

Prüfung eines Datenbestandes

auf Abweichungen einzelner Zahlen vom erwarteten mathematisch-statistischen Verhalten,
die nicht mit einem Zufall erklärbar sind

(Prüfung auf Manipulationen des Datenbestandes)

Inhaltsverzeichnis

1.	Grundlagen der Auswertung	2
2.	Auswertung der ersten führenden Ziffer	3
3.	Auswertung der ersten beiden führenden Ziffern	5
4.	Auswertung der 1.+2. Vor- und Nachkomma-Ziffer	8
5.	ausgewertete Daten mit Kennzeichnung besonders auffälliger Zahlen	12
6.	Schlussfolgerungen aus der Auswertung	13
7.	Anhang	14
7.1.	Ziffern-Wahrscheinlichkeiten lt. Benford-Verteilung.....	14
7.2.	Chi-Quadrat-Verteilung der verwendeten Freiheitsgrade	16
7.3.	Z-Werte zur Normalverteilung für einige Wahrscheinlichkeiten	16

1. Grundlagen der Auswertung

Für die Prüfung der Daten werden die statistischen Gesetzmäßigkeiten:

- der Benford-Verteilung der ersten beiden Ziffern,
- die Normalverteilung der ersten beiden Vor- und Nachkommstellen,
- der Chi-Quadrat-Test sowie
- der Gaußsche Z-Test angewendet.

Für die Auswertung wurde lediglich eine Aufzählung der zur Verfügung gestellten Zahlen verwendet. Somit konnten Beziehungen zu anderen Bezugsdaten (z.B. Rechnungs-, Lieferanten-, Bestell-Nummern usw.) nicht geprüft werden.

Es wurden 100% der Daten ausgewertet, d.h., eine mögliche Filterung der Daten wurde nicht vorgenommen.

Signifikante Abweichungen der Häufigkeit einzelner Zahlen von der erwarteten Häufigkeit entsprechend der statistischen Gesetzmäßigkeiten werden farblich gekennzeichnet.

In der Auswertung wird von einer 95%igen Sicherheit der Aussage, d.h., einer Irrtumswahrscheinlichkeit von 5%, ausgegangen.

In den Grafiken sind diejenigen Ziffern / Zahlen mit Rot markiert, deren Häufigkeit in den ausgewerteten Daten mit einer Sicherheit von 95% nicht zufällig entstanden sein kann.

2. Auswertung der ersten führenden Ziffer

Vorgehensweise

- Aus dem Datenbestand wird ermittelt, wie oft die Ziffern 0 - 9 als erste Ziffer vorkommen.
- Aus der erwarteten Häufigkeit pro Ziffer und der vorgefundenen Häufigkeit wird ein Chi-Quadrat Wert ermittelt.
- Die Summe der einzelnen Chi-Quadrat-Werte ist größer als der zur Irrtumswahrscheinlichkeit gehörende Wert.
- Für die Gesamtheit aller Daten kann festgestellt werden, dass die in den Daten vorgefundene Häufigkeit der ersten Ziffern mit einer Sicherheit von 95% nicht mit zufälligen Abweichungen erklärbar ist.
- Für jede einzelne Ziffer wurde mit dem Z-Test ermittelt, ob die Differenz aus erwarteter und vorgefundener Häufigkeit noch zufällig entstanden sein kann.
- Für die Ziffern 3, 8 und 9 kann festgestellt werden, dass die Häufigkeit dieser Ziffern zufällig ist.
- Die Ziffern 1, 2 und 7 wurden zu selten verwendet, als dass die Häufigkeit dieser Ziffern noch mit einem Zufall zu erklären wäre.
- Die Ziffern 4, 5 und 6 wurden dagegen häufiger als erwartet verwendet.

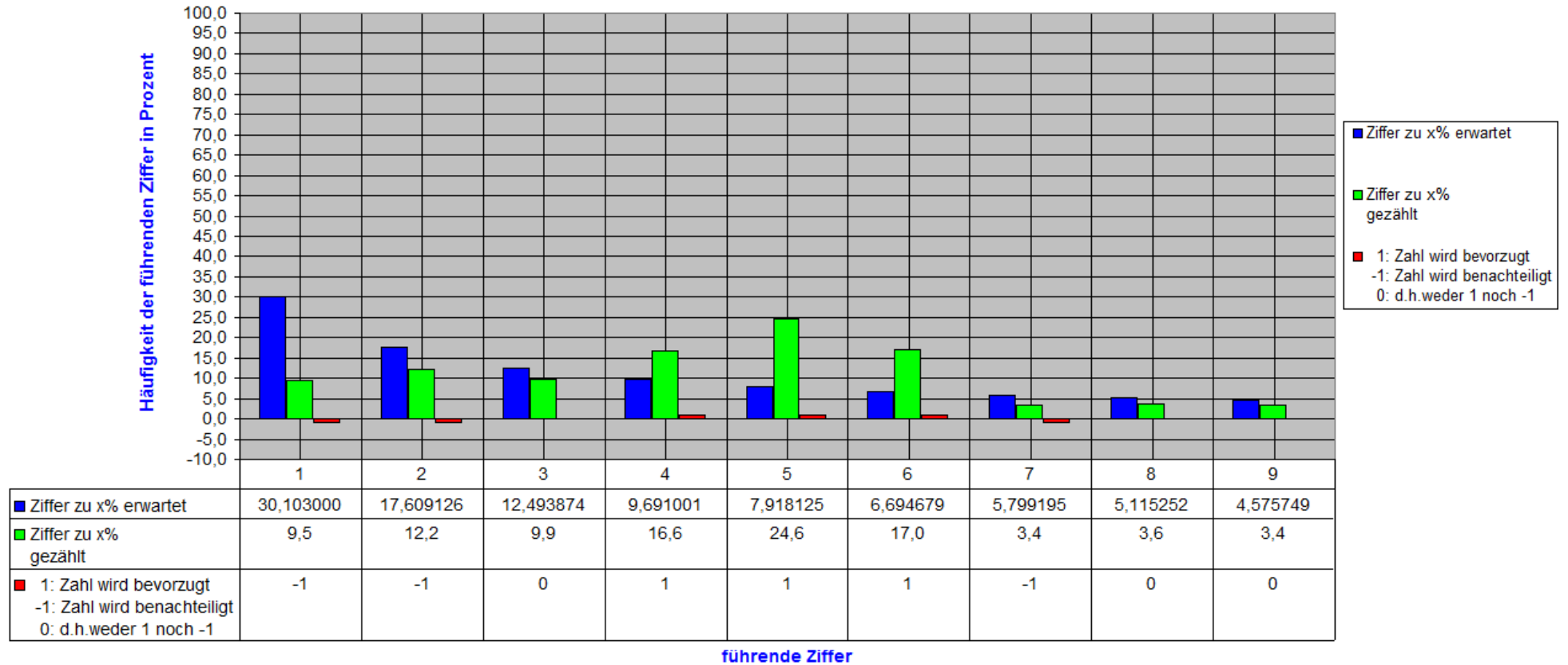
Tabelle: Auswertung der 1. führenden Ziffer

Ermittlung auffälliger Daten mit Statistischen Methoden (Benford-Gesetz, Chi-Quadrat-Test, Gauß-Test)				
Prüfung der ersten führenden Ziffer				
	Benford-Wert		Chi-Wert	Gauß-Test
führende linke Ziffer	Ziffer zu x% erwartet	Ziffer zu x% gezählt	Chi-Quadrat-Wert je Ziffer	Z-Wert
1	30,103000	9,5	67,16	9,77
2	17,609126	12,2	8,05	3,05
3	12,493874	9,9	2,44	1,66
4	9,691001	16,6	23,67	5,02
5	7,918125	24,6	164,24	13,38
6	6,694679	17,0	75,03	8,92
7	5,799195	3,4	5,14	2,18
8	5,115252	3,6	2,04	1,42
9	4,575749	3,4	1,64	1,16
		Summe	349,41	1,96 Z-Wert
		Chi-Wert	15,51	
Ergebnis pro Ziffer: grün + rot: es sind signifikante Abweichungen vorhanden grün: Ziffer wird bevorzugt rot: Ziffer wird zu selten verwendet weiß: d.h. weder grün noch rot				
Vorgaben: Chi-Wert für 8 Freiheitsgrade; Chi-Wert(15,51) u. Z-Wert(1,96) für eine Irrtumswahrscheinlichkeit von 5%				
Ergebnis für alle Ziffern:				
Farbe: grün: keine signifikanten Abweichungen rot: signifikante Abweichungen sind vorhanden				

Grafik: Auswertung der 1. führenden Ziffer

Ermittlung auffälliger Daten mit Statistischen Methoden:
Prüfung der Häufigkeit der ersten führenden Ziffer

Ergebnis: (rot markierte Ziffern): Die Häufigkeit dieser Ziffern kann mit 95%iger Sicherheit nicht mit zufälligen Abweichungen erklärt werden



3. Auswertung der ersten beiden führenden Ziffern

- Aus dem Datenbestand wird ermittelt, wie oft die Ziffern 10 - 99 als führende Ziffern verwendet wurden.
- Wie zu den führenden Ziffern 0 - 9 wurde per Chi-Quadrat- und Z-Test auf eine Signifikanz der Abweichungen getestet.
- Zu den rot und grün markierten Zahlen kann festgestellt werden, dass die in den Daten vorgefundene Häufigkeit der ersten Ziffern mit einer Sicherheit von 95% nicht mit zufälligen Abweichungen erklärbar ist.
- die rot markierten Zahlen wurden zu selten verwendet, die grün markierten Zahlen hingegen bevorzugt.
- In nachfolgenden Prüfungen ist zu ermitteln, warum die Ziffern 1 und 2 seltener und die Zahlen im Bereich von 45 - 66 zu häufig verwendet wurden.

Tabelle: Auswertung der ersten beiden führenden Ziffern

Prüfung der ersten beiden führenden Ziffern					
führende zwei Ziffern	Benford-Wert		Chi-Wert	Gauß- Test	Ergebnis pro Zahl: grün + rot: es sind signifikante Abweichungen vorhanden grün: Ziffer wird bevorzugt rot: Ziffer wird zu selten verwendet weiß: d.h. weder grün noch rot
	Ziffern zu x% erwartet	Ziffern zu x% gezählt	Chi-Quadrat- Wert je Zahl	Z-Wert	
10	4,139269	1,6	81	2,28	rot
11	3,778856	0,3	169	3,39	
12	3,476211	0,5	121	2,92	
13	3,218468	0,3	121	3,05	
14	2,996322	0,5	81	2,6	
15	2,802872	1,4	25	1,51	
16	2,632894	0,5	64	2,34	
17	2,482358	0,3	64	2,55	
18	2,348110	0,5	49	2,11	
19	2,227639	0,3	49	2,36	
20	2,118930	0,5	36	1,91	rot
21	2,020339	0,3	36	2,19	
22	1,930516	1,1	9	0,98	weiß
23	1,848341	1,9	0	0,08	
24	1,772877	0,5	25	1,58	
25	1,703334	0,3	25	1,92	
26	1,639042	0,5	16	1,45	
27	1,579427	0,3	25	1,8	
28	1,523997	0,8	9	0,89	
29	1,472326	0,3	16	1,69	
30	1,424044	1,6	1	0,12	

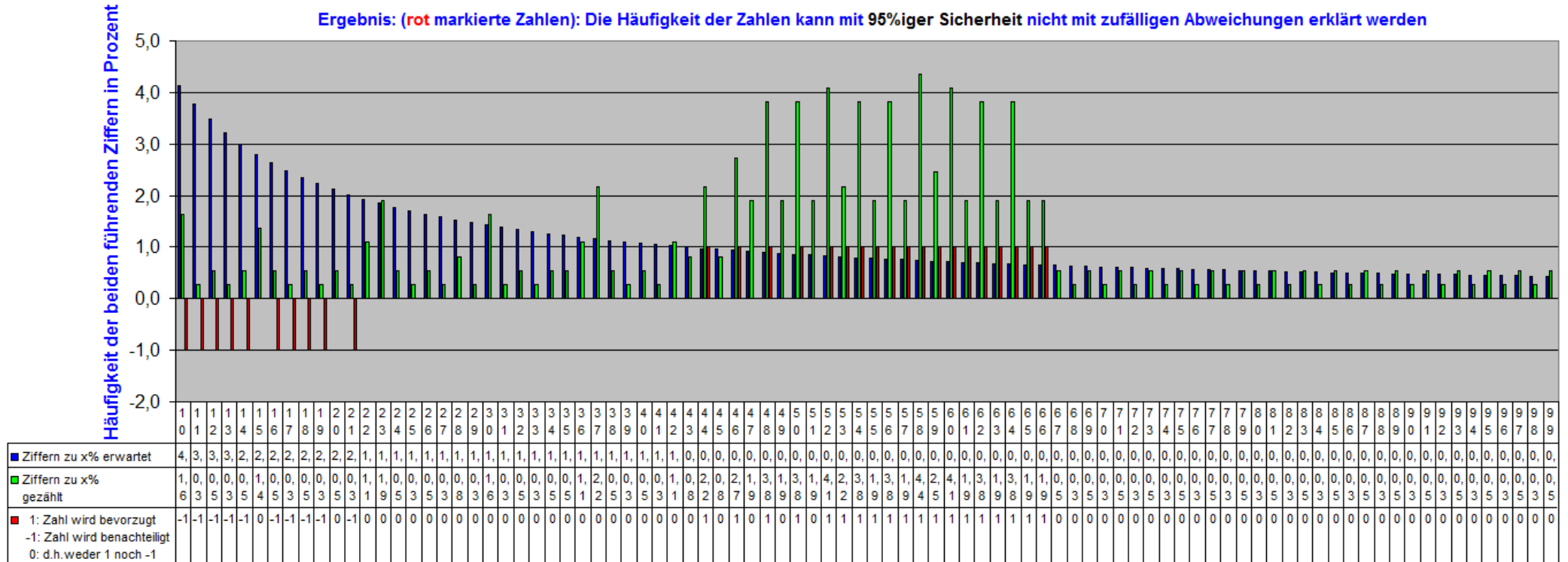
...

40	1,072387	0,5	4	0,73
41	1,046543	0,3	9	1,2
42	1,021917	1,1	0	0,13
43	0,998422	0,8	1	0,09
44	0,975984	2,2	16	2,08
45	0,954532	0,8	1	0
46	0,934003	2,7	49	3,3
47	0,914338	1,9	16	1,72
48	0,895484	3,8	121	5,66
49	0,877392	1,9	16	1,84
50	0,860017	3,8	121	5,85
51	0,843317	1,9	16	1,94
52	0,827253	4,1	144	6,61
53	0,811789	2,2	25	2,63
54	0,796893	3,8	121	6,21
55	0,782534	1,9	16	2,15
56	0,768683	3,8	121	6,38
57	0,755314	1,9	16	2,25
58	0,742402	4,4	169	7,77
59	0,729924	2,5	36	3,57
60	0,717858	4,1	144	7,34
...				
95	0,454763	0,5	0	0,26
96	0,450050	0,3	1	0,12
97	0,445434	0,5	0	0,29
98	0,440912	0,3	1	0,09
99	0,436481	0,5	0	0,32
Summe:	100,000000			

Grafik: Auswertung der ersten beiden führenden Ziffern

**Ermittlung auffälliger Daten mit Statistischen Methoden:
Prüfung der Häufigkeit der ersten beiden führenden Ziffern**

Ergebnis: (rot markierte Zahlen): Die Häufigkeit der Zahlen kann mit 95%iger Sicherheit nicht mit zufälligen Abweichungen erklärt werden



■ Ziffern zu x% erwartet
 ■ Ziffern zu x% gezählt
 ■ 1: Zahl wird bevorzugt
 -1: Zahl wird benachteiligt
 0: d.h. weder 1 noch -1

4. Auswertung der 1.+2. Vor- und Nachkomma-Ziffer

- Aus dem Datenbestand wird ermittelt, wie oft die Ziffern 0 - 9 als Vor- und Nachkommaziffer verwendet wurden.
- Wie zu den führenden Ziffern 0 - 9 wurde per Chi-Quadrat- und Z-Test auf eine Signifikanz der Abweichungen getestet.
- Die Summen aus den einzelnen Chi-Quadrat-Werten liegen deutlich über den zum Signifikanzniveau zulässigen Wert.
- Die Häufigkeit einzelner Ziffern an diesen Ziffern-Positionen weicht ebenfalls zu deutlich von den üblichen Häufigkeiten ab.

Ermittlung auffälliger Daten mit Statistischen Methoden (Chi-Quadrat-Test, Gauß-Test)

Prüfung der 1.Vorkomma-Ziffer

1. Vorkomma-Ziffer	erwartete Häufigkeit	gemessene Häufigkeit	Chi-Wert	Gauß-Test	Ergebnis pro Ziffer:
	Ziffer zu x% erwartet	Ziffer zu x% gezählt	Chi-Quadrat-Wert je Ziffer	Z-Wert	grün + rot: es sind signifikante Abweichungen vorhanden grün: Ziffer wird bevorzugt rot: Ziffer wird zu selten verwendet weiß: d.h. weder grün noch rot
0	10	11,5	1,1	5,54	
1	10	10,9	0,4	5,64	
2	10	16,2	18,9	4,83	
3	10	9,2	0,3	5,88	
4	10	13,3	5,5	5,26	
5	10	6,6	5,7	6,29	
6	10	9,9	0,0	5,79	
7	10	7,0	4,4	6,23	
8	10	8,8	0,7	5,95	
9	10	6,6	5,7	6,29	
Summe			42,6	1,96 Z-Wert	
Chi-Wert			16,92		

Vorgaben:
Chi-Wert für 9 Freiheitsgrade;
Chi-Wert(16,92) u. Z-Wert(1,96) für eine Irrtumswahrscheinlichkeit von 5%

Ergebnis für alle Ziffern:		grün: keine signifikanten Abweichungen rot: signifikante Abweichungen sind vorhanden
-----------------------------------	--	---

Prüfung der 2.Vorkomma-Ziffer

	erwartete Häufigkeit	gemessene Häufigkeit	Chi-Wert	Gauß-Test	Ergebnis pro Ziffer: grün + rot: es sind signifikante Abweichungen vorhanden grün: Ziffer wird bevorzugt rot: Ziffer wird zu selten verwendet weiß: d.h. weder grün noch rot	
2. Vorkomma-Ziffer	Ziffer zu x% erwartet	Ziffer zu x% gezählt	Chi-Quadrat-Wert je Ziffer	Z-Wert		
0	10	1,6	25,7	6,01		
1	10	4,6	10,6	5,49		
2	10	5,2	8,5	5,40		
3	10	7,9	1,6	4,92		
4	10	15,0	9,1	3,69		
5	10	30,5	154,5	0,99		
6	10	20,7	42,1	2,70		
7	10	5,2	8,5	5,40		
8	10	4,6	10,6	5,49		
9	10	4,6	10,6	5,49		
		Summe	281,8	1,96 Z-Wert		
		Chi-Wert	16,92			

Vorgaben:
Chi-Wert für 9 Freiheitsgrade;
Chi-Wert(16,92) u. Z-Wert(1,96) für eine Irrtumswahrscheinlichkeit von 5%

Ergebnis für alle Ziffern:		grün: keine signifikanten Abweichungen rot: signifikante Abweichungen sind vorhanden
-----------------------------------	--	---

Prüfung der 1.Nachkomma-Ziffer

1. Nachkomma Ziffer	erwartete Häufigkeit	gemessene Häufigkeit	Chi-Wert	Gauß- Test	Ergebnis pro Ziffer: grün + rot: es sind signifikante Abweichungen vorhanden grün: Ziffer wird bevorzugt rot: Ziffer wird zu selten verwendet weiß: d.h. weder grün noch rot
	Ziffer zu x% erwartet	Ziffer zu x% gezählt	Chi-Quadrat- Wert je Ziffer	Z-Wert	
0	10	10,0	0,0	5,72	
1	10	6,2	6,8	6,29	
2	10	10,6	0,2	5,62	
3	10	17,5	26,8	4,58	
4	10	5,0	12,1	6,48	
5	10	7,3	3,6	6,13	
6	10	18,1	31,5	4,49	
7	10	4,8	13,1	6,51	
8	10	5,4	10,2	6,41	
9	10	15,2	12,9	4,93	
		Summe	117,0	1,96 Z-Wert	
		Chi-Wert	16,92		

Vorgaben:
Chi-Wert für 9 Freiheitsgrade;
Chi-Wert(16,92) u. Z-Wert(1,96) für eine Irrtums-
Wahrscheinlichkeit von 5%

Ergebnis für alle Ziffern:		grün: keine signifikanten Abweichungen rot: signifikante Abweichungen sind vorhanden
---	--	--

Prüfung der 2.Nachkomma-Ziffer

2. Nachkomma Ziffer	erwartete Häufigkeit	gemessene Häufigkeit	Chi-Wert	Gauß-Test	Ergebnis pro Ziffer: grün + rot: es sind signifikante Abweichungen vorhanden grün: Ziffer wird bevorzugt rot: Ziffer wird zu selten verwendet weiß: d.h. weder grün noch rot
	Ziffer zu x% erwartet	Ziffer zu x% gezählt	Chi-Quadrat-Wert je Ziffer	Z-Wert	
0	10	0,0	44,6	6,96	
1	10	11,2	0,7	5,19	
2	10	11,2	0,7	5,19	
3	10	8,1	1,7	5,69	
4	10	12,6	2,9	4,98	
5	10	14,1	7,6	4,73	
6	10	10,1	0,0	5,37	
7	10	8,7	0,7	5,58	
8	10	13,5	5,3	4,84	
9	10	10,5	0,1	5,30	
		Summe	64,2		
		Chi-Wert	16,92	1,96 Z-Wert	

Vorgaben:
Chi-Wert für 9 Freiheitsgrade;
Chi-Wert(16,92) u. Z-Wert(1,96) für eine Irrtumswahrscheinlichkeit von 5%

Ergebnis für alle Ziffern:		grün: keine signifikanten Abweichungen rot: signifikante Abweichungen sind vorhanden
-----------------------------------	--	---

5. ausgewertete Daten mit Kennzeichnung besonders auffälliger Zahlen

In der nachfolgenden Übersicht sind Auszüge des ausgewerteten Datenbestandes dargestellt. In der rechten Spalte sind Zeilen mit rot markiert, in denen Werte vorhanden sind, zu denen es mehrere Auffälligkeiten gibt.

Daten	Auswertung						
	grün markierte Daten gehören zu den bevorzugten Zahlen (Ziffern)						gehäufte Auffälligkeit
	1.führende linke Ziffer	1.+2.führende linke Ziffer	2.Vorkomma-Ziffer	1.Vorkomma-Ziffer	1.Nachkomma-Ziffer	2.Nachkomma-Ziffer	
99,16							
99,83							
100,5							
101,17							
101,84							
102,51							
46,9							
47,57							
48,24							
48,91							
49,58							
50,25							
50,92							
51,59							
52,26							
52,93							
53,6							
54,27							
54,94							
55,61							
56,28							
56,95							
57,62							
58,29							
58,96							
59,63							
60,3							
60,97							
61,64							
62,31							
62,98							
63,65							
64,32							
64,99							

6. Schlussfolgerungen aus der Auswertung

- Im Datenbestand gibt es beim Vergleich der erwarteten und tatsächlich vorgefundenen Häufigkeiten Abweichungen, die nicht mehr einem Zufall erklärbar sind.
- Da diese Abweichungen an allen geprüften Ziffern-Stellen auftreten, besteht die begründete Vermutung, dass die Daten bewusst geändert wurden.
- In nachfolgenden Prüfungen wäre zu ermitteln, warum einzelne Zahlenbereiche zu oft vorkommen.

7. Anhang

7.1. Ziffern-Wahrscheinlichkeiten lt. Benford-Verteilung

führende 1.Ziffer (x)	Wahrscheinlichkeit P für das Auftreten der Ziffer: $P = \text{LOG}_{10}(1+1/x)$
1	0,301030
2	0,176091
3	0,124939
4	0,096910
5	0,079181
6	0,066947
7	0,057992
8	0,051153
9	0,045757
10	0,041393
11	0,037789
12	0,034762
13	0,032185
14	0,029963
15	0,028029
16	0,026329
17	0,024824
18	0,023481
19	0,022276
20	0,021189
21	0,020203
22	0,019305
23	0,018483
24	0,017729
25	0,017033
26	0,016390
27	0,015794
28	0,015240
29	0,014723
30	0,014240
31	0,013788
32	0,013364
33	0,012965
34	0,012589
35	0,012234
36	0,011899
37	0,011582
38	0,011281
39	0,010995
40	0,010724
41	0,010465
42	0,010219
43	0,009984
44	0,009760
45	0,009545
46	0,009340
47	0,009143

48	0,008955
49	0,008774
50	0,008600
51	0,008433
52	0,008273
53	0,008118
54	0,007969
55	0,007825
56	0,007687
57	0,007553
58	0,007424
59	0,007299
60	0,007179
61	0,007062
62	0,006949
63	0,006839
64	0,006733
65	0,006631
66	0,006531
67	0,006434
68	0,006340
69	0,006249
70	0,006160
71	0,006074
72	0,005990
73	0,005909
74	0,005830
75	0,005752
76	0,005677
77	0,005604
78	0,005532
79	0,005463
80	0,005395
81	0,005329
82	0,005264
83	0,005201
84	0,005140
85	0,005080
86	0,005021
87	0,004963
88	0,004907
89	0,004853
90	0,004799
91	0,004746
92	0,004695
93	0,004645
94	0,004596
95	0,004548
96	0,004501
97	0,004454
98	0,004409
99	0,004365

7.2. Chi-Quadrat-Verteilung der verwendeten Freiheitsgrade

Freiheitsgrad	P = (1 - a), a = 0,05	P = (1 - a), a = 0,01	P = (1 - a), a = 0,01
1	3,841	6,635	10,828
2	5,991	9,210	13,816
3	7,815	11,345	16,266
4	9,488	13,277	18,467
5	11,070	15,086	20,515
6	12,592	16,812	22,458
7	14,067	18,475	24,322
8	15,507	20,090	26,124
9	16,919	21,666	27,877
10	18,307	23,209	29,588

7.3. Z-Werte zur Normalverteilung für einige Wahrscheinlichkeiten

Beispiel-Fragestellung:

Gesucht ist der z-Wert, unterhalb dem 95% aller möglichen z-Werte liegen.
(Hinw.: 1-seitiger Test)

Lösung:

Es wird eine Wahrscheinlichkeit (1-a) gesucht, die möglichst nah an den Wert 0,95 herankommt.

In der Tabelle der Standard-Normalverteilung sind das die Werte 0,9495 und 0,9505 mit den z-Werten 1,64 und 1,65, durch Interpolation erhält man den gesuchten Wert z=1,645.

	Fläche (1- α)									
	0,65	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
z-Wert	0,385	0,52	0,67	0,842	1,04	1,282	1,645	1,96	2,326	2,576