Faustregeln für das Sammelbilderproblem

Vorbemerkung: Bereits in einem früheren Beitrag (Januar 2016) auf www.mathematik-ist-schoen.de hatte ich mich mit dem Problem der vollständigen Serie beschäftigt. Auf die wichtigsten Aspekte dieses Beitrags gehe ich zunächst noch einmal ein.

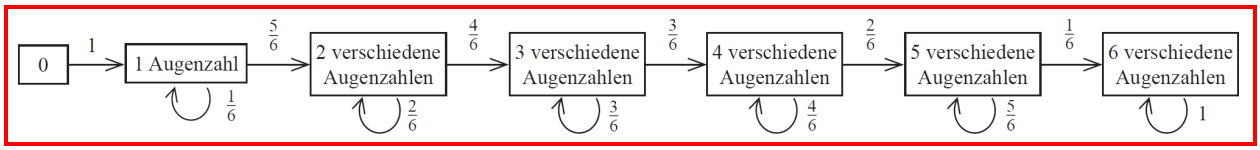
Beim *Warten auf eine vollständige Serie* geht es darum, einen Zufallsversuch so lange durchzuführen, bis jedes der möglichen Ergebnisse mindestens einmal aufgetreten ist.

Bei den Überlegungen wird vorausgesetzt, dass alle möglichen Ergebnisse des betrachteten Zufallsversuchs mit gleichen Wahrscheinlichkeiten auftreten. Wie es sich auswirkt, wenn dies nicht der Fall ist, habe ich in einem Beitrag veröffentlicht, der in der Zeitschrift **MNU, Heft 1/2016** abgedruckt wurde. Das pdf dieses Artikels stelle ich gleichzeitig mit diesem Beitrag auf meiner Homepage zum download ein.

- **Warten auf eine vollständige Serie beim regelmäßigen Hexaeder (Würfel)**

Eine vollständige Serie beim Würfeln bedeutet, dass jede der sechs Augenzahlen 1, 2, 3, 4, 5, 6 mindestens einmal gefallen ist; dabei spielt die Reihenfolge, in der die verschiedenen Augenzahlen auftreten, keine Rolle.

Allgemein spricht man davon, dass der Zufallsprozess verschiedene **Zustände** annehmen kann. Dies kann man mithilfe eines **Übergangsdiagramms** beschreiben – hier für das Würfeln:



Mit Sicherheit (also mit der Wahrscheinlichkeit 1 = 100 %) fällt beim ersten Wurf des Würfels eine Augenzahl, die bisher noch nicht gefallen ist.

Dass beim nächsten Wurf die gleiche Augenzahl wieder fällt, hat nur die Wahrscheinlichkeit $^1/_6$; also ist die Wahrscheinlichkeit, dass sich der Prozess nach dem 2. Wurf im Zustand 2 (= 2 verschiedene Augenzahlen) befindet, gleich $^5/_6$.

Wenn das System im Zustand 2 ist, tritt beim nächsten Wurf mit der Wahrscheinlichkeit $^2/_6$ eine Augenzahl auf, die bereits gefallen ist, und mit der Wahrscheinlichkeit $^4/_6$ geht das System in Zustand 3 über, usw.

Der Übergang mit der geringsten Wahrscheinlichkeit ist der Übergang vom Zustand 5 zum Zustand 6, da ja bereits 5 verschiedene Augenzahlen gefallen sind und nur *eine* Augenzahl noch fehlt.

Wenn ein Ergebnis mit der Wahrscheinlichkeit $^1/_6$ auftritt, benötigt man *im Mittel* 6 Versuche, bis es tatsächlich vorliegt. (Die Häufigkeitsinterpretation der Wahrscheinlichkeit besagt: Wenn man den Versuch 600-mal durchführt, kann man mit ungefähr 100-maligem Auftreten des Ergebnisses rechnen, *im Mittel* tritt es also bei jedem 6. Versuch auf.)

Beim Warten auf eine vollständige Serie dauert also der letzte Schritt *im Mittel* am längsten.

Wenn ein bestimmtes Ergebnis mit der Wahrscheinlichkeit $^1/_3$ auftritt, benötigt man *im Mittel* 3 Versuche, bis es tatsächlich vorliegt, usw.

- **Mittlere Wartezeit auf eine vollständige Serie beim Würfeln**

Insgesamt benötigt man also für das Warten auf eine vollständige Serie beim Würfeln im Mittel

$$\mu = \frac{6}{6}+\frac{6}{5}+\frac{6}{4}+\frac{6}{3}+\frac{6}{2}+\frac{6}{1} = 6\cdot\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}\right) = 6\cdot\frac{60+30+20+15+12+10}{60} = \frac{147}{10} = 14{,}7$$

Würfe. Man beachte: Dies ist ein Mittelwert!

Es könnte sein, dass man bereits nach 6 Würfen alle 6 Augenzahlen geworfen hat; es könnte aber auch seeeeeehr lange dauern, bis es endlich so weit ist.

- **Iterative Berechnung der Wahrscheinlichkeiten**

Die Berechnung der Wahrscheinlichkeit P(*m*; *k*), dass sich das System nach *m* Stufen im Zustand *k* befindet, erfolgt rekursiv aus den Wahrscheinlichkeiten $P(m-1;k-1)$ und $P(m-1;k)$ der vorherigen Stufe.

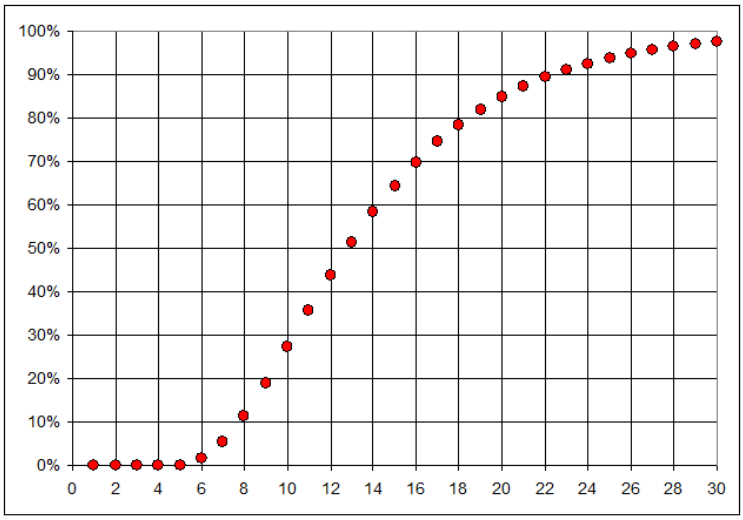
Diese werden mit den Übergangswahrscheinlichkeiten gewichtet, die sich in Abhängigkeit vom jeweiligen Zufallsgerät und dem betrachteten Zustand ergeben:

$P(1;1)=1$ und $P(m;1)=\frac{1}{n}\cdot P(m-1;1)$ für $m \geq 2$

$P(m;k)=\frac{k}{n}\cdot P(m-1;k)+\frac{n-k+1}{n}\cdot P(m-1;k-1)$ für $2 \leq k \leq m$

Mithilfe dieser beiden Rekursionsformeln können die gesuchten Wahrscheinlichkeiten iterativ berechnet werden. Die folgende Tabelle zeigt die Entwicklung der Wahrscheinlichkeiten bei den ersten 30 Würfen, daneben die zugehörige graphische Darstellung.

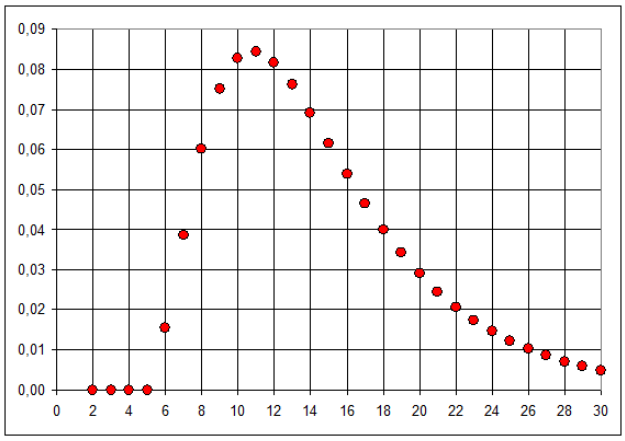
m \ k	1	2	3	4	5	6
1	1	0	0	0	0	0
2	0,167	0,833	0	0	0	0
3	0,028	0,417	0,556	0	0	0
4	0,005	0,162	0,556	0,278	0	0
5	0,001	0,058	0,386	0,463	0,093	0
6	0,000	0,020	0,231	0,502	0,231	0,015
7	0,000	0,007	0,129	0,450	0,360	0,054
8	0,000	0,002	0,069	0,365	0,450	0,114
9	0,000	0,001	0,036	0,278	0,497	0,189
10	0,000	0,000	0,019	0,203	0,506	0,272
11	0,000	0,000	0,009	0,145	0,490	0,356
12	0,000	0,000	0,005	0,101	0,456	0,438
13	0,000	0,000	0,002	0,070	0,414	0,514
14	0,000	0,000	0,001	0,048	0,368	0,583
15	0,000	0,000	0,001	0,032	0,323	0,644
16	0,000	0,000	0,000	0,022	0,280	0,698
17	0,000	0,000	0,000	0,015	0,240	0,745
18	0,000	0,000	0,000	0,010	0,205	0,785
19	0,000	0,000	0,000	0,007	0,174	0,819
20	0,000	0,000	0,000	0,004	0,148	0,848
21	0,000	0,000	0,000	0,003	0,124	0,873
22	0,000	0,000	0,000	0,002	0,105	0,893
23	0,000	0,000	0,000	0,001	0,088	0,911
24	0,000	0,000	0,000	0,001	0,074	0,925
25	0,000	0,000	0,000	0,001	0,062	0,938
26	0,000	0,000	0,000	0,000	0,052	0,948
27	0,000	0,000	0,000	0,000	0,043	0,957
28	0,000	0,000	0,000	0,000	0,036	0,964
29	0,000	0,000	0,000	0,000	0,030	0,970
30	0,000	0,000	0,000	0,000	0,025	0,975



Wir lesen ab: Bereits nach 13 Würfen ist die Wahrscheinlichkeit für das Vorliegen einer vollständigen Serie etwas größer als 50 %. In ungefähr 50 % der Versuchsdurchführungen benötigt man zwischen 10 und 17 Würfe (unteres und oberes Quartil der Verteilung).

Würde man eine Wette abschließen, dass spätestens nach 14 Würfen eine vollständige Serie vorliegt, dann ständen die Chancen für einen Wettsieg etwa wie 7 zu 5 (0,583 : 0,417), obwohl der Erwartungswert μ der Anzahl der notwendigen Würfe für eine vollständige Serie 14,7 beträgt!

Der Grund für das Abweichen des Erwartungswerts μ (= gewichteter Mittelwert einer Verteilung) vom Median x_{Q2} (= 50 %-Wert der Verteilung) ist die fehlende Symmetrie der zugrundeliegenden Wahrscheinlichkeits-verteilung, wie man auch aus der folgenden Grafik ablesen kann. Hier ist dargestellt, wie sich die Wahrscheinlichkeit für das Vorliegen einer vollständigen Serie beim Hexaeder *von Wurf zu Wurf* verändert.

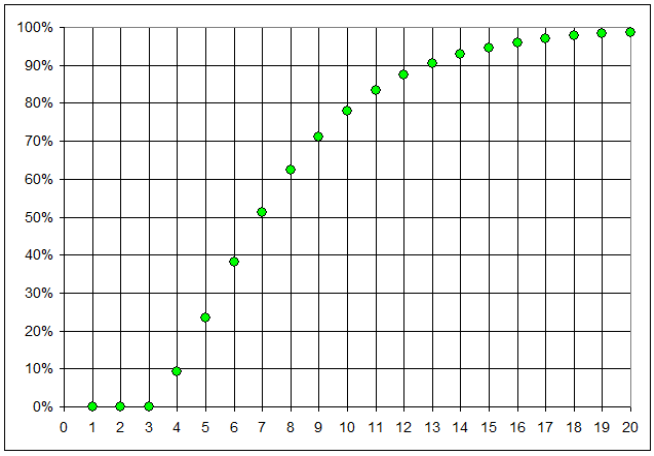


- **Warten auf eine vollständige Serie bei anderen LAPLACE'schen Zufallsgeräten**

Die folgenden Grafiken zeigen die Wahrscheinlichkeiten für das Vorliegen einer vollständige Serie bei verschiedenen Zufallsgeräten, jeweils mit Angabe des Erwartungswerts μ der Anzahl der notwendigen Versuchsdurchführungen, des Medians x_{Q2} (= 2. Quartil) und des 1. Quartils x_{Q1} und des 3. Quartils x_{Q3}.

- **Warten auf eine vollständige Serie beim regelmäßigen Tetraeder (n = 4)**

Aus der folgenden Grafik kann entnommen werden, dass die Wahrscheinlichkeit, nach spätestens 7 Würfen eines regelmäßigen Tetraeders eine vollständige Serie zu haben, etwas größer ist als 50 %, d. h., wenn man eine Wette darauf abschließen würde, wäre man geringfügig im Vorteil.

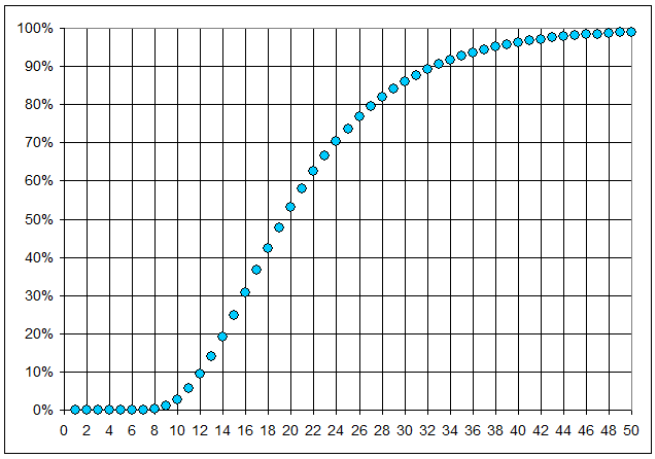


Bemerkenswert ist die Tatsache, dass die Wahrscheinlichkeit, nach spätestens 8 Würfen eine vollständige Serie zu haben, bereits 62 % ist. Auch dies steht scheinbar im Widerspruch zu der Erkenntnis, dass man *im Mittel* $\mu \approx 8{,}3$ Würfe benötigt, um bei den Tetraederwürfen eine vollständige Serie zu erhalten!

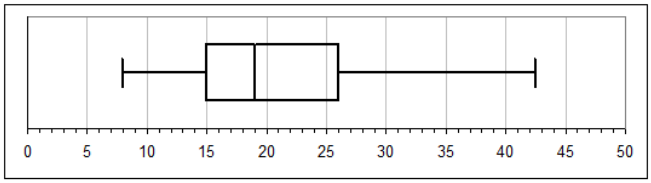
Erwartungswert: $\mu = 4 \cdot \left(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} \right) \approx 8{,}3$ Quartile: $x_{Q1} = 5$; $x_{Q2} = 7$; $x_{Q3} = 11$.

- **Warten auf eine vollständige Serie beim regelmäßigen Oktaeder (n = 8)**

Erwartungswert: $\mu = 8 \cdot \left(\frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{8} \right) \approx 21{,}7$ Quartile: $x_{Q1} = 15$; $x_{Q2} = 19$; $x_{Q3} = 26$.



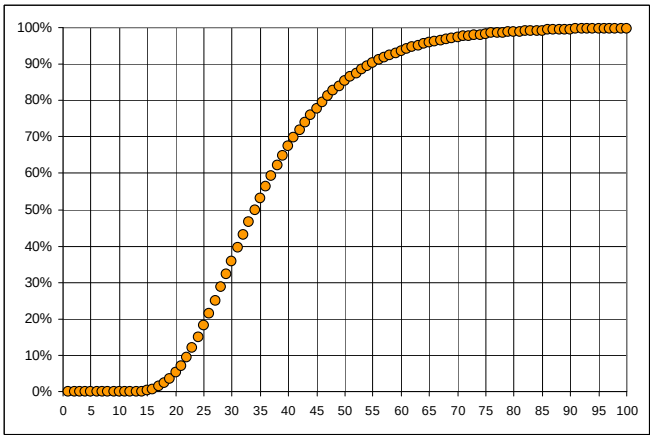
Das zugehörige Boxplot sieht wie folgt aus:

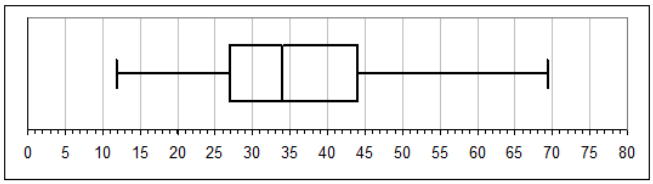


Hinweis: Da kein Maximum existiert, ist es üblich, den oberen Whisker durch das 1,5-fache der Quartilsdifferenz festzulegen.

- **Warten auf eine vollständige Serie beim regelmäßigen Dodekaeder (*n* = 12)**

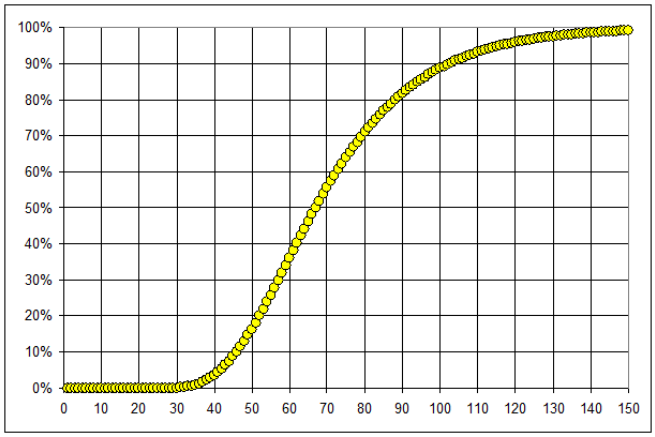
Erwartungswert: $\mu = 12 \cdot \left(\frac{1}{1} + \frac{1}{2} + ... + \frac{1}{12} \right) \approx 37{,}2$ Quartile: $x_{Q1} = 27$; $x_{Q2} = 34$; $x_{Q3} = 44$

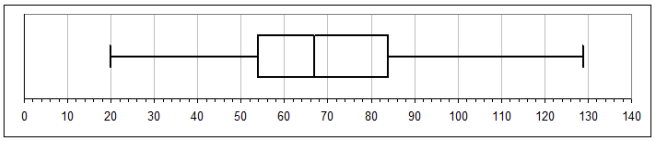




- **Warten auf eine vollständige Serie beim regelmäßigen Ikosaeder (*n* = 20)**

Erwartungswert: $\mu = 20 \cdot \left(\frac{1}{1} + \frac{1}{2} + ... + \frac{1}{20} \right) \approx 72$ Quartile: $x_{Q1} = 54$; $x_{Q2} = 67$; $x_{Q3} = 84$





Warten auf eine vollständige Serie bei anderen Zufallsgeräten

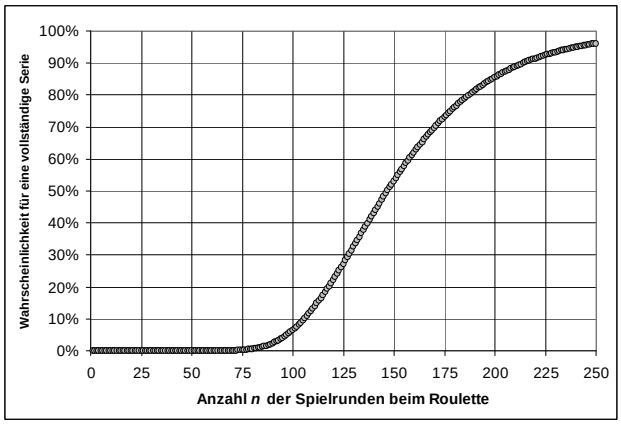
Als Zufallsgeräte kommen natürlich nicht nur Polyederwürfel infrage, sondern allgemein geeignete Glücksräder mit n gleichgroßen nummerierten Sektoren, wie z. B. das Roulette, oder auch Kartenspiele (und zwar als Ziehen mit Zurücklegen).

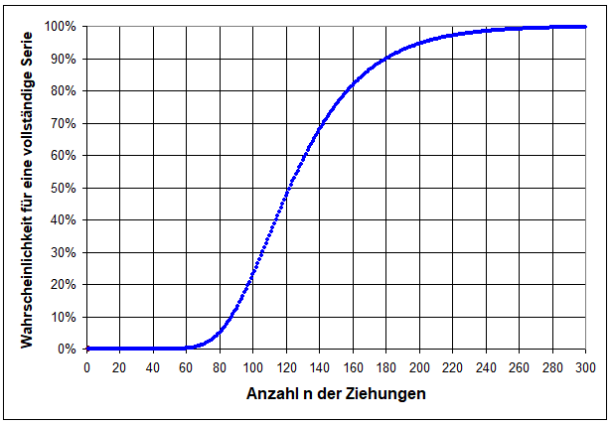
- **Warten auf eine vollständige Serie beim Roulettespiel (n = 37)**

Erwartungswert: $\mu = 37 \cdot \left(\frac{1}{1} + \frac{1}{2} + ... + \frac{1}{37}\right) \approx 155{,}5$ Quartile: $x_{Q1} = 123$; $x_{Q2} = 147$; $x_{Q3} = 178$

- **Ziehen mit Zurücklegen bei einem Skatkartenspiel = Sammeln einer Serie mit 32 Sammelbildern**

Erwartungswert: $\mu = 32 \cdot \left(\frac{1}{1} + \frac{1}{2} + ... + \frac{1}{32}\right) \approx 130$ Quartile: $x_{Q1} = 102$; $x_{Q2} = 122$; $x_{Q3} = 150$





Untersuchung, ob man die Quartile näherungsweise ohne umfangreiche Rechnung bestimmen kann

Für die Suche nach einer allgemeinen Regel für die Lage der Quartile in Abhängigkeit von n wird – wie bei den o. a. Beispielen – eine Tabellenkalkulation mit variabler Anzahl n verwendet, bei der die Wahrscheinlichkeiten mithilfe der o. a. Rekursionsformeln berechnet werden.

Die folgende Tabelle enthält dann für verschiedene Werte von n diejenigen Werte von m, bei denen die Wahrscheinlichkeiten *ungefähr* den drei Quartilen m_{25}, m_{50}, m_{75} entsprechen.

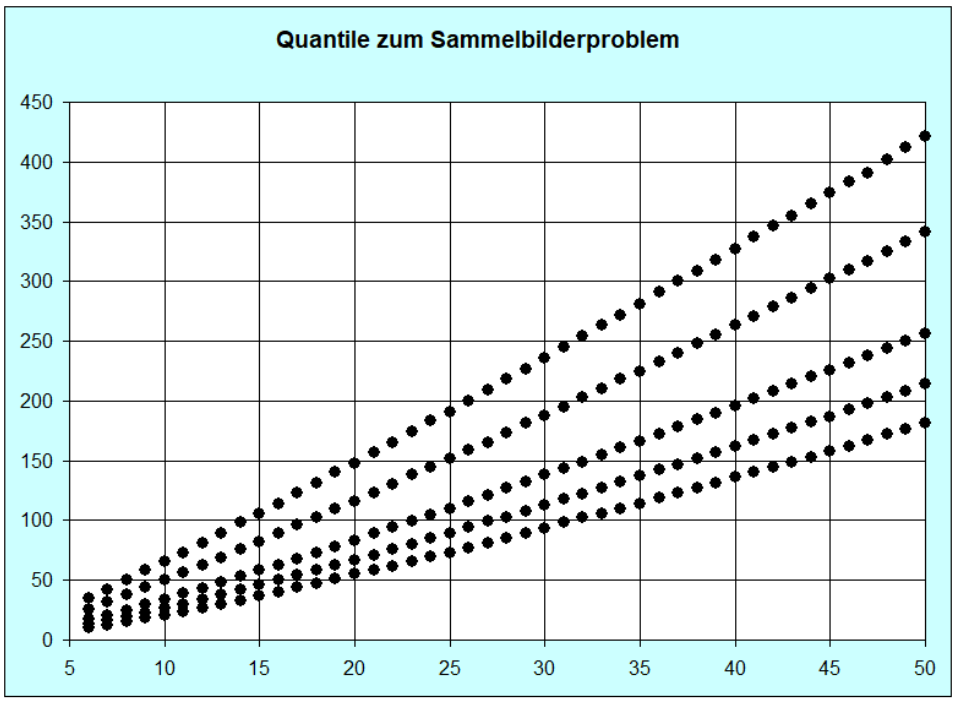
Außerdem sind das 95 %-Quantil und das 99 %-Quantil bestimmt worden, an deren Werten man ablesen kann, mit welch langen Wartezeiten man durchaus rechnen muss.

Diese letzten beiden Werte sind im Hinblick auf eine kritische Beobachtung eines Sammelvorgangs wichtig. Wenn nämlich zu befürchten ist, dass beispielsweise beim Sammeln von Bildern nicht alle Bilder einer Serie dem entsprechenden Kaufobjekt *mit gleichen Häufigkeiten* beigefügt sind, dann verlängert sich die Wartezeit.

Das 95 %-Quantil und das 99 %-Quantil dienen daher als **kritische Werte**, bei deren Überschreitung man von *ungewöhnlichen* Abweichungen vom Mittelwert sprechen könnte, d. h., Ergebnisse, die oberhalb dieser Quantile liegen, können Anlass geben, an der Hypothese der gleichmäßigen Verteilung der Sammelbilder zu zweifeln (im Sinne eines Hypothesentests von Stichprobenergebnissen).

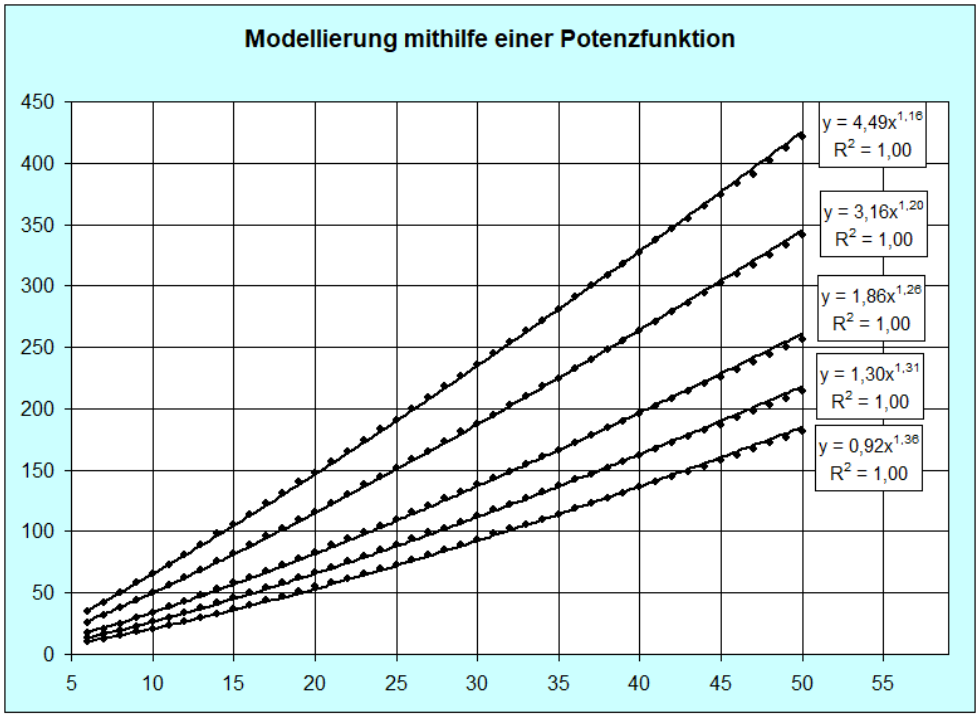
n	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
m_{25}	10	12	15	18	21	24	27	30	33	37	40	44	47	51	55
m_{50}	13	16	19	23	27	30	34	38	42	46	50	54	58	63	67
m_{75}	17	21	25	30	34	39	43	48	53	58	63	68	73	78	83
m_{95}	26	32	38	44	50	56	63	69	76	82	89	96	103	110	116
m_{99}	35	42	50	58	66	73	81	89	98	106	114	123	131	140	148

Diese Tabelle wurde als Grundlage für eine Regressionsrechnung benutzt, die bei der Tabellenkalkulation als Option zur Verfügung steht (für $6 \le n \le 40$).



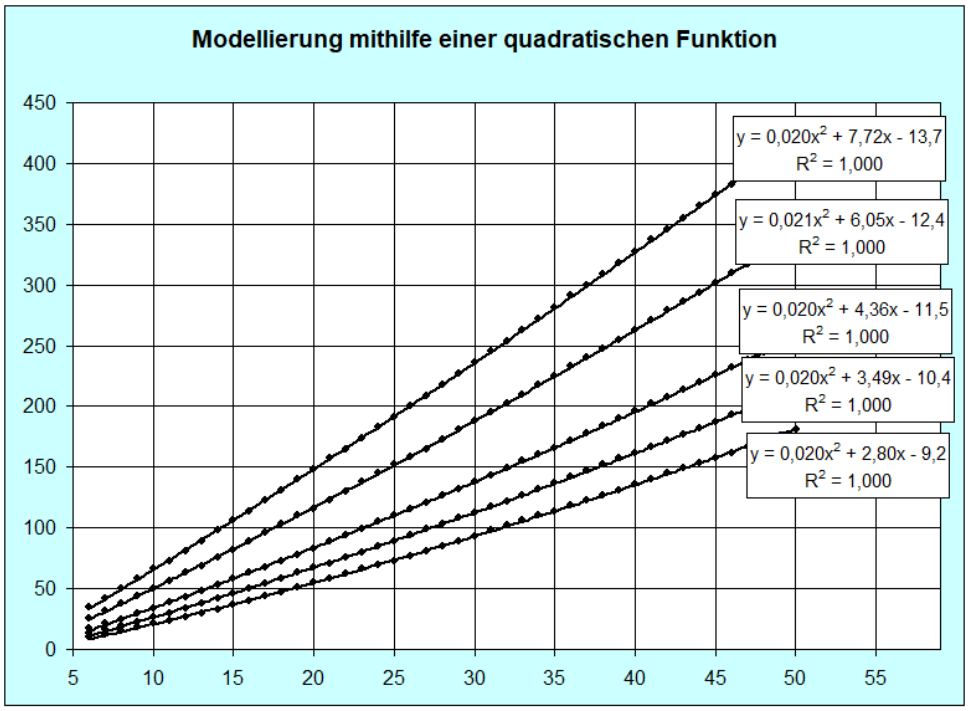
Man stellt fest, dass eine Modellierung mithilfe einer *Potenzfunktion* oder auch mithilfe einer *quadratischen Funktion* hervorragende Anpassungen bietet (Bestimmtheitsmaß $R^2 \geq 99{,}9\ \%$).

Der folgenden Grafik sind jeweils diejenigen fünf Modellierungsfunktionen für m_{25}, m_{50}, m_{75}, m_{95} und m_{99} zu entnehmen, die sich aus der Auswertung für $n = 6$ bis $n = 50$ ergeben.



Zahlenbeispiele zur Kontrolle $n = 20$: $m_{50} \approx 1{,}30 \cdot 20^{1{,}31} \approx 65{,}8$ (*exakt*: 67)

$n = 50$: $m_{50} \approx 1{,}30 \cdot 50^{1{,}31} \approx 218{,}6$ (*exakt*: 214)



Zahlenbeispiele zur Kontrolle $n = 20$: $m_{50} \approx 0{,}020 \cdot 20^2 + 3{,}49 \cdot 20 - 10{,}4 \approx 67{,}4$ (*exakt*: 67)

$n = 50$: $m_{50} \approx 0{,}020 \cdot 50^2 + 3{,}49 \cdot 50 - 10{,}4 \approx 214{,}1$ (*exakt*: 214)

Wenn man für diese Fälle jeweils eine Faustregel entwickeln möchte, also eine Regel, die man überschlagsweise auch ohne Rechenhilfsmittel bestimmen kann, dann kommt als Grundlage nur die Modellierung durch eine quadratische Funktion infrage.

Denn eine auftretende Potenz wie beispielsweise $x^{1,20}$ bei der Modellierung zum 95 %-Quantil eignet sich wohl kaum für das Kopfrechnen.

Bei der quadratischen Modelllierung hingegen kann man folgende Überschlagsrechnung vornehmen:

$$0{,}021 \cdot x^2 + 6{,}05 \cdot x - 12{,}4 \approx \tfrac{1}{50} \cdot x^2 + 6{,}05 \cdot x - 12{,}4 \approx \tfrac{1}{50} \cdot x \cdot (x + 300) - 10$$

Beispiele:

$n = 20$: $\tfrac{1}{50} \cdot 20 \cdot (20 + 300) - 10 = 128 - 10 = 118$ (*exakt*: 116)

$n = 30$: $\tfrac{1}{50} \cdot 30 \cdot (30 + 300) - 10 = 198 - 10 = 188$ (*exakt*: 188)

$n = 40$: $\tfrac{1}{50} \cdot 40 \cdot (40 + 300) - 10 = 272 - 10 = 262$ (*exakt*: 263)

$n = 50$: $\tfrac{1}{50} \cdot 50 \cdot (50 + 300) - 10 = 350 - 10 = 340$ (*exakt*: 341)

Ähnlich brauchbare Ergebnisse erhält man auch für Faustregeln zu den anderen Quantile.

Insgesamt lassen sich also die folgenden Faustregeln zum Sammelbilderproblem formulieren:

- 25 %-Quantil: $0{,}020 \cdot x^2 + 2{,}80 \cdot x - 9{,}2 \approx \tfrac{1}{50} \cdot x^2 + 2{,}80 \cdot x - 9{,}2 \approx \tfrac{1}{50} \cdot x \cdot (x + 140) - 10$
- 50 %-Quantil: $0{,}020 \cdot x^2 + 3{,}49 \cdot x - 10{,}4 \approx \tfrac{1}{50} \cdot x^2 + 3{,}49 \cdot x - 10 \approx \tfrac{1}{50} \cdot x \cdot (x + 175) - 10$
- 75 %-Quantil: $0{,}020 \cdot x^2 + 4{,}36 \cdot x - 11{,}5 \approx \tfrac{1}{50} \cdot x^2 + 4{,}36 \cdot x - 11{,}5 \approx \tfrac{1}{50} \cdot x \cdot (x + 220) - 10$
- 95 %-Quantil: $0{,}021 \cdot x^2 + 6{,}05 \cdot x - 12{,}4 \approx \tfrac{1}{50} \cdot x^2 + 6{,}05 \cdot x - 12{,}4 \approx \tfrac{1}{50} \cdot x \cdot (x + 300) - 10$
- 99 %-Quantil: $0{,}020 \cdot x^2 + 7{,}72 \cdot x - 13{,}7 \approx \tfrac{1}{50} \cdot x^2 + 7{,}72 \cdot x - 13{,}7 \approx \tfrac{1}{50} \cdot x \cdot (x + 390) - 10$