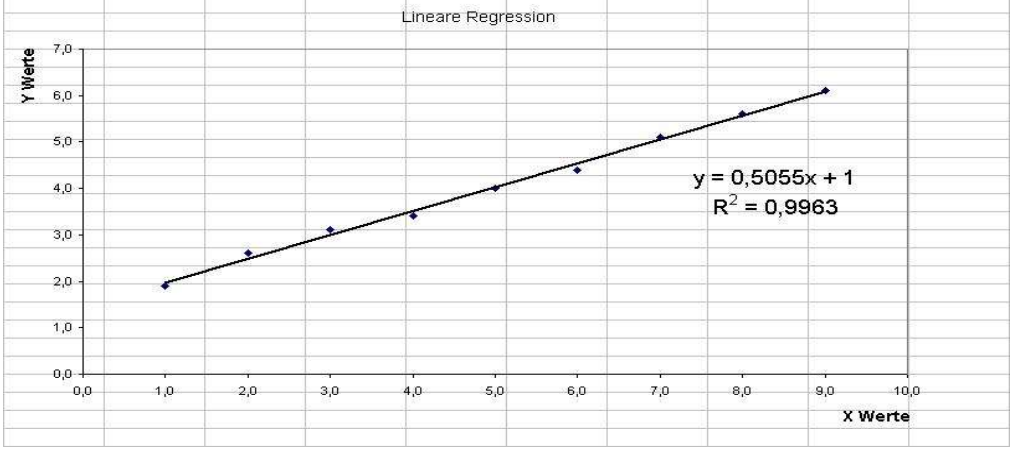
Der Quellcode des entsprechenden BASIC-Programmes befindet sich ebenfalls im Anhang des Protokolls.

Durchführung und Auswertung:

a) Lineare Regeression

Die Durchführung der linearen Regression lieferte folgendes Ergebnis:
(oberste Zahlenreihe: x-Koordinaten, zweite Zahlenreihe: y-Koordinaten)

X:	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0
Y:	1,9	2,6	3,1	3,4	4,0	4,4	5,1	5,6	6,1



Man erkennt, dass das Programm gut funktioniert hat, die Punkte werden durch die Gerade mit einem Korrelationskoeffizienten von 99,6 % angenähert.

b) Zufallszahlenverteilung

Die Durchführung der Zufallszahlen-Verteilung ergibt folgendes Ergebnis:
(oberste Zahlenreihe: Anzahl der Stichproben, Werte nach unten: Ergebnisse für die Ereignisse „0“ - „9“)

Man erkennt, dass bei höherer Stichprobenzahl die Verteilung der Zahlen immer mehr der Rechtecksform entspricht. Das Diagramm oben rechts zeigt das Ergebnis für 1000 Stichproben, das unten rechts für 1 Million Stichproben. Die Annäherung zur Rechtecksform ist deutlich erkennbar.

1000	10000	100000	1000000
93	1015	10034	100223
90	971	9817	99532
113	1053	10051	99926
106	1032	9973	100021
118	985	10045	100678
112	978	10046	99491
95	1014	10080	99659
99	981	9939	100223
83	985	9980	100134
91	986	10027	100113

c) Bernulli- und Gaußverteilung, Stichprobenumfang

Es wurden folgende Fälle betrachtet:

N = 50, K = 10; 100; 1000; 10000; 100000; 1000000

was zu einer Gaussverteilung führt,
sowie

N = 10, K = 1000000

was zu einer Bernoulli-Verteilung führt.
Ergebnistabelle siehe Anhang, Graphen der Häufigkeitsverteilung:

Aus dem Graphen erkennt man, dass für höhere Anzahl der Würfel die Gaussverteilung entstanden ist, während bei N = 10 selbst bei einer Stichprobenzahl von 1 Million deutlich die Bernoulli-Verteilung zu erkennen ist.
Die Mittelwerte für beide Verteilungen errechnet sich nach der Formel 30.16 im Skript:

$$\bar{n} = \frac{1}{K} \cdot \sum_i^K n_i$$

Für die Grundgesamtheit errechnet sie sich nach der Formel:

$$\bar{n} = N \cdot p$$

Für die Bernoulliverteilung erhält man die Standardabweichungen durch die Formen 30.18:

$$\sigma = \sqrt{\frac{1}{K} \cdot \left(\sum_i^K n_i^2 - \frac{1}{K} \cdot \left(\sum_i^K n_i\right)^2\right)}$$

Für die Gaußverteilung berechnet sich die Standardabweichungen durch die Formel 30.20:

$$\sigma = \sqrt{\frac{1}{K-1} \cdot \sum_i^K (n_i - \bar{n})^2}$$

v.d.Hoff
Wagner

Anhang: Quelltext der Programme:

Regressionsprogramm:

```
DIM X(1000), Y(1000)
INPUT "Anzahl der Messwert-Paare (max. 1000): ";Z
IF Z > 1000 THEN STOP
FOR I = 1 TO Z
        INPUT X(I)
        INPUT Y(I)
NEXT I
SUMX2=0
SUMX=0
SUMXY=0
SUMY=0
FOR I = 1 TO Z
        SUMX2 = SUMX2 + X(I)*X(I)
        SUMX = SUMX + X(I)
        SUMXY = SUMXY + X(I)*Y(I)
        SUMY = SUMY + Y(I)
NEXT I
Qx = SUMX2 - (SUMX * SUMX)/Z
Qxy = SUMXY - (SUMX * SUMY)/Z
B = Qxy/Qx
A = (SUMY - B*SUMX)/Z
CALL LRPLOT(X(),Y(),Z,A,B)
END
```

Bernoulli-Gauss-Programm:

```
DIM H(1000)
FOR I = 1 TO 1000
        H(I) = 0
NEXT I
INPUT "Anzahl Würfel N = ";N
INPUT "Anzahl Würfe K = ";K
FOR L = 1 TO K
        I = 0
        FOR J = 1 TO N
        X = RND
        IF X > 0 AND X <= 1/6 THEN I = I+1
NEXT J
H(I+1) = H(I+1) + 1
NEXT L
CALL RNDPLOT(H())
END
```

Anmerkung: Die von den Programmen aufgerufenen Funktionen RNDPLOT, LRPLOT und RND sind BASIC-Funktionen, die nicht eigens definiert worden sind.

Tabellenwerk und Graphen für den Versuch: Bernoulli- und Gaußverteilung, Stichprobenumfang:

50/10	ni		50/100	ni		50/1000	ni		50/10000	ni	
4	4	1	2	2	1	2	12	6	0	0	1
5	0	0	3	0	0	3	54	18	1	8	8
6	12	2	4	12	3	4	128	32	2	96	48
7	14	2	5	40	8	5	390	78	3	528	176
8	16	2	6	114	19	6	690	115	4	1580	395
9	18	2	7	105	15	7	1071	153	5	3820	764
10	0	0	8	152	19	8	1192	149	6	6738	1123
11	0	0	9	144	16	9	1215	135	7	10108	1444
12	0	0	10	60	6	10	1130	113	8	12096	1512
13	13	1	11	55	5	11	847	77	9	12564	1396
Summe ni:	77		12	24	2	12	696	58	10	11120	1112
Summe ni^2:	1105,00		13	39	3	13	442	34	11	9493	863
			14	14	1	14	210	15	12	6732	561
			15	30	2	15	120	8	13	3913	301
			Summe ni:	791		16	112	7	14	2072	148
			Summe ni^2	79427		17	34	2	15	1290	86
						Summe ni:	**8343**		16	704	44
						Summe ni^2	7438063		17	204	12
									18	54	3
									19	57	3
									Summe ni:	**83177**	
									Summe ni^2	750020007	

50/100000	ni		50/1000000	ni		10/1000000	H	
0	0	12	0	0	0	0	0,161772	161772
1	92	92	1	1124	1124	1	0,323009	323009
2	1032	516	2	10784	5392	2	0,291225	291225
3	5232	1744	3	51645	17215	3	0,154629	154629
4	15852	3963	4	160880	40220	4	0,053905	53905
5	37605	7521	5	372915	74583	5	0,013084	13084
6	68046	11341	6	672324	112054	6	0,002105	2105
7	98007	14001	7	984816	140688	7	0,000247	247
8	121600	15200	8	1207776	150972	8	0,000024	24
9	125802	13978	9	1266102	140678			
10	116010	11601	10	1154310	115431			
11	92092	8372	11	926871	84261			
12	65724	5477	12	654360	54530			
13	40729	3133	13	416741	32057			
14	23422	1673	14	234668	16762			
15	12165	811	15	122700	8180			
16	5424	339	16	56896	3556			
17	2635	155	17	23902	1406			
18	828	46	18	10134	563			
19	361	19	19	3040	160			
20	80	4	20	1120	56			
21	42	2	21	315	15			
Summe ni:	**832780**		22	44	2			
Summe ni^2	7,52E+10		23	23	1			
			Summe ni:	**8333490**	**Summe ni^2**	7,52E+12		

Für 50/1000000

n(Anzahl "6")	2	3	4	5	6	7	8	9	10	11	12	13
Verteilung BASIC	5392	17215	40220	74583	112054	140688	150972	140678	115431	84261	54530	32057
Rekursionsfaktor	3,200	2,350	1,840	1,500	1,257	1,075	0,933	0,820	0,727	0,650	0,585	0,529
P(Rekursion)	0,0054	0,0172	0,0405	0,0745	0,1118	0,1405	0,1510	0,1410	0,1156	0,0841	0,0546	0,0319
P(Gauss)	0,0054	0,0172	0,0405	0,0745	0,1118	0,1405	0,1510	0,1410	0,1156	0,0841	0,0546	0,0319
Verteilung Gauss	5384	17230	40490	74502	111753	140490	151027	140958	115586	84062	54641	31944
Fehler [%]	-0,142	0,087	0,668	-0,108	-0,269	-0,141	0,036	0,199	0,134	-0,236	0,202	-0,355
P(BASIC)	0,005392	0,017215	0,04022	0,074583	0,112054	0,140688	0,150972	0,140678	0,115431	0,084261	0,05453	0,032057
Vert Basic *n	10784	51645	160880	372915	672324	984816	1207776	1266102	1154310	926871	654360	416741

14	15	16	17	18	19	20	21	22	23
16762	8180	3556	1406	563	160	56	15	3	1
0,480	0,438	0,400	0,367	0,337	0,310	0,286	0,264	0,243	0,225
0,0169	0,0081	0,0035	0,0014	0,0005	0,0002	0,0001	0,0000	0,0000	0,0000
0,0169	0,0081	0,0035	0,0014	0,0005	0,0002	0,0001	0,0000	0,0000	0,0000
16885	8105	3546	1418	520	175	54	16	4	1
0,726	-0,931	-0,289	0,867	-8,260	8,661	-3,124	3,321	26,667	-0,409
0,016762	0,00818	0,003556	0,001406	0,000563	0,00016	0,000056	0,000015	0,000003	0,000001
234668	122700	56896	23902	10134	3040	1120	315	66	23

Für 10/100000

(Anzahl "6")	1	2	3	4	5	6	7	8
erteilung BASIC	323009	291225	154629	53905	13084	2105	247	24
ekursionsfaktor	0,9	0,533333333	0,35	0,24	0,166666667	0,114285714	0,075	0,044444444
(Rekursion)	0,323011166	0,290710049	0,15504536	0,054265876	0,01302381	0,002170635	0,000248073	1,86054E-05
(Gauss)	0,323011166	0,290710049	0,15504536	0,054265876	0,01302381	0,002170635	0,000248073	1,86054E-05
erteilung Gauss	323011	290710	155045	54266	13024	2171	248	19
ehler [%]	-0,165	0,001	-0,177	0,269	0,665	-0,462	3,024	0,432
(BASIC)	0,323009	0,291225	0,154629	0,053905	0,013084	0,002105	0,000247	0,000024
ert. Basic * n	323009	582450	463887	215620	65420	12630	1729	192

I	K	Mittelwert	Standarabweichung
50	∞	8 1/3	2,6352
50	1000000	8,332388	2,635
50	100000	8,3278	2,639
50	10000	8,3177	2,6233
50	1000	8,343	2,6424
50	100	7,91	2,4041
50	10	7,7	2,406
10	∞	1 2/3	1,1785
10	1000000	1,664937	1,1775

Graph: