Wilfried Rohm wilfried.rohm@schule.at

Der Run-Test - Einführung in statistische Fragestellungen

Kopfregion

- Mathematische / Fachliche Inhalte in Stichworten:

  Statistische Tests, Binomialverteilung, Simulation, Zufall, Exponentialverteilung, Wartezeitenmodelle

- Kurzzusammenfassung

  Der Run-Test ist eine im Schulunterricht interessante Variante, typische Fragestellungen der Wahrscheinlichkeitsrechnung und Statistik zu behandeln. In diesem File wird eine entsprechende Simulation des Münzwurfes durchgeführt. Ausserdem werden die Simulationen statistisch ausgewertet und mit den Werten verglichen, die man mit Hilfe eines theoretischen Modells erhält. Dieses wird auch kurz hergeleitet.
  Schließlich wird noch eine Verbindung zu Wartezeitenmodellen bzw. zur Exponentialverteilung als Lebensdauerverteilung aufgezeigt.
  Eine ausführlichere Beschreibung der Hintergrunde und Einsatzmöglichkeiten befindet sich im Artikel "Statistik mit Zufallszahlen" von Wilfried Rohm auf der Archivseite www.ammu.at

- Didaktische Überlegungen / Zeitaufwand:

  siehe der Artikel "Statistik mit Zufallszahlen" von Wilfried Rohm auf der Archivseite www.ammu.at

- Lehrplanbezug (bzw. Gegenstand / Abteilung / Jahrgang):

  Angewandte Mathematik, Einsatz in allen Jahrgängen und Abteilungen möglich

- Mathcad-Version:

  erstellt mit Mathcad 11 und Mathcad 14

Kopfregion

Beim Run-Test geht es um folgende Fragestellung:
Eine Münze wird n-mal geworfen (hier n=50).
Das erste Mal als Gedankenexperiment, wobei man versuche, den "Zufall" gedanklich so gut wie möglich nachzubilden. Gezählt wird die Anzahl der "Runs", d.h. der gleichbleibenden Läufe (=Anzahl der Wechsel + 1).
z.B: 11010001011000 sind 8 Runs.
Das zweite Mal soll die Simulation mit Münzwurf tatsächlich durchgeführt werden.
Hinweis: In dem im Vorspann angeführten Artikel auf der Archivseite www.ammu.at wird näher erläutert, warum und wie die "gedachten" Simulationen von den "tatsächlichen" Simulationen meist gut unterschieden werden können.

In diesem File wird nur die Simulation (mit dem Zufallszahlengenerator des Programms) durchgeführt und nach verschiedenen Gesichtspunkten ausgewertet. Typische Fragen, die direkt und indirekt im vorliegenden File beantwortet werden sollen, können beispielsweise sein:

a) Ist die Verteilung der Runs "rein zufällig" oder folgt sie gewissen Gesetzmäßigkeiten?
b) Wieviel Run´s gibt es durchschnittlich ?
c) Wie oft kommen Runs vor, die aus einer Folge von mindestens 5 gleichen Ergebnissen bestand?
d) Wie oft kommen Runs vor, die aus einer Folge von mindestens 10 gleichen Ergebnissen bestand?
e) Welche Länge hatte der längste aufgetretene Run ?
f) Wenn es Gesetzmäßigkeiten gibt: Kann man mit ihnen rechnen? Wie aussagekräftig sind die Ergebnisse im Vergleich zur Simulation?

$N := 1000$ **Anzahl der Simulationen**

Simulation / rechn. Auswertung

$n := 50$ n... Anzahl der Würfe eiines Simulationsdurchganges; n = 50 voreingestellt

$i := 0..n-1$ $j := 0..N-1$ $r_{i,j} := \mathrm{rund}(\mathrm{rnd}(1))$

Es werden N Simulationen durchgeführt mit je n=50 Münzwürfen

$r^T =$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0	0	1	1	0	1	1	0	0	1	0	0	1	0
1	0	0	0	1	0	1	1	1	1	0	0	1	1	1
2	1	1	0	0	0	1	0	1	1	0	1	1	1	1

Auswertung der Simulation:

```
run(j) := | wechsel ← 1
          | for  i ∈ 0.. n − 2
          |    wechsel ← wechsel + 1  if  r[i,j] ≠ r[i+1,j]
          | return  wechsel
```

Hier werden (über die Anzahl der festgestellten Wechsel) die Anzahl der Runs je Simulation ermittelt.

$runs_j := run(j)$

Umdefinition des obigen Funktionsergebnisses in ein Feld, damit leichter eine Klasseneinteilung mit dem hist-Befehl erfolgen kann (siehe unten)

```
anzBereich(a, b) := | anzahl ← 0
                    | for  j ∈ 0.. N − 1
                    |    anzahl ← anzahl + 1  if  runs[j] ≥ a ∧ runs[j] ≤ b
                    | return  anzahl
```

Hier wird ermittelt, wieviele Runs in einem bestimmten Bereich vorkommen (wird unten bei der statistischen Auswertung verwendet)

```
simulation(j) := | anz ← 1
                 | max ← 0
                 | fünf ← 0
                 | zehn ← 0
                 | for  i ∈ 0.. n − 2
                 |    | if  r[i,j] = r[i+1,j]
                 |    |    | anz ← anz + 1
                 |    |    | fünf ← fünf + 1  if  anz = 5
                 |    |    | zehn ← zehn + 1  if  anz = 10
                 |    |    | max ← anz  if  anz > max
                 |    | anz ← 1  otherwise
                 | ergebnisse[0] ← fünf
                 | ergebnisse[1] ← zehn
                 | ergebnisse[2] ← max
                 | return  ergebnisse
```

eser Programmteil bezieht sich auf eine
ızelne Simulation mit 50 Münzwürfen - es
erden ermittelt:
ieviele Runs der Länge 5 oder mehr ?
ieviele Runs der Länge 10 oder mehr?
aximale Länge der Runs

```
versuche := | fünfer ← 0
            | zehner ← 0
            | max ← 0
            | for  j ∈ 0.. N − 1
            |    | fünfer ← fünfer + simulation(j)_0
            |    | zehner ← zehner + simulation(j)_1
            |    | max ← simulation(j)_2  if  simulation(j)_2 > max
            | versuche_0 ← fünfer
            | versuche_1 ← zehner
            | versuche_2 ← max
            | return  versuche
```

…esem Programmteil werden
1000 Simulationen
…glich Fünferserien,
…erien und Maximallänge
…ewertet.

$\text{…ferserien} := \text{versuche}_0$

$\text{…nerserien} := \text{versuche}_1$

$\text{maximallänge} := \text{versuche}_2$

$G_{bi}(x, n, p) := \text{pbinom}(x, n, p)$

$g_{bi}(x, n, p) := \text{dbinom}(x, n, p)$

Umdefinition der Wahrscheinlichkeiten der Binomialverteilung auf die in der HTL übliche Bezeichungsweise

$x_i := i$

Intervallvektor für hist-Befehl

$H := \text{hist}(x, \text{runs})$

$H^T =$

	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0

$h := \frac{H}{N}$

Das sind die relativen Häufigkeiten der Runs

$\text{theorie}_{i+1} := \text{dbinom}(i, n-1, 0.5)$

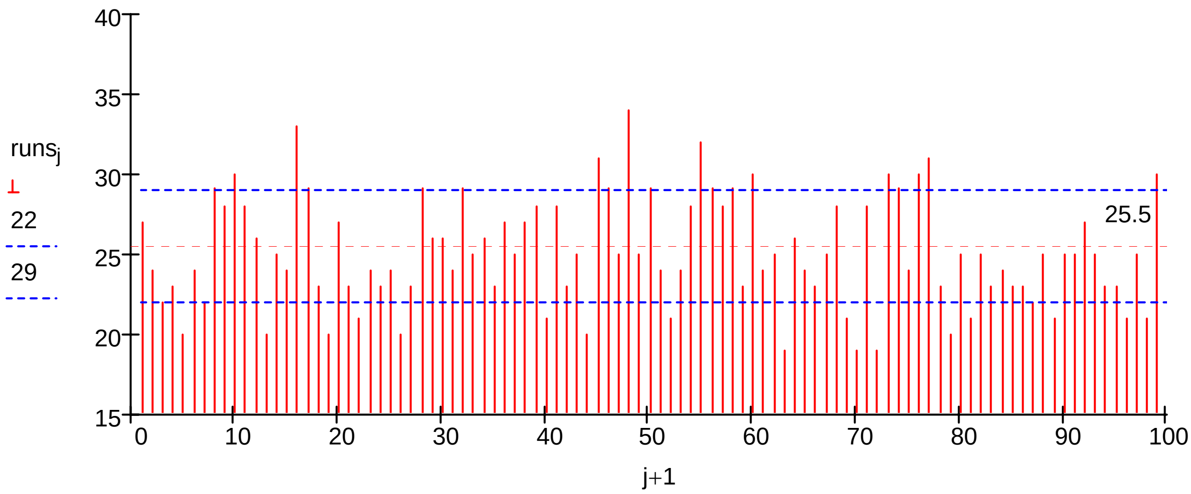
Das ist die theoretische Verteilung der Runs (Begründung siehe unten in der Region "Wahrscheinlichkeiten")

Simulation / rechn. Auswertung

Grafiken und Ergebnisse

Grafik: Die ersten 100 Versuche

Aus Übersichtsgründen werden hier standardmäßig nur die ersten 100 Versuche dargestellt

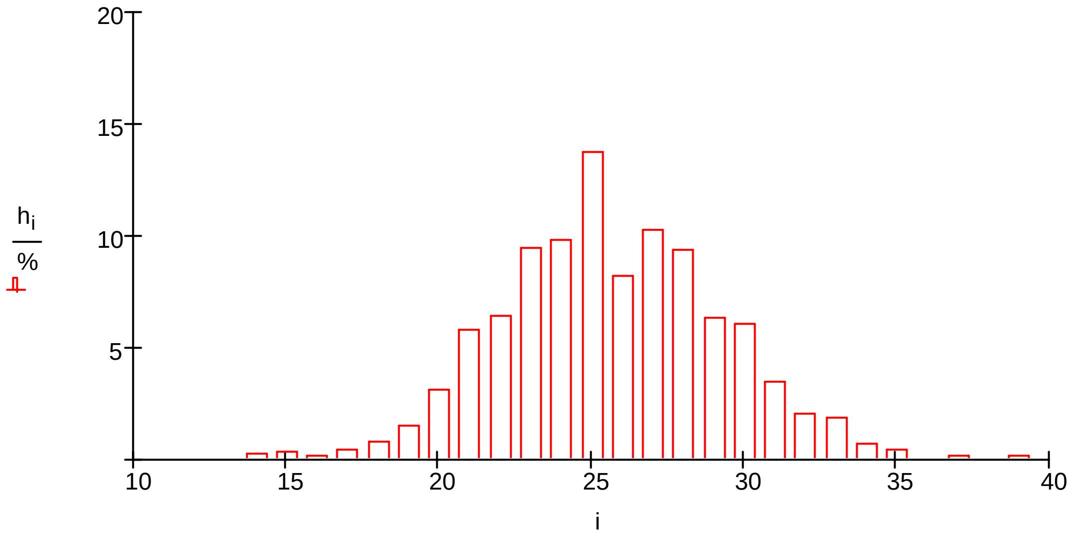


Hinweis:
In der Grafik sind der Erwartungswert μ=25.5 und der Bereich [μ-σ , μ+σ] eingezeichnet! (Berechnung dieser Größen siehe in der Region "Wahrscheinlichkeitsmodell"

Grafik: Die ersten 100 Versuche

Run-Häufigkeiten

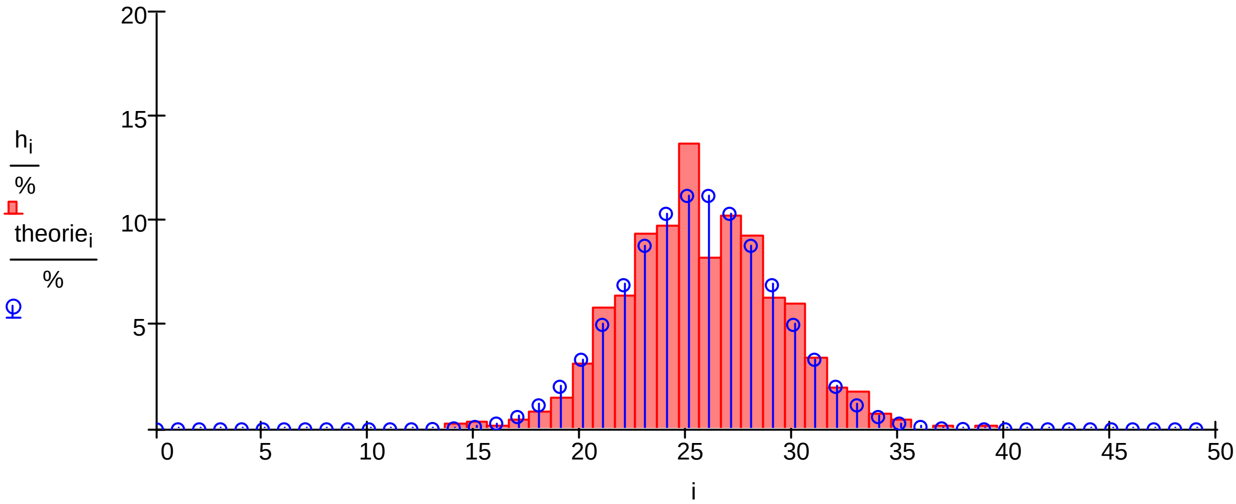
Das sind die relativen Häufigkeiten (in %) der Runs in den $N = 1000$ Versuchen



Run-Häufigkeiten

Run-Häufigkeiten - Vergleich

Hier werden die relativen Häufigkeiten mit den theoretischen Werten verglichen. Dieses Modell ist eine Binomialverteilung (siehe unten bei der Region "Wahrscheinlichkeitsmodell")



Run-Häufigkeiten - Vergleich

Statistiken

Die folgenden Auswertungen beziehen sich auf die Anzahl von Runs bei 50 Münzwürfen:

Die Mittlere Anzahl der Runs beträgt $\text{mittelwert}(\text{runs}) = 25.63$ $N = 1000$

Die Standardabweichung der Runs beträgt $\text{Stdev}(\text{runs}) = 3.561$

Zum Vergleich: **Die theoretischen Werte sind μ = 25,5 und σ = 3,5.**

Einige beobachtete Werte: Runlängen $\geq$ 5 $\text{fünferserien} = 1468$

Runlängen $\geq$ 10 $\text{zehnerserien} = 34$

beobachtete Maximallänge $\text{maximallänge} = 13$

Anzahl der Runs im Bereich [22 - 29] $\dfrac{\text{anzBereich}(22, 29)}{N} = 73.3\%$

Anmerkungen:
Wie unten in der Region "Wahrscheinlichkeitsmodell" gezeigt wird, folgt die Anzahl der Runs einer Binomialverteilung. Daraus ergeben sich folgende theoretischen Werte für Erwartungswert μ und Standardabweichung σ:

$\mu = 25.5$ $\sigma = 3.5$

Der Vergleich mit den Auswertungen der Simulation zeigt eine entsprechende Übereinstimmung, die aber natürlich von der Anzahl der durchgeführten Simulationen abhängigt ist.

Ferner werden die Anzahl der beobachteten "Fünfersreien" (=Runlänge mindestens 5) sowie "Zehnerserien" (=Runlänge mimndestens 10) und der jeweils längste Run ausgegeben. Dies deswegen, weil man bei "erdachten 0-1-Zufallsfolgen" beobachtne kann, dass diese Serien bzw. Längen meistens unterschätzt werden.
Das ist ja auch der Grund, weshalb bei erdachten 0-1-Zufallszahlenfolgen im Schnitt mehr Runs geschätzt werden als sie dem "wirklichen" Zufall entsprechen!

Statistiken

Wahrscheinlichkeitsmodell

Im Folgenden werden noch einige interessante Größen und Wahrscheinlichkeiten berechnet, die sich aus der Theorie (Binomialverteilung) ergeben. Dabei ist zu berücksichtigen, dass **die Anzahl der (Runs-1) einer Binomialverteilung folgen mit den Parametern n=49 und p=0.5.** Das (-1) gegenüber n=50 ist letztlich darauf zurückzuführen, dass ein Run immer stattfindet.

Begründung : Wir betrachten am einfachsten statt der Runs die **Anzahl der Wechsel**, die das Ende eines Runs bedeuten. Die Anzahl der Wechsel muss gleich der Anzahl der Runs minus 1 sein, weil anfangs jedenfalls irgendwie gestartet wird. Daher gibt es n=49 verschiedene mögliche Positionen für einen "Wechsel".
Für jede Position gilt beim Münzwurf die "Wechselwahrscheinlichkeit" p=0.5.
Wenn wir nun nach der Wahrscheinlichkeit fragen, mit der **x** Wechsel auftreten, entspricht dies dem **Modell der Binomialverteilung mit n=49 und p=0,5** und wir können sagen:
p(Wechselzahl = x) = g_{bi}(x ; n=49 ; p=0,5) = p(Runzahl = x+1)

Da bei einer Binomialverteilung mit den Parametern n und p der Erwartungswert mit $\mu = n \cdot p$ und die Standardabweichung mit $\sigma = n \cdot p \cdot (1 - p)$ berechnet werden, ergeben sich für die Anzahl der Runs folgende theoretischen Werte:

$\mu = 49 \cdot 0.5 + 1 = 25.5$ und $\sigma = \sqrt{49 \cdot 0.5 \cdot (1 - 0.5)} = 3.5$

Will man nun beispielsweise die Wahrscheinlichkeit berechnen , dass die Anzahl der Runs > 30 ist, wird so vorgegangen:
p (Runs>30) = p(Runs-1>29) = 1 - p (Runs-1 <= 29) = 1-G_{bi} (29; n=49 ; p=0.5)

Daraus ergeben sich die folgenden interessanten Wahrscheinlichkeiten, die auch mit den Ergebnissen der Simulation verglichen werden sollten (siehe oben unter dem Punkt "Statistiken" - andere Bereichsergebnisse können mit der weiter oben definierten Funktion anzBereich(a,b) ermittelt werden.

Wie groß ist die Wahrscheinlichkeit, dass die Anzahl der Runs größer als 25 ist?

$1 - G_{bi}(24, 49, 0.5) = 50\%$

Wie groß ist die Wahrscheinlichkeit, dass die Anzahl der Runs 30 oder mehr beträgt?

$1 - G_{bi}(28, 49, 0.5) = 12.6\%$

Wie groß ist die Wahrscheinlichkeit, dass die Anzahl der Runs 32 oder mehr beträgt?

$1 - G_{bi}(30, 49, 0.5) = 4.3\%$

Wie groß ist die Wahrscheinlichkeit, dass die Anzahl der Runs 35 oder mehr beträgt?

$1 - G_{bi}(33, 49, 0.5) = 0.47\%$

Wie groß ist die Wahrscheinlichkeit, dass die Anzahl der Runs 21 oder weniger beträgt?

$G_{bi}(20, 49, 0.5) = 12.6\%$

Letzteres Beispiel zeigt, dass die Wahrscheinlichkeiten symmetrisch um dem Erwartungswert 25,5 sind.

Wie groß ist die Wahrscheinlichkeit für den Bereich 22 - 29 ? $G_{bi}(28, 49, 0.5) - G_{bi}(20, 49, 0.5) = 74.7\%$

Wahrscheinlichkeitsmodell

Feinauswertung (aktivieren!)

Hier wird ein Versuch hinsichtlich aller Runlängen ausgewertet -
der unten ermittelte Vektor "länge" enthält die Anzahl der Einser-Runs, Zweier-Runs usw.
Da die Funktion recht langsam arbeitet, muss sie bei Bedarf extra aktiviert werden, damit unten die entsprechende grafische Auswertung möglich ist.

```
runlength(j) := | for  k ∈ 0 .. 50
                |   länge_k ← 0
                | k ← 1
                | for  i ∈ 0 .. n − 2
                |   | k ← k + 1  if  r_{i,j} = r_{i+1,j}
                |   | otherwise
                |   |   | länge_k ← länge_k + 1
                |   |   | k ← 1
                | länge_k ← länge_k + 1  if  r_{n−1,j} = r_{n−2,j}
                | länge_1 ← länge_1 + 1  if  r_{n−1,j} ≠ r_{n−2,j}
                | return  länge
```

Bei Bedarf aktivieren !

$k := 0 .. 50$

$werte_k := \sum_j (runlength(j))_k$

Hier werden dei Runlängen aller Versuche passend addiert und im Feld "werte" gespeichert.

werte =

	0
0	0
1	13155
2	6400
3	3048
4	1559
5	761
6	364
7	192
8	82

9	35
10	18
11	12
12	3
13	1
14	0
15	0

$$alle := \sum_{i=1}^{n} werte_i \qquad h := \frac{werte}{alle}$$

Hier werden die relativen Häufigkeiten h_i für das Auftreten der einzelnen Runlängen angegeben.
Gemäß dem Gesetz der großen Zahlen entspricht dies annähernd der Wahrscheinlichkeit, dass der nächste Run von eben dieser Länge ist.

Unten erfolgt eine grafische Darstellung dieser Werte.

h =

	0
0	0
1	0.513
2	0.25
3	0.119
4	0.061
5	0.03
6	0.014
7	$7.491\cdot10^{-3}$
8	$3.199\cdot10^{-3}$
9	$1.366\cdot10^{-3}$
10	$7.023\cdot10^{-4}$
11	$4.682\cdot10^{-4}$
12	$1.171\cdot10^{-4}$
13	$3.902\cdot10^{-5}$
14	0
15	0

Wenn man die relativen Häufigkeiten der Anzahl der Runs betrachtet, beobachtet man nicht zufällig ab h_1 eine stark abfallende Tendenz , die schon auf den ersten Blick gut durch eine Exponentialfunktion zu beschreiben sein dürfte.
Die Anzahl der Runs kann nämlich auch als eine "Wartezeitenverteilung" oder auch als "Lebensdauerexperiment" gedeutet werden - die Runlängen entsprechen dann "Wartezeiten" bzw. "Lebensdauern".
Bei gleichbleibenden Wahrscheinlichkeiten für jedes einzelne Element bezüglich "Ausfall" (so wie es hier gegeben ist), kann man nachweisen, dass derartige "Wartezeiten" oder "Lebensdauern" mit Hilfe der Exponentialverteilung beschrieben werden können (allerdings ist hier natürlich zu berücksichtigen, dass nur diskrete Werte vorliegen!)
Im folgenden soll eine passende Exponentialverteilung auf 2 Arten gefunden werden.

1) Methode der kleinsten Quadrate:

$i := 0..30 \qquad t_i := i$

$$sumq(h0, T_{dach}) := \sum_{i=1}^{30} \left(h_i - h0 \cdot e^{-\frac{t_i}{T_{dach}}} \right)^2$$

Zielgröße , die ien Minimum werden soll.

$h0 := 1 \qquad T_{dach} := 2$

Schätzwerte für numerischen Lösungsblock

Vorgabe

$sumq(h0, T_{dach}) = 0$

$$\begin{pmatrix} h0 \\ T_{dach} \end{pmatrix} := Suchen(h0, T_{dach}) \qquad \begin{pmatrix} h0 \\ T_{dach} \end{pmatrix} = \begin{pmatrix} 1.056 \\ 1.386 \end{pmatrix}$$

Die Exponentialverteilung hat für die Überlebenswahrscheinlichkeit folgende Formel:

$$R(t) = e^{\frac{-t}{T}}$$

Das heisst: Obige Größe h0 ist eigentlich gleich 1 zu setzen. Daher bleibt als Parameter nur mehr T_{dach} über, welcher die sogenannte "charakteristische Lebensdauer" repräsentiert.

Diese ist aber im Fall der Exponentialverteilung gleich der durchschnittlichen Lebensdauer - daher wird nun so vorgegangen:

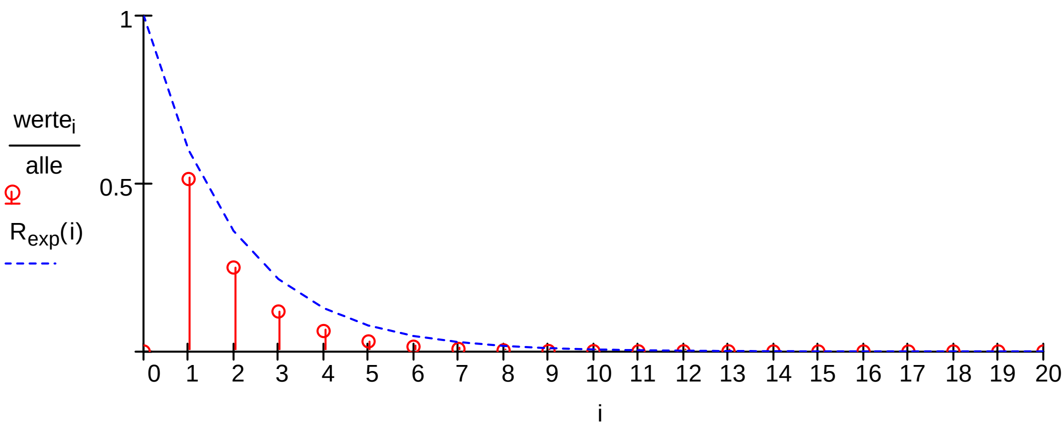
2) Ermittlung der charakteristischen (= durchschnittlichen) Lebensdauer:

$$T_{dach} := \frac{\sum_{i=1}^{n} (i \cdot werte_i)}{alle} \qquad T_{dach} = 1.951$$

$$R_{exp}(t) := e^{\frac{-(t)}{T_{dach}}}$$

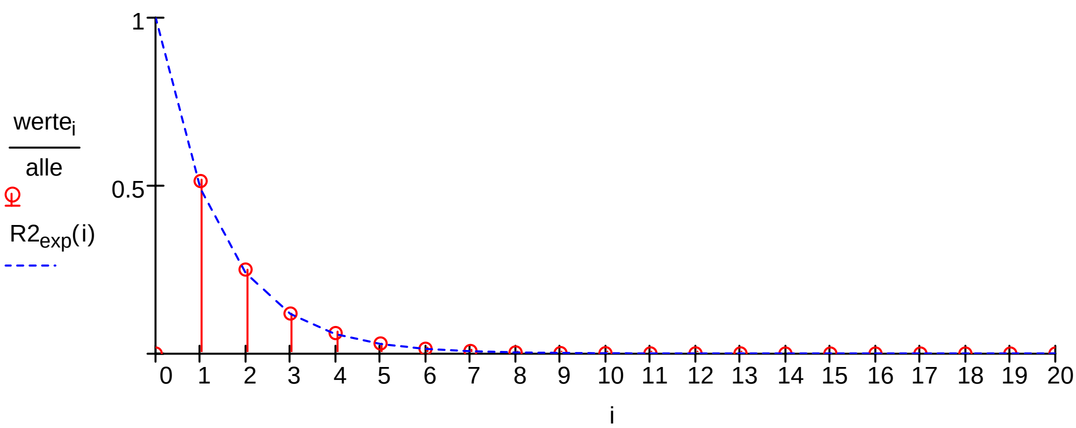
Dieser Wert weicht relativ deutlich von dem oben ermittelten Wert (T_{dach} ungefähr gleich 1,4) ab.

Der Grund dafür ist darin zu sehen, dass das Modell der Exponentuialverteilung hier nicht ganz passen kann, weil ein Run der Länge=0 nicht auftreten kann und daher ein Run der Länge 1 einer "Mindestlebensdauer" entspricht, weswegen das Modell der Exponentialverteilung nicht ganz passen kann - wie man auch in der folgenden Grafik erkennt:



Setzt man hingegen für T_{dach} den Wert ein, der (ungefähr) mit Hilfe der Methode der kleinsten Quadrate ermittelt wurde, so passt die "Exponentialverteilung" optisch deutlich besser.

$$T_{dach} := 1.4 \qquad R2_{exp}(t) := e^{\frac{-(t)}{T_{dach}}}$$



Feinauswertung (aktivieren!)