

Protokoll zu

Versuch 30: Statistik

1. Versuchsziel

Mit Hilfe von in „BASIC“ geschriebenen Computerprogrammen sollen Ausgleichsgeraden erstellt werden sowie über einen Zufallsgenerator der Unterschied zwischen einer Bernoulli- und einer Gaußverteilung demonstriert werden. Des weiteren sind die bei einer Gaußverteilung erhaltenen Mittelwerte und Standardabweichungen in Abhängigkeit vom Stichprobenumfang zu diskutieren.

2. Theoretische Aspekte und dazugehörige Aufgaben

a) Lineare Regression

Erhält man bei einer Messung verschiedene Messwerte, die annähernd durch eine Linie verbunden werden können, so ist es oft günstig, diese durch eine Ausgleichsgerade zu mitteln. Dieses Verfahren nennt man lineare Regression.

Die allgemeine Gleichung für eine solche Ausgleichsgerade besitzt die Form

$$y = a + bx \quad ,$$

wobei b = Steigung

a = y-Achsenabschnitt für $x = 0$

Betrachtet man nun einen Messpunkt mit y -Wert = y_2 , der nicht auf der Ausgleichsgeraden liegt, so gilt für dessen y -Differenz bezüglich seines auf der Ausgleichsgeraden liegenden Referenzpunktes:

$$\Delta y = a + bx - y_2$$

Diese Differenz wird im Folgenden als „Fehler“ bezeichnet.

Für eine optimale Ausgleichsgerade muss nun gelten, dass die Summe der Fehlerquadrate über alle Messpunkte möglichst klein wird, d.h. die partiellen Ableitungen von

$$S = \sum_i^N (a + bx_i - y_i)^2$$

nach a und b müssen minimiert werden:

Es ergibt sich für

$$a = (\sum_i^N y_i - b \sum_i^N x_i) / N$$

$$\text{und } b = Q_{xy}/Q_x$$

mit

$$Q_x = \sum_i^N x_i^2 - (\sum_i^N x_i)^2 / N$$

$$Q_y = \sum_i^N y_i^2 - (\sum_i^N y_i)^2 / N$$

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

Praktikumsaufgabe:

Es ist eine lineare Regression mit folgenden x- und y-Werten durchzuführen:

x	1,0	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10
y	1,6	1,9	2,6	3,1	3,4	4,0	4,4	5,1	5,6	6,1

Der hierzu nötige Quelltext für das in BASIC erstellte Programm befindet sich im Anhang.

Nachstehend einige Erläuterungen diesbezüglich:

- ↻ mit *DIM x(1000), y(1000)* wird das zugehörige Feld für die x-Werte und y-Werte definiert, es ist auf eine Wertepaaranzahl von 1000 beschränkt.
- ↻ *INPUT „Anzahl der Messwerte (max.1000)“; Z* legt die Eingabeaufforderung bei Programmstart fest
- ↻ Falls die begrenzte Anzahl von *Z = 1000* überschritten werden sollte, so wird das Programm an dieser Stelle durch den Befehl *IF Z > 1000 THEN STOP* beendet
- ↻ Im Folgenden werden Operationsschleifen bezüglich der x- und y-Werte definiert, welche für die Berechnung von Q_x , Q_y und Q_{xy} notwendig sind: es werden zunächst die zu definierenden Variablen auf null zurückgesetzt, dann folgt der Ausdruck für die Schleifen, und schließlich die Befehlsformeln für die Summen der „x-Werte“, der „y-Werte“ und der „x-bzw. y-Werte zum Quadrat“ sowie der „Produkte xy“.

```
xs = 0
ys = 0 usw.
FOR I = 1 TO Z
INPUT „x-Wert“; x(I)
INPUT „y-Wert“; y(I)
xs = xs + x(I)
ys = ys + y(I)
```

```

xxs = xxs + x(I) * x(I)
yys = yys + y(I) * y(I)
xys = xys + x(I) * y(I)

```

- ⌘ *NEXT I* beendet die Formulierung der „Schleife“.
- ⌘ Schließlich folgen die aus den obigen Parametern zusammengesetzten Variablen Q_x , Q_y und Q_{xy} fest, welche folgendermaßen definiert werden:

$q_x = xxs - x_s * x_s / Z$ (dieser Ausdruck entspricht vollkommen (30.6.) im Skript)

usw..

- ⌘ Zum Schluss werden nun die Variablen für die Steigung b und den y -Abschnitt a formuliert:

```

B = qxy / qx
A = (ys - B * xs) / Z

```

- Die Kommandos *PRINT A* und *PRINT B* bewirken, dass am Ende des Programms die Werte für die Steigung b und den y -Abschnitt a am Bildschirm angezeigt werden.
- ⌘ Die letzten Zeilen des Codes rufen das Graphikprogramm auf und beenden schließlich das Programm.

b) Zufallszahlenverteilung

Im Folgenden sollen durch ein einfaches Programm Zufallszahlen zwischen 0,0 und 1,0 „gezogen“ werden. Da alle Zahlen in diesem Intervall mit gleicher Wahrscheinlichkeit auftreten, ist bei genügend hoher „Loszahl“ eine „Rechteckverteilung“ der Häufigkeit dieser Zahlen zu erwarten. Der Quellcode für das hierzu erforderliche Programm befindet sich wieder im Anhang, Erläuterungen hierzu (bereits aus Teil a) bekannte Formulierungen wurden jetzt nicht mehr berücksichtigt):

- ⌘ *X = RND* fordert das Programm auf, Zufallszahlen zu „ziehen“.
- ⌘ *IF X > (J-1) / 10 AND X <= J / 10 THEN H(J) = H(J) + 1* legt fest, dass alle auftretenden Zahlen I durch den Wert J einem bestimmten Intervall (Klasse) zugeordnet werden, wobei diese Intervalle jeweils eine Größe von $1/10$ besitzen und zwischen 0,0 und 1,0 liegen (also 0,0 – 0,1; 0,1 – 0,2; ..., 0,9 – 1,0).
- ⌘ *CALL RNDPLOT(H())* ruft das Graphikprogramm auf.

c) Bernoulli- und Gaußverteilung, Stichprobenumfang

Man betrachte den Fall, dass mit N Würfeln K mal gewürfelt wird. Es ist nun zu untersuchen, wie die Häufigkeitsverteilung des Ereignisses „eine 6“ zu würfeln von N und K abhängt.

Unter K versteht man hierbei auch den Stichprobenumfang.
Für unendlich hohe K spricht man von der sog. Grundgesamtheit.
In Abhängigkeit von der Größe von N erhält man verschieden aussehende Häufigkeitsverteilungen, welche im Fall für kleine N (z.B. $N = 10$) als Bernoulli-Verteilung und für größere N (z.B. $N = 50$) als Gaußverteilung bezeichnet werden.

Um diese Verteilungen durch ein BASIC Programm darstellen zu können, ist wieder der im Anhang beigefügte Quellcode zu betrachten.
Erklärungen hierzu:

- ☞ zunächst werden die Eingabeaufforderungen für die Anzahl der Würfel N und der Würfe K festgelegt:
INPUT „ANZAHL DER Würfel“; K
INPUT „ANZAHL DER Würfel“; N
- ☞ anschließend werden zwei Schleifen für die Hintereinanderausführung der K Würfe mit N Würfeln definiert, wobei
FOR I = 1 TO K
für die Schleife bezüglich der K Würfe mit jeweils I Sechsen steht und
FOR J = 1 TO N
für die N Würfel pro Wurf K, d.h. diese Schleife ist in die I-Schleife integriert und steht deshalb zwischen
FOR I NEXT I.
- ☞ *X = RND*
ruft wiederum den Zufallsgenerator auf,
- ☞ wobei *IF X < (1 / 6) THEN R = R + 1*
festlegt, dass jede Zahl, die kleiner als $1 / 6$ ist, als Würfelzahl 6 gewertet wird und am Ende eines Wurfs durch
H (R+1) = H(R+1) + 1
der jeweiligen Klasse für die pro Wurf erzielte Anzahl an Sechsen zugeordnet wird.
- ☞ Die Schleifen werden jeweils mit
NEXT J und *NEXT I* beendet

Es wurden mit dem Programm folgende Fälle simuliert:

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

Dies Fälle ergeben eine Gaußverteilung.

$N = 10, K = 1000000;$

Dies ergibt eine Bernoulli-Verteilung.

Es ist nun die Gaußverteilung mit der Bernoulli-Verteilung für die letzten beiden Fälle zu vergleichen.

Hierzu zunächst die Ergebnisse des Experiments:

Tabelle:

<i>Anzahl „Sechser“ pro Wurf</i>	<i>Häufigkeitsverteilung auf 1000000 Würfe für N = 50 (Gauß)</i>	<i>Häufigkeitsverteilung auf 1000000 Würfe für N = 10 (Bernoulli)</i>
0	1124	161772
1	5392	323001
2	17210	291224
3	40220	254630
4	74582	53905
5	112055	13084
6	140687	2105
7	150973	247
8	140678	24
9	115430	0
10	84262	0
11	54530	
12	32057	
13	16762	
14	8180	
15	3550	
16	1406	
17	563	
18	160	
19	56	
20	25	
21	23	

Mittelwerte und Standardabweichungen hierzu:

für $N = 50$:

Mittelwert: $\check{n} = (\sum_i^N n_i) / N = 7,33$

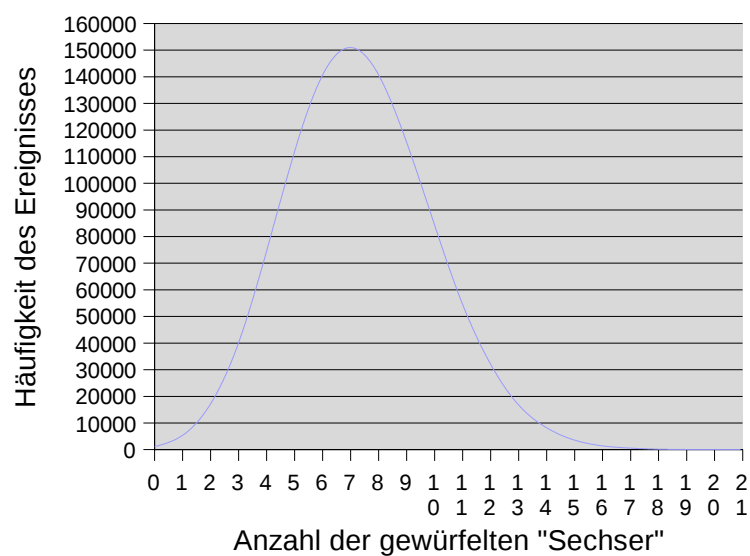
Standardabweichung $\sigma = \sqrt{\sum_i^N (n_i - \check{n})^2 / (N - 1)}$
 $= \sqrt{4048509,38 / 999999} = 2,01$

für $N = 10$:

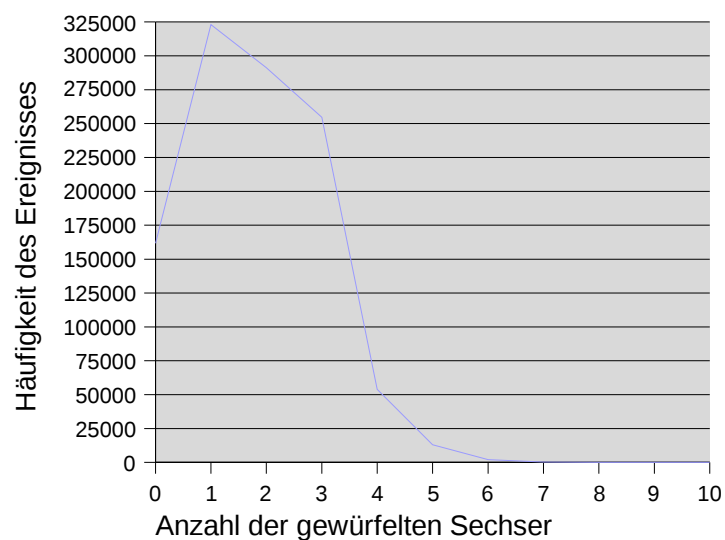
$\check{n} = (\sum_i^N n_i) / N = 1,97$

Standardabweichung $\sigma = \sqrt{\sum_i^N (n_i - \check{n})^2 / (N - 1)} = 1,26$

Gaußverteilung für $N = 50$



Bernoulli-Verteilung für $N = 10$



Vergleicht man die Gaußverteilung mit der Bernoulli-Verteilung, so stellt man fest, dass die Gaußverteilung die Form einer vom Nullpunkt aus nach rechts verschobenen, auf den Kopf gestellten Normalparabel besitzt, während die Kurve der Bernoulli-Verteilung zwar ähnlich der obigen ist, jedoch auf der linken Seite „abgeschnitten“ wirkt. Dies ist leicht zu verstehen, da im zweiten Fall die Anzahl der Würfel wesentlich geringer ist, wodurch die Wahrscheinlichkeit, genau 0 mal einem Sechs zu würfeln, sehr niedrig wird. Für beide Verteilungen gilt auf Grund der kleinen „a priori“ Wahrscheinlichkeit von 1/6, dass die Wahrscheinlichkeit für das Auftreten vieler „Sechser“ sehr gering und schließlich für Werte größer als 8 bzw. 21 gegen null geht. Im Folgenden sollen nun die theoretischen Werte der Häufigkeitsverteilungen für die Grundgesamtheit (Idealfall: $N \rightarrow \infty$) sowohl für Bernoulli als auch Gauß berechnet werden und mit den tatsächlich erhaltenen Werten verglichen werden.

1. Theoretische Gaußverteilung

Hierzu ist zunächst der Mittelwert für die Grundgesamtheit zu berechnen:

$\bar{n} = Np$, wobei p = „a priori“ Wahrscheinlichkeit = 1/6
 N = Anzahl Würfel

$\rightarrow \bar{n} = 50 \cdot 1/6 = 8,33$

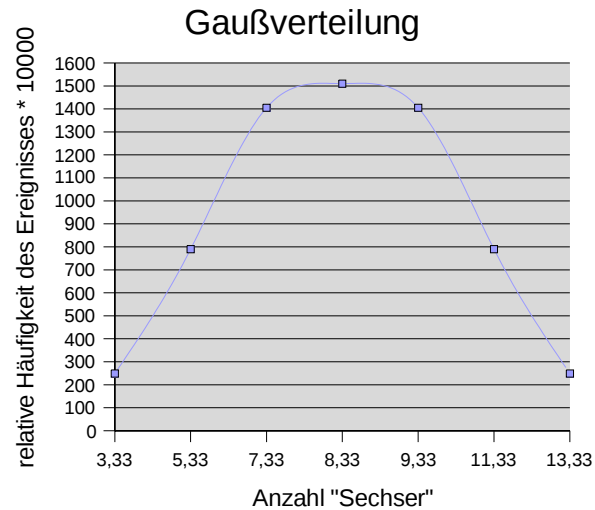
Nun folgt die Berechnung für die Standardabweichung:

$\sigma = \sqrt{Npq} = \sqrt{50 \cdot 1/6 \cdot 5/6} = 2,64$,

wobei q = Wahrscheinlichkeit für Nicht-Eintreten des Ereignisses =
 $= 5/6 = 1 - 1/6$

Durch Einsetzen des Mittelwerts und der Standardabweichung in die Formel für die Berechnung der Kurve der Gaußverteilung (30.19. im Skript) erhält man untenstehende Tabelle sowie das zugehörige Diagramm:

x	3,33	5,33	7,33	8,33	9,33	11,33	13,33
W	249	790	1405	1510	1405	790	249



2. Theoretische Bernoulli-Verteilung

Wie oben sind zunächst der Mittelwert und die Standardabweichungen für eine Würfelanzahl von $N = 10$ zu berechnen:

$$\check{n} = 10 * 1/6 = 1,66$$

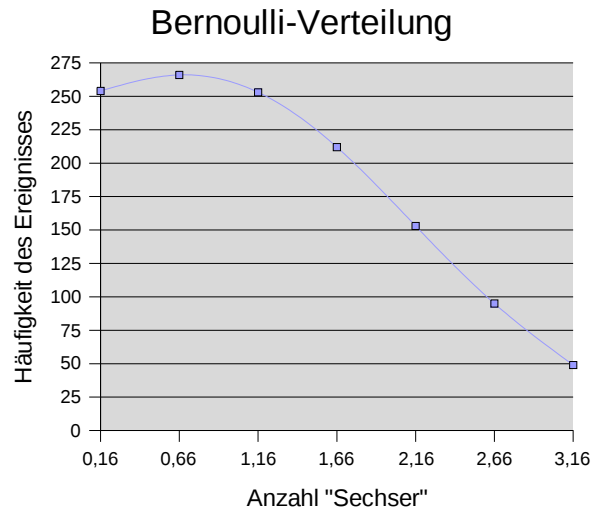
$$\sigma = \sqrt{Npq} = \sqrt{10 * 1/6 * 5/6} = 1,18$$

Für die Berechnung der Häufigkeitsverteilung ist hier die Rekursionsformel vorteilhafter:

$$W = \{[(N - x) * p] / [(x + 1) * q]\} * W_{\text{Gauß}}$$

Daraus ergeben sich folgende Tabelle sowie das zugehörige Diagramm:

x	0,16	0,66	1,16	1,66	2,16	2,66	3,16
W	254	266	253	212	153	95	49



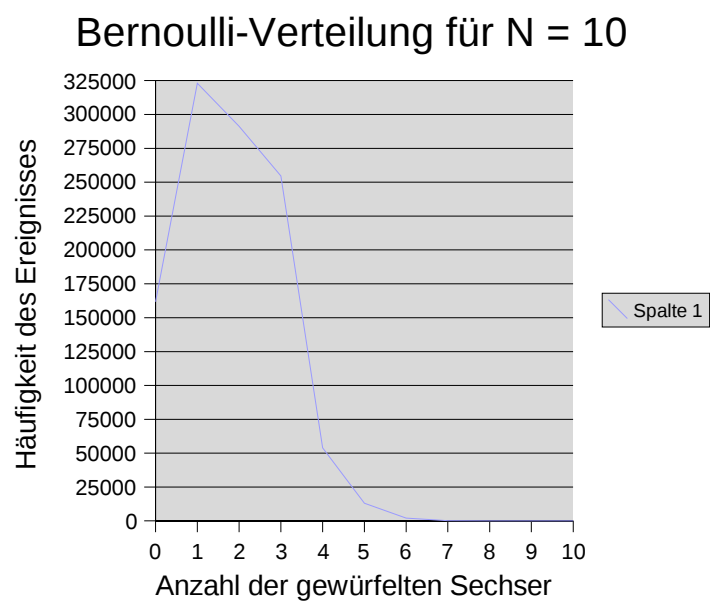
3. Vergleich der theoretischen Ergebnisse mit den tatsächlichen

Zusammenfassende Ergebnistabelle:

<i>Mittelwert Gauß theoretisch</i>	8,33
<i>Mittelwert Gauß praktisch</i>	7,33
<i>Standardabweichung Gauß theoretisch</i>	2,64
<i>Standardabweichung Gauß praktisch</i>	2,01
<i>Mittelwert Bernoulli theoretisch</i>	1,66
<i>Mittelwert Bernoulli praktisch</i>	1,97
<i>Standardabweichung Bernoulli theor.</i>	1,18
<i>Standardabweichung Bernoulli prakt.</i>	1,26

Die zusammenfassende Ergebnistabelle zeigt deutlich, dass die theoretischen Formeln zur Beschreibung der Verteilungsverhältnisse im Versuch relativ gute Ergebnisse liefern, obgleich sie dennoch immer noch geringfügig davon abweichen.

Der Grund dafür liegt klar in der Tatsache, dass es sich eben um eine begrenzte Anzahl an „Würfen“ handelt, und nicht um eine „Grundgesamtheit“.



<i>y</i>	<i>150</i>	<i>236</i>	<i>309</i>	<i>338</i>	<i>309</i>	<i>236</i>	<i>150</i>
----------	------------	------------	------------	------------	------------	------------	------------

<i>0</i>	<i>161772</i>
----------	---------------

<i>0</i>	<i>161772</i>
<i>1</i>	<i>323001</i>
<i>2</i>	<i>291224</i>
<i>3</i>	<i>254630</i>
<i>4</i>	<i>53905</i>
<i>5</i>	<i>13084</i>
<i>6</i>	<i>2105</i>
<i>7</i>	<i>247</i>
<i>8</i>	<i>24</i>
<i>9</i>	<i>0</i>
<i>10</i>	<i>0</i>