Über Tests von Zufallszahlengeneratoren

Bernd Paul Jäger
Inst. f. Biometrie u. Med. Informatik
Ernst-Moritz-Arndt-Universität

Walther-Rathenau-Str. 48
17487 Greifswald
bjaeger@biometrie.uni-greifswald.de

Torsten Philipp
Fachhochschule Stralsund,
Fachbereich Elektrotechnik und Informatik
Zur Schwedenschanze 15
18435 Stralsund
torsten-philipp@web.de

Paul Eberhard Rudolph
Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere (FBN)
Wilhelm-Stahl-Allee 2
18196 Dummerstorf
pe.rudolph@fbn-dummerstorf.de

Karl-Ernst Biebler
Inst. f. Biometrie u. Med. Informatik
Ernst-Moritz-Arndt-Universität
Walther-Rathenau-Str. 48
17487 Greifswald
biebler@biometrie.uni-greifswald.de

Zusammenfassung

Zahlreiche statistische Tests, darunter auch die Familie der Run-Tests (Abb.1.1-1.3), wurden ersonnen, um die ordnungsgemäße Funktion von Zufallszahlengeneratoren zu überprüfen, die (pseudo)gleichverteilte Zufallszahlen erzeugen. Ein einzelner Test prüft aber nur bestimmte Konstellationen ab, etwa die, dass eine Sequenz von Zufallszahlen nicht zu lange monoton wächst, dass die empirische Verteilung der Gleichverteilung entspricht oder dass Permutationen genau so häufig gefunden werden wie man erwartet. Deshalb ist es heute Standard, eine ganze Batterie von Tests [2] nacheinander anzuwenden, um Besonderheiten in vielfältiger Hinsicht zu erkennen.

Der so genannte Run-Test nach Knuth [1] untersucht Längen von streng monotonen Sequenzen innerhalb einer Folge von gleichverteilten Zufallszahlen aus dem Intervall [0, 1). Nachfolgend wird o.B.d.A. nur der Fall streng monoton fallender Sequenzen behandelt.

Die Verteilungsfunktion der zufälligen Längen L der Sequenzen ist bekannt:

$P(L = k) = k / (k + 1)!$

L kann einerseits sämtliche natürliche Zahlen k als Werte annehmen, andererseits werden die Wahrscheinlichkeiten für größer werdende k rasch klein. Es gelten für den Erwartungswert $E(L) = e - 1 \approx 1.71828$ und für die Varianz $V(L) = 3e - e^2 \approx 0.76579$. Man wird folglich keine sehr langen Sequenzen beobachten.

Der Run-Test wird fälschlicherweise bereits in [1] und auch andernorts auf Zufallszahlengeneratoren für gleichverteilte ganze Zahlen angewandt. Für den Zufallszahlengenerator in SAS führt eine solche unkorrekte Bewertung zu einem Negativurteil.

In dieser Arbeit werden für gleichverteilte ganze Zahlen die exakte Verteilung der Längen L streng monoton wachsender Sequenzen angegeben, in einem SAS Makro berechnet sowie ihre Grenzverteilung für $n \rightarrow \infty$ ermittelt. Zur Anwendung wird diese asymptotische Verteilung nicht empfohlen. Der Zufallszahlengenerator in SAS besteht den unter Verwendung der exakten Verteilung von L durchgeführten Run-Test und ist somit nicht zu beanstanden.

Schlüsselwörter: Zufallszahlengenerator, Gleichverteilung, Run-Test

1 Einleitung: Run-Tests

Der Begriff Run-Test zur Überprüfung, ob ein Zufallszahlengenerator gleichverteilte Zufallszahlen erzeugt, ist in der Literatur nicht einheitlich. Es gibt mindestens drei Tests solchen Namens, die sowohl als exakte Tests als auch in der asymptotischen Form besprochen werden sollen.

1.1 Run-Test nach Knuth

Beim Run-Test nach Knuth [1] werden zufällige Längen L streng monotoner Sequenzen untersucht.

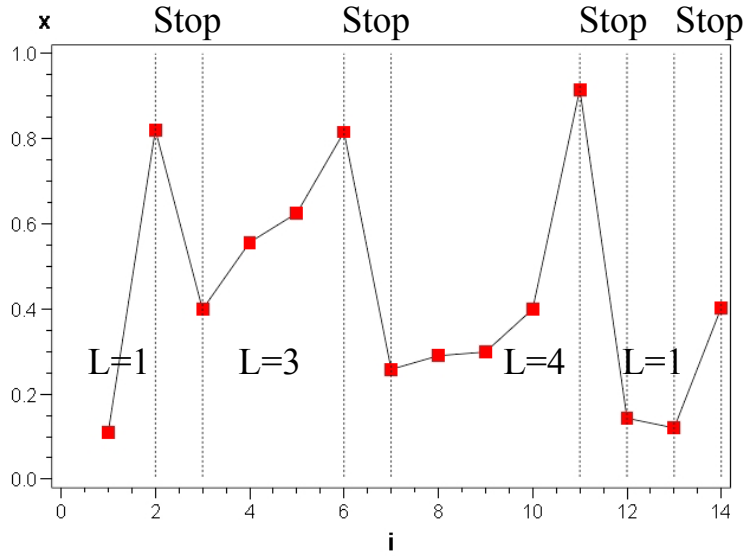


Abbildung 1.1: Illustration zum Run-Test nach Knuth, Zufallsgröße ist die Länge der monotonen Sequenzen, die durch eine Stoppzahl getrennt werden, bei der das Monotonieverhalten wechselt

Die Zufallsgröße L ist die Länge der monoton wachsenden oder fallenden Teilsequenzen, Runs genannt, die durch eine „Stoppzahl“ beendet werden. Bei der Stoppzahl wird das Monotonieverhalten aufeinander folgender Zufallszahlen erstmals unterbrochen. In der Abbildung ändert sich die Monotonie nach der zweiten, sechsten, elften und 13. Zufallszahl. Es entstehen die zufälligen Längen der Runs von 1, 3, 4, 1,….. Die Verteilung dieser zufälligen Längen kann durch kombinatorische Überlegungen hergeleitet werden. In einem genügend umfangreichen Simulationsexperiment wird die empirische Verteilung der Längen der Runs bestimmt. Ein anschließender χ^2-Test prüft, ob die in einer Folge von generierten Zufallszahlen beobachteten Anzahlen an monotonen Sequenzen der jeweiligen Länge mit den erwarteten Anzahlen übereinstimmen.

1.2 Run-Test 2

Ein weiterer Run-Test (im Folgenden Run-Test 2 genannt) untersucht von Sequenzen von Zufallszahlen (aufeinander folgende Zufallszahlen) vorgegebener Länge n die Anzahl von monoton wachsenden (oder fallenden) Teilfolgen, Runs genannt. Im Unterschied zum Run-Test nach Knuth treten keine „Stoppzahlen“ auf.
Vorgegeben wird eine Sequenzlänge n. Die Sequenz setzt sich ihrerseits aus Teilsequenzen zusammen, die streng monoton sind. Die folgende Abbildung gibt die Situation für eine Sequenz der Länge n =12 an. Sie zerfällt in sechs Teilsequenzen (Runs) mit den Längen 1, 1, 3, 2, 3 und 1. Zufallsgröße ist die Anzahl an Runs in der untersuchten Sequenz. Auch hier gilt, dass kurze Runs häufig vorkommen und lange Runs gegen die Gleichverteilung sprechen.

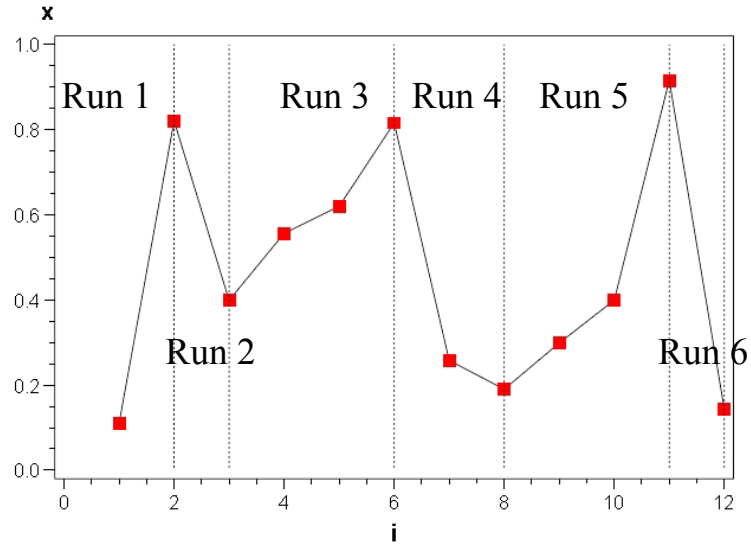


Abbildung 1.2: Illustration zum Run-Test 2, Zufallsgröße ist die Anzahl von Runs ($n_r = 6$) bei gegebener Sequenzlänge (n = 12)

Jede Sequenz kann in eine Permutation der Länge n umgewandelt werden, wenn man die Zahlen durch ihren Rang ersetzt. Die Sequenz in Abb. 1.2 beispielsweise

0.10978, 0.82053, 0.39895, 0.55639, 0.62032, 0.81566,
0.25788, 0.19015, 0.29876, 0.39940, 0.91591, 0.14322

geht durch diese Transformation über in die Permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 1 & 11 & 6 & 8 & 9 & 10 & 4 & 3 & 5 & 7 & 12 & 2 \end{pmatrix},$$

bei der man die gleichen Subsequenzen (Runs) in gleichen Längen wie in der Ausgangssequenz feststellen kann. Die Herleitung der Verteilungsfunktion für die Anzahlen N_r der Runs bei Sequenzen der Länge n kann über die Anzahlen der Runs der zugehörigen Permutation erfolgen.

1.3 Bedingter Run-Test

Ein bedingter Run-Test unter Rückgriff auf den Wald-Wolfowitz-Test [6] untersucht Zufallszahlenfolgen vorgegebener Länge n von aufeinander folgenden Zufallszahlen aus dem Intervall [0, 1). Bezüglich des Mittelwertes zerfällt die Menge von n Zufallszahlen in zwei Teilmengen. Es liegen davon n_1 Zufallszahlen oberhalb und $n_2 = n - n_1$ unterhalb des arithmetischen Mittels. Die Wahrscheinlichkeit ist Null, dass zwei Zufallszahlen gleich sind oder dass eine der n Zufallszahlen mit dem Mittelwert zusammenfällt. In diesen Teilmengen vom zufälligen Umfang n_1 und n_2 werden zusammenhängende Sequenzen, die Runs, ausgezählt. Diese müssen nicht notwendigerweise monoton wachsend oder fallend sein. Die Zufallsgröße ist die Anzahl K der Runs, wobei K als Summe der Anzahl der Runs oberhalb des Mittels K_1 und der Anzahl der Runs unterhalb des Mittels K_2 aufgefasst werden kann: $K = K_1 + K_2$. Die bedingte Verteilung der Anzahl der Runs, unter der Voraussetzung, dass n_1 Zufallszahlen oberhalb und $n_2 = n - n_1$ unterhalb des arithmetischen Mittels liegen, ist zu bestimmen.
Diese Verteilung wird zurückgeführt auf die Verteilung der Iterationszahlen, die Wald und Wolfowitz [6] ausführlich untersucht haben. Dazu ist es notwendig, die Folge der n Zufallszahlen in eine Folge der Werte von 0 und 1 umzuwandeln, wobei man für Zahlen oberhalb des arithmetischen Mittels eine 1 und für solche unterhalb des arithmetischen Mittels eine 0 schreibt (oder auch umgekehrt). Eine vollständige und ausführliche Herleitung der Wahrscheinlichkeitsverteilung der Iterationszahlen findet man bei Fisz [7].

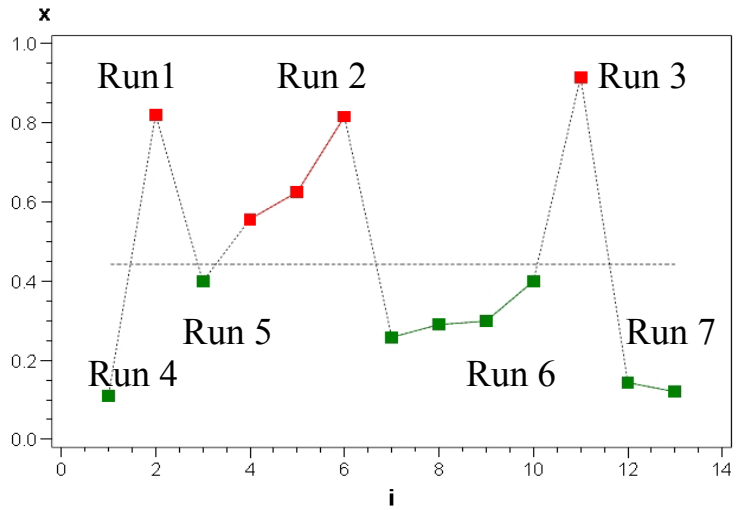


Abbildung 1.3: Illustration zum bedingten Run-Test nach Wald und Wolfowitz (Sequenzlänge n = 13, Anzahl Runs $k = k_1 + k_2 = 3 + 4 = 7$)

2 Der Run Test nach KNUTH

Dieser Run-Test untersucht die Länge von streng monotonen Sequenzen innerhalb einer Folge von gleich verteilten Zufallszahlen, die o.B.d.A. aus dem Intervall [0, 1) stammen. Die Sequenzen können streng monoton wachsend oder auch streng monoton fallend sein. Zwei gleiche aufeinander folgende Zufallszahlen kann es auf Grund der stetigen Verteilung nicht geben.

Es werden o.B.d.A. nur die streng monoton wachsenden Sequenzen betrachtet. Die Verteilungen der Längen der streng monoton fallenden und der streng monoton wachsenden Sequenzen sind identisch. Eine Sequenz der Länge L = l beginnt mit x_i an der Stelle i und endet an der Stelle x_{i+l}. Die darauf folgende Zufallszahl x_{i+l+1} ist die „Stoppzahl“. Für diese Zufallszahlen gelten die folgenden Ungleichungen

$$x_i < x_{i+1} < \ldots < x_{i+l} > x_{i+l+1}.$$

Die folgende Sequenz beginnt nach der Stoppzahl. Als mögliche Sequenzlängen kommen alle natürlichen Zahlen in Frage. Für die Zufallsgröße L gilt:

$$P(L = l) = \frac{l}{(l+1)!}$$

Die Beweisidee wird kurz skizziert. Wegen $x_i < x_{i+1} < \ldots < x_{i+l}$ in einem Run ist die Folge der Ränge der Zufallszahlen des Runs 1 < 2 <…< l. Wegen $x_{i+l} > x_{i+l+1}$ ist die Rangzahl der Stoppzahl entweder vor der von x_i oder der von x_{i+l} oder …. oder der von x_{i+l} einzuordnen. Das sind l mögliche Anordnungen (Permutationen) von insgesamt (l +1)!

Für die Zufallsgröße L gelten weiterhin

$$E(L) = \sum_{i=1}^{\infty} i \cdot P(L = i) = \sum_{i=1}^{\infty} \frac{i^2}{(i+1)!} = e - 1 \approx 1.7182818 \quad \text{und}$$

$$V(L) = \sum_{i=1}^{\infty} (i - E(L))^2 \cdot P(L = i) = \sum_{i=1}^{\infty} \frac{i \cdot (i - (e-1))^2}{(i+1)!} = 3e - e^2 \approx 0.7657894 .$$

Man erwartet bei gleichverteilten Zufallszahlen nur kleine Sequenzlängen. Werden die Sequenzlängen zu groß oder treten wesentlich andere relative Häufigkeiten auf, als sie durch die Verteilungsfunktion vorgegeben sind, so wird die Gleichverteilungshypothese abgelehnt. Statistischer Test: Durchgeführt wird ein χ^2-Anpassungstest, der bei einer großen Stichprobe die Häufigkeiten der Sequenzlängen mit der erwarteten Anzahl der Sequenzlängen nach deren Verteilungsgesetz vergleicht. Schwach besetzte Sequenzlängenkategorien werden zusammengefasst.

Dazu ist die folgende Aussage hilfreich, die mit vollständiger Induktion über n leicht bewiesen werden kann:

$$\sum_{i=1}^{n} P(L = i) = \sum_{i=1}^{n} \frac{i}{(i+1)!} = 1 - \frac{1}{(n+1)!} .$$

Daraus lässt sich leicht die für den χ^2-Anpassungstest benötigte Reihe ermitteln:

$$\sum_{i=n}^{\infty} P(L = i) = 1 - \sum_{i=1}^{n-1} \frac{i}{(i+1)!} = 1 - \left(1 - \frac{1}{n!}\right) = \frac{1}{n!} .$$

Beispiel 2.1:

Mit dem SAS-Zufallszahlengenerator UNIFORM wurden n = 1 000 000 gleichverteilte Zufallszahlen aus dem Intervall [0, 1) erzeugt. Die ermittelte Anzahl von Sequenzen ist

nur noch 367 702, weil zum einen die Stoppzahlen abgezogen werden, zum anderen mehrere Zufallszahlen zu einer streng monoton wachsenden Teilsequenz gehören. Die Ergebnisse sind in der folgenden Tabelle zusammengefasst. Bei einem Freiheitsgrad von f = 8 wird der kritische Wert 14.0671 durch die Prüfgröße $\chi^2 = 13.0003$ nicht erreicht.

Tabelle 2.1: Beobachtete und erwartete Häufigkeiten von Run-Längen streng monoton wachsender Sequenzen aus einer Folge von 1 000 000 gleichverteilter Zufallszahlen aus dem Intervall [0; 1)

Länge L der Sequenz	P(L = l)	Beobachtete Häufigketen B_i	Erwartete Häufigkeiten $E_i = n \cdot P(L = l)$	$(B_i - E_i)^2 / E_i$
1	1 / 2 = 0.50000	183 443	183 851.00	0.90543
2	2 / 6 = 0.33333	122 676	122 567.33	0.09634
3	3 / 24 = 0.12500	46 493	45 962.75	6.11724
4	4 / 120 = 0.03333	12 038	12 256.73	3.90351
5	5 / 720 = 0.00694	2 530	2 553.49	0.21602
6	6 / 5040 = 0.00119	444	437.74	0.08951
7	7 / 40320 = 0.00017	65	63.84	0.02118
≥8	1 / 40320 = 0.00002	13	9.12	1.65112
	Summe	367 702	367702.00	13.0003

Wesentlich schwieriger ist dieser Run-Test für gleichverteilte ganze Zahlen von 1 bis k. Die Zufallsgröße L wird wie im stetigen Fall bestimmt. Man beachte, dass die Abbruchbedingung bereits eintritt, wenn die folgende Zahl mit der vorangehenden übereinstimmt. Im Gegensatz zum obigen Run-Test, bei dem die Wahrscheinlichkeit Null ist, dass zwei aufeinander folgende Zufallszahlen in einer Sequenz gleich sind, hat dieses Ereignis bei gleichverteilten ganzen Zahlen von 1 bis k die Wahrscheinlichkeit $1/k^2$.
Die Zufallsgröße L kann im Gegensatz zum stetigen Fall auch nur endlich viele Werte annehmen, und zwar von 1 bis k. Spätestens nach k + 1 Schritten muss sich eine der Zufallszahlen wiederholen.
Die Herleitung der Verteilungsfunktion wird für k = 6 dargestellt. Dem entspricht als Zufallszahlengenerator der gewöhnlichen Spielwürfel.
Im ersten Schritt werden alle Sequenzen der Länge L = 1 angegeben. Die Sequenz und die Stoppzahl bilden ein Paar von Zufallszahlen, insgesamt also $36 = k^2$ Paare. In der Tabelle 2.2 sind alle Sequenzen der Länge 1 und ihre Stoppzahlen angegeben.
Da es genau $21 = 1 + 2 + \ldots + 6$ solcher Sequenzen gibt, ist die Wahrscheinlichkeit

$$P(L_6 = 1) = \frac{21}{6^2} \approx 0.58333 .$$

Aus Tabelle 2.3 ist eine Formel für gleichverteilte Zufallszahlen von 1 bis n ableitbar,

$$P(L_n = 1) = \frac{\frac{n}{2}(n+1)}{n^2} = \frac{1}{2}\left(1 + \frac{1}{n}\right),$$

so dass sofort folgt

$$\lim_{n\to\infty} P(L_n = 1) = \frac{1}{2} = P(L = 1).$$

Tabelle 2.2: Anzahlbestimmung streng monoton wachsenden Sequenzen der Länge 1

monotone Sequenz	Stoppzahl	Anzahl Paare
1	1	1
2	1, 2	2
3	1, 2, 3	3
4	1, 2, 3, 4	4
5	1, 2, 3, 4, 5	5
6	1, 2, 3, 4, 5, 6	6
	Summe	21

Tabelle 2.3: Anzahlbestimmung streng monoton wachsenden Sequenzen der Länge 2

monotone Sequenz	Stoppzahl	Anzahl Tripel
1, 2	1, 2	2
1, 3	1, 2, 3	3
1, 4	1, 2, 3, 4	4
1, 5	1, 2, 3, 4, 5	5
1, 6	1, 2, 3, 4, 5, 6	6
2, 3	1, 2, 3	3
2, 4	1, 2, 3, 4	4
2, 5	1, 2, 3, 4, 5	5
2, 6	1, 2, 3, 4, 5, 6	6
3, 4	1, 2, 3, 4	4
3, 5	1, 2, 3, 4, 5	5
3, 6	1, 2, 3, 4, 5, 6	6
4, 5	1, 2, 3, 4, 5	5
4, 6	1, 2, 3, 4, 5, 6	6
5, 6	1, 2, 3, 4, 5, 6	6
	Summe	70

Die Tabelle 2.3 enthält sämtliche streng monoton wachsenden Sequenzen der Länge 2 mit den möglichen Stoppzahlen für jede Sequenz. Aufgeführt sind 70 Tripel, bestehend aus der Sequenz der Länge 2 und der Stoppzahl. Insgesamt gibt es 6^3 mögliche Tripel, so dass

$$P(L_6 = 2) = \frac{70}{6^3} \approx 0.32407.$$

Tabelle 2.4: Bestimmung der Anzahl aller streng monoton wachsenden Sequenzen der Länge 3

monotone Sequenz	Stoppzahl	Anzahl Quadrupel
1, 2, 3	1, 2, 3	3
1, 2, 4	1, 2, 3, 4	4
1, 2, 5	1, 2, 3, 4, 5	5
1, 2, 6	1, 2, 3, 4, 5, 6	6
1, 3, 4	1, 2, 3, 4	4
1, 3, 5	1, 2, 3, 4, 5	5
1, 3, 6	1, 2, 3, 4, 5, 6	6
1, 4, 5	1, 2, 3, 4, 5	5
1, 4, 6	1, 2, 3, 4, 5, 6	6
1, 5, 6	1, 2, 3, 4, 5, 6	6
2, 3, 4	1, 2, 3, 4	4
2, 3, 5	1, 2, 3, 4, 5	5
2, 3, 6	1, 2, 3, 4, 5, 6	6
2, 4, 5	1, 2, 3, 4, 5	5
2, 4, 6	1, 2, 3, 4, 5, 6	6
2, 5, 6	1, 2, 3, 4, 5, 6	6
3, 4, 5	1, 2, 3, 4, 5	5
3, 4, 6	1, 2, 3, 4, 5, 6	6
3, 5, 6	1, 2, 3, 4, 5, 6	6
4, 5, 6	1, 2, 3, 4, 5, 6	6
	Summe	105

Die Berechnung der übrigen Wahrscheinlichkeiten geht aus den Tabellen 2.4 bis 2.6 hervor.

$$P(L_6 = 3) = \frac{105}{6^4} \approx 0.08102$$

$$P(L_6 = 4) = \frac{84}{6^5} \approx 0.01080$$

$$P(L_6 = 5) = \frac{35}{6^6} \approx 0.00075 \text{ und}$$

$$P(L_6 = 6) = \frac{1}{6^6} \approx 0.00002 \ .$$

Tabelle 2.5: Bestimmung der Anzahl aller streng monoton wachsenden Sequenzen der Länge 4

monotone Sequenz	Stoppzahl	Anzahl Pentupel
1, 2, 3, 4	1, 2, 3, 4	4
1, 2, 3, 5	1, 2, 3, 4, 5	5
1, 2, 3, 6	1, 2, 3, 4, 5, 6	6
1, 2, 4, 5	1, 2, 3, 4, 5	5
1, 2, 4, 6	1, 2, 3, 4, 5, 6	6
1, 2, 5, 6	1, 2, 3, 4, 5, 6	6
1, 3, 4, 5	1, 2, 3, 4, 5	5
1, 3, 4, 6	1, 2, 3, 4, 5, 6	6
1, 3, 5, 6	1, 2, 3, 4, 5, 6	6
1, 4, 5, 6	1, 2, 3, 4, 5, 6	6
2, 3, 4, 5	1, 2, 3, 4, 5	5
2, 3, 4, 6	1, 2, 3, 4, 5, 6	6
2, 3, 5, 6	1, 2, 3, 4, 5, 6	6
2, 4, 5, 6	1, 2, 3, 4, 5, 6	6
3, 4, 5, 6	1, 2, 3, 4, 5, 6	6
	Summe	84

Tabelle 2.6: Bestimmung der Anzahl aller streng monoton wachsenden Sequenzen der Länge 5

Sequenz	Stoppzahl	Anzahl 6-Tupel
1, 2, 3, 4, 5	1, 2, 3, 4, 5	5
1, 2, 3, 4, 6	1, 2, 3, 4, 5, 6	6
1, 2, 3, 5, 6	1, 2, 3, 4, 5, 6	6
1, 2, 4, 5, 6	1, 2, 3, 4, 5, 6	6
1, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	6
2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	6
	Summe	35

Die folgende Tabelle 2.7 enthält neben der Zusammenfassung der Berechnungsschritte der Wahrscheinlichkeitsverteilung für die Längen der streng monotonen Sequenzen $P(L_6 = k)$ auch die zugehörige Wahrscheinlichkeitsverteilung $P(L = k)$ für den stetigen Fall. Die Differenzen sind noch beachtlich.

Tabelle 2.7: Wahrscheinlichkeitsverteilung für die Längen *k* der streng monoton wachsenden Sequenzen $P(L_6 = k)$ und zugehörige Wahrscheinlichkeitsverteilung $P(L = k)$ des stetigen Falls

k	$P(L_6 = k)$	$P(L = k)$
1	$21 / 6^2 = 0.58333$	0.50000
2	$70 / 6^3 = 0.32407$	0.33333

3	$105 / 6^4 = 0.08102$	0.12500
4	$84 / 6^5 = 0.01080$	0.03333
5	$35 / 6^6 = 0.00075$	0.00694
6	$1 / 6^6 = 0.00002$	0.00119
Summe	1	

Mit größer werdendem Vorrat an ganzen Zufallszahlen (1, 2, ..., n) wird die Wahrscheinlichkeitsmasse zum einen auf immer mehr Portionen aufgeteilt. Zum anderen wirkt aber ein gegenläufiger Prozess. Hält man k fest, so konvergieren die Wahrscheinlichkeiten $P(L_n = k)$ der Sequenzlängen im diskreten Fall gegen die Grenzwahrscheinlichkeiten des stetigen Modells

$$\lim_{n\to\infty} P(L_n = k) = \frac{k}{(k+1)!} = P(L = k).$$

Oben wurde bereits gezeigt, dass diese Grenzwerteigenschaft für k = 1 gilt:

$$\lim_{n\to\infty} P(L_n = 1) = \frac{1}{2} = P(L = 1).$$

Der Nachweis, dass diese Grenzwerteigenschaft auch für alle anderen Sequenzlängen k gilt, ist schwierig und wird hier nicht erbracht.

Die Tabelle 2.8 illustriert die Konvergenz von $P(L_n = k)$ für n = 3, …15, 20, 30, 50 und 100 für praktisch relevante streng monotone Sequenzlängen von 1 bis 8, die restlichen seltenen Längen sind zur Kategorie „9+“ zusammengefasst worden (vorletzte Spalte).

Selbstverständlich gelten die Grenzwerteigenschaften nicht nur für die Wahrscheinlichkeiten $P(L_n = k)$ gegen $P(L = k)$, sondern auch für den Erwartungswert (siehe Tab. 2.8) und die Varianz,

$$\lim_{n\to\infty} E(L_n) = E(L) = e - 1$$

und

$$\lim_{n\to\infty} V(L_n) = V(L) = 3e - e^2.$$

Tabelle 2.8: Konvergenz für wachsendes n von $P(L_n = k)$ gegen $P(L = k)$ innerhalb der Spalten

n	$P(L_n = k)$									$E(L_n)$
	k = 1	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9+	
3	.66667	.29630	.03704	-	-	-	-	-	-	1.37039
4	.62500	.31250	.05859	.00390	-	-	-	-	-	1.44137
5	.60000	.32000	.07200	.00768	.00032	-	-	-	-	1.48832
6	.58333	.32407	.08102	.01080	.00075	.00002	-	-	-	1.52160
7	.57143	.32653	.08746	.01333	.00119	.00005	.00000	-	-	1.54635

8	.56250	.32813	.09229	.01538	.00160	.00010	.00000	.00000	-	1.56575
9	.55556	.32922	.09602	.01707	.00198	.00015	.00001	.00000	.00000	1.58121
10	.55000	.33000	.09900	.01848	.00231	.00020	.00001	.00000	.00000	1.59374
11	.54545	.33058	.10143	.01967	.00260	.00024	.00002	.00000	.00000	1.60416
12	.54167	.33102	.10344	.02068	.00287	.00029	.00002	.00000	.00000	1.61298
13	.53846	.33136	.10514	.02156	.00311	.00033	.00003	.00000	.00000	1.62058
14	.53571	.33163	.10660	.02233	.00332	.00037	.00003	.00000	.00000	1.62712
15	.53333	.33185	.10785	.02301	.00352	.00040	.00004	.00000	.00000	1.63290
20	.52500	.33250	.11222	.02544	.00424	.00055	.00006	.00001	.00000	1.65342
30	.51667	.33296	.11654	.02797	.00505	.00072	.00008	.00001	.00000	1.67486
50	.51000	.33320	.11995	.03007	.00576	.00089	.00011	.00001	.00000	1.69152
100	.50500	.33330	.12249	.03168	.00634	.00103	.00013	.00002	.00001	1.70483
∞	.50000	.33333	.12500	.03333	.00694	.00119	.00017	.00002	.00002	1.71828

3 Run-Test 2

In Abschnitt 1.2 der Einleitung wurde bereits an einem Beispiel ausgeführt, dass eine Sequenz in eine Permutation der Länge n umgewandelt werden kann, wenn man die Zahlen durch ihren Rang ersetzt.

Die Herleitung der Verteilungsfunktion für die Anzahlen N_r der Runs bei Sequenzen der Länge n kann über die Anzahlen der Runs der zugehörigen Permutation erfolgen. Dabei kann man sich N_r als Summe der auf- bzw. absteigenden Runs vorstellen, $N_r = N_1 + N_2$. Damit ist dieser Test allerdings uninteressant geworden, weil er ähnliche Konstellationen im Output des Zufallszahlengenerators wie der Permutationstest abprüft. Treten nämlich die Permutationen mit den erwarteten Häufigkeiten auf unter der Annahme, dass der Zufallszahlengenerator gleichverteilte Zufallszahlen liefert, dann natürlich auch die durch Transformationen daraus erhaltenen Anzahlen von Runs.

Die Wahrscheinlichkeitsfunktion, der Erwartungswert und die Varianz von N_r wird in Tabelle 3.1 für die Sequenzlänge n = 4 exemplarisch vorgeführt. Ein entsprechendes SAS-Programm berechnet Verteilungen für nicht zu große Sequenzlängen n. (Man beachte, dass n! mögliche Permutationen zu berücksichtigen sind.).

Interessant ist allerdings, dass Erwartungswert und Varianz der Zufallsgröße nur von der Sequenzlänge n abhängen. Es gelten

$$E(N_r) = \frac{2n-1}{3} \text{ und } V(N_r) = \frac{16n-29}{90}.$$

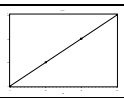
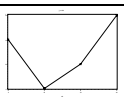
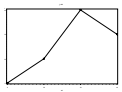
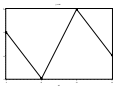
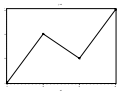
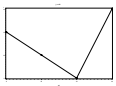
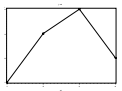
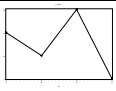
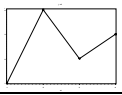
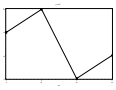
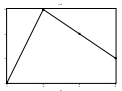
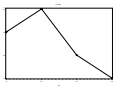
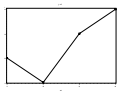
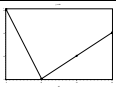
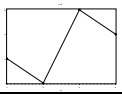
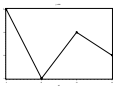
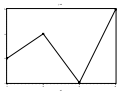
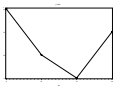
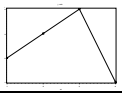
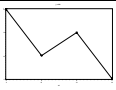
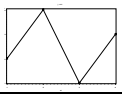
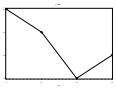
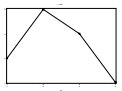
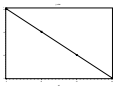
Dann ist für große Sequenzlängen n die Prüfgröße

$$U_1 = \left(N_r - \frac{2n-1}{3} \right) / \sqrt{\frac{16n-29}{90}}$$

asymptotisch N(0, 1)-verteilt. Damit hat man einen praktikablen und schnell durchführbaren Test des Zufallszahlengenerators für große Sequenzlängen, bei denen ein Permutationstest unsinnig wäre.

Für die Sequenzlänge k = 4 wird über die zugehörigen 4! = 24 möglichen Permutationen die Herleitung der Wahrscheinlichkeiten $P(N_r = i)$ gezeigt.

Tabelle 3.1: Herleitung der Verteilung der Anzahl der Runs N_r bei Sequenzlänge k = 4

Permutation		N_1	N_2	N_r	Permutation		N_1	N_2	N_r
$\begin{pmatrix}1&2&3&4\\1&2&3&4\end{pmatrix}$		1	0	1	$\begin{pmatrix}1&2&3&4\\3&1&2&4\end{pmatrix}$		1	1	2
$\begin{pmatrix}1&2&3&4\\1&2&4&3\end{pmatrix}$		1	1	2	$\begin{pmatrix}1&2&3&4\\3&1&4&2\end{pmatrix}$		1	2	3
$\begin{pmatrix}1&2&3&4\\1&3&2&4\end{pmatrix}$		2	1	3	$\begin{pmatrix}1&2&3&4\\3&2&1&4\end{pmatrix}$		1	1	2
$\begin{pmatrix}1&2&3&4\\1&3&4&2\end{pmatrix}$		1	1	2	$\begin{pmatrix}1&2&3&4\\3&2&4&1\end{pmatrix}$		1	2	3
$\begin{pmatrix}1&2&3&4\\1&4&2&3\end{pmatrix}$		2	1	3	$\begin{pmatrix}1&2&3&4\\3&4&1&2\end{pmatrix}$		2	1	3
$\begin{pmatrix}1&2&3&4\\1&4&3&2\end{pmatrix}$		1	1	2	$\begin{pmatrix}1&2&3&4\\3&4&2&1\end{pmatrix}$		1	1	2
$\begin{pmatrix}1&2&3&4\\2&1&3&4\end{pmatrix}$		1	1	2	$\begin{pmatrix}1&2&3&4\\4&1&2&3\end{pmatrix}$		1	1	2
$\begin{pmatrix}1&2&3&4\\2&1&4&3\end{pmatrix}$		1	2	3	$\begin{pmatrix}1&2&3&4\\4&1&3&2\end{pmatrix}$		1	2	3
$\begin{pmatrix}1&2&3&4\\2&3&1&4\end{pmatrix}$		2	1	3	$\begin{pmatrix}1&2&3&4\\4&2&1&3\end{pmatrix}$		1	1	2
$\begin{pmatrix}1&2&3&4\\2&3&4&1\end{pmatrix}$		1	1	2	$\begin{pmatrix}1&2&3&4\\4&2&3&1\end{pmatrix}$		1	2	3
$\begin{pmatrix}1&2&3&4\\2&4&1&3\end{pmatrix}$		2	1	3	$\begin{pmatrix}1&2&3&4\\4&3&1&2\end{pmatrix}$		1	1	2
$\begin{pmatrix}1&2&3&4\\2&4&3&1\end{pmatrix}$		1	1	2	$\begin{pmatrix}1&2&3&4\\4&3&2&1\end{pmatrix}$		0	1	1

Man erhält für die Sequenzlänge k = 4 aus Tabelle 3.1 die Wahrscheinlichkeiten

$P(N_r = 1) = 2/24 \approx 0.08333,$

$P(N_r = 2) = 12/24 = 0.50000$ und

$P(N_r = 3) = 10/24 \approx 0.41667,$

damit den Erwartungswert $E(N_r) = 7/3$ und die Varianz $V(N_r) = 35/90$

4 Bedingter Run-Test

Beim bedingten Run-Test wird eine Sequenz von n Zufallszahlen gezogen und das arithmetische Mittel m gebildet. Bezüglich des Wertes von m zerfällt die Sequenz in zwei Teilmengen mit den zufälligen Umfängen N_1 und N_2, die Zufallszahlen die oberhalb und diejenigen die unterhalb des arithmetischen Mittels m liegen. N_1 und N_2 können die Werte von 1 bis n-1 annehmen, denn es muss mindestens eine Zufallszahl ober-

halb und auch mindestens eine Zufallszahl unterhalb von m liegen (vergleiche Abb. 4.1).
Die Verteilungen von N_1 und N_2 sind identisch bei einem Zufallszahlengenerator, der gleichverteilte Zufallszahlen aus dem Intervall von 0 bis 1 erzeugt. Davon kann man sich durch Abb. 4.1 überzeugen lassen. Ausführlich findet man die Herleitung der Formeln dieses Abschnittes in Fisz [7]

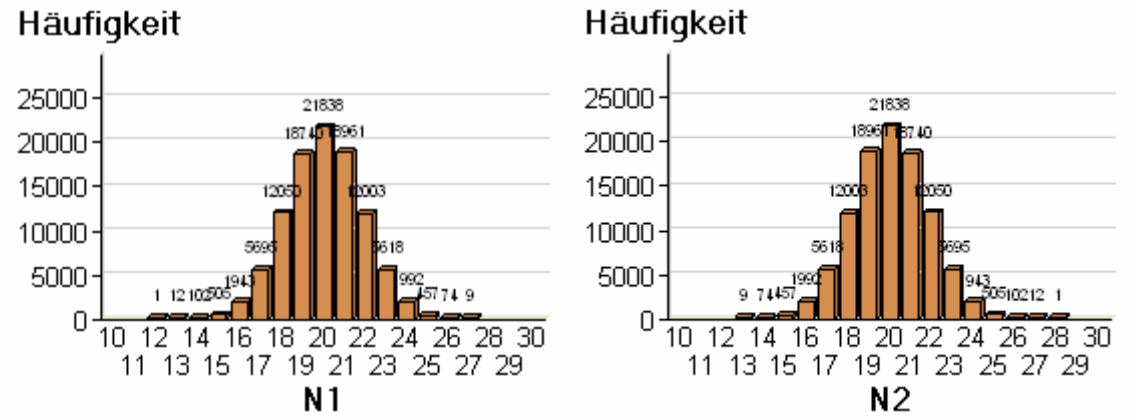


Abbildung 4.1: Empirische Verteilung von N_1 (links) und von N_2 (rechts) für Sequenzlänge n = 40

Die bedingte Verteilung der Anzahl der Runs unter eben diesen Bedingungen ist

$$P\left(K = k \mid N_1 = n_1, N_2 = n_2\right) = 2 \cdot \frac{\binom{n_1 - 1}{\frac{k}{2} - 1} \cdot \binom{n_2 - 1}{\frac{k}{2} - 1}}{\binom{n}{n_1}}$$

für k gerade und

$$P\left(K = k \mid N_1 = n_1, N_2 = n_2\right) = \frac{\binom{n_1 - 1}{\frac{k-1}{2}}\binom{n_2 - 1}{\frac{k-3}{2}} + \binom{n_1 - 1}{\frac{k-3}{2}}\binom{n_2 - 1}{\frac{k-1}{2}}}{\binom{n}{n_1}}$$

für k ungerade und der bedingte Erwartungswert und die bedingte Varianz sind

$$E\left(K \mid N_1 = n_1, N_2 = n_2\right) = \frac{2n_1n_2 + n}{n}$$

und

$$V\left(K \mid N_1 = n_1, N_2 = n_2\right) = \frac{2n_1n_2\left(2n_1n_2 - n\right)}{(n-1)n^2}.$$

Für große Sequenzlängen n ist

$$Z = \frac{K - \left(\frac{2n_1 n_2 + n}{n} \right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{(n-1)n^2}}}$$

asymptotisch standardnormalverteilt. Wald und Wolfowitz [6] haben aber bereits 1940 bewiesen, dass die obige bedingte Verteilung für $n_1 = \alpha n_2$ und $n_1 \to \infty$ gegen

$$N\left(\frac{2n_1}{1+\alpha}, \frac{4\alpha n_1}{(1+\alpha)^3} \right)$$

konvergiert. Diese Normalverteilung kann bereits für $n_1 > 20$ und $n_2 > 20$ verwandt werden.

Beispiel 4.1:
Es werden 100 000 zufällige Sequenzen der Länge n = 40 betrachtet, $(x_1, x_{2,}..., x_{40})$. Die Zufallsgrößen N_1 und N_2 können prinzipiell die Werte von 1 bis 39 annehmen, allerdings sind extreme Konstellationen selten. In Tab. 4.1 erkennt man, dass bei 10000 simulierten Sequenzen N_1 und N_2 nicht unter 12 fallen und nicht über 28 ansteigen. Einige der bedingten Häufigkeitsfunktionen zeigen die folgenden Abbildungen.

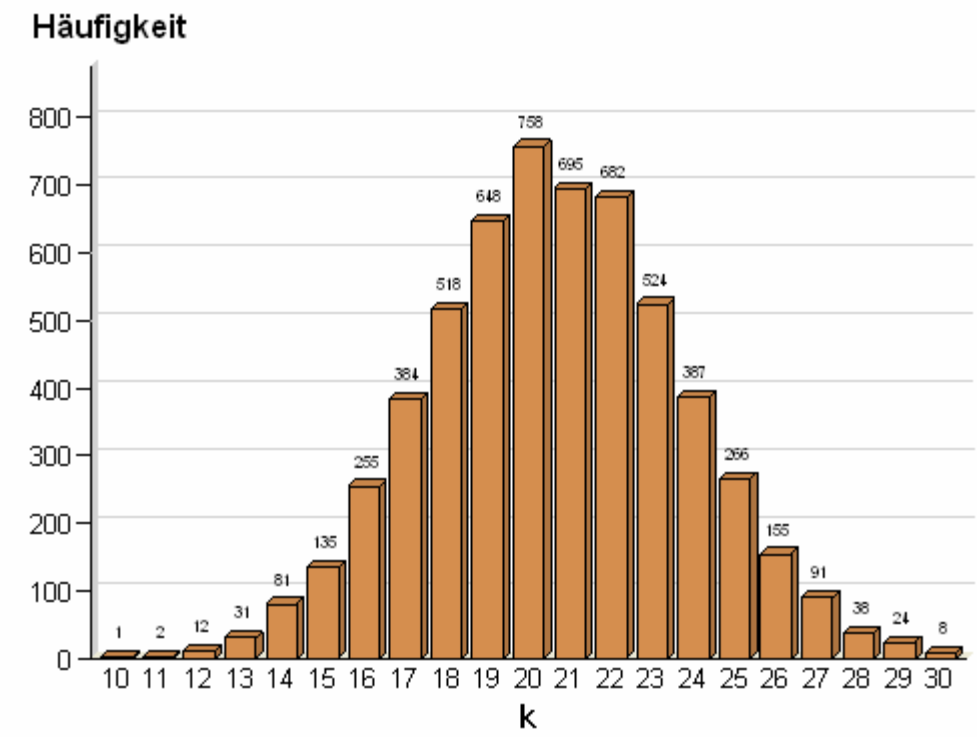


Abbildung 4.2: Bedingte Verteilung der Anzahl an Runs mit $n_1 = 17$, $n_2 = 23$, $E(K \mid N_1 = 23, N_2 = 17) = 20.55$ und $V(K \mid N_1 = 23, N_2 = 17) = 9.29$

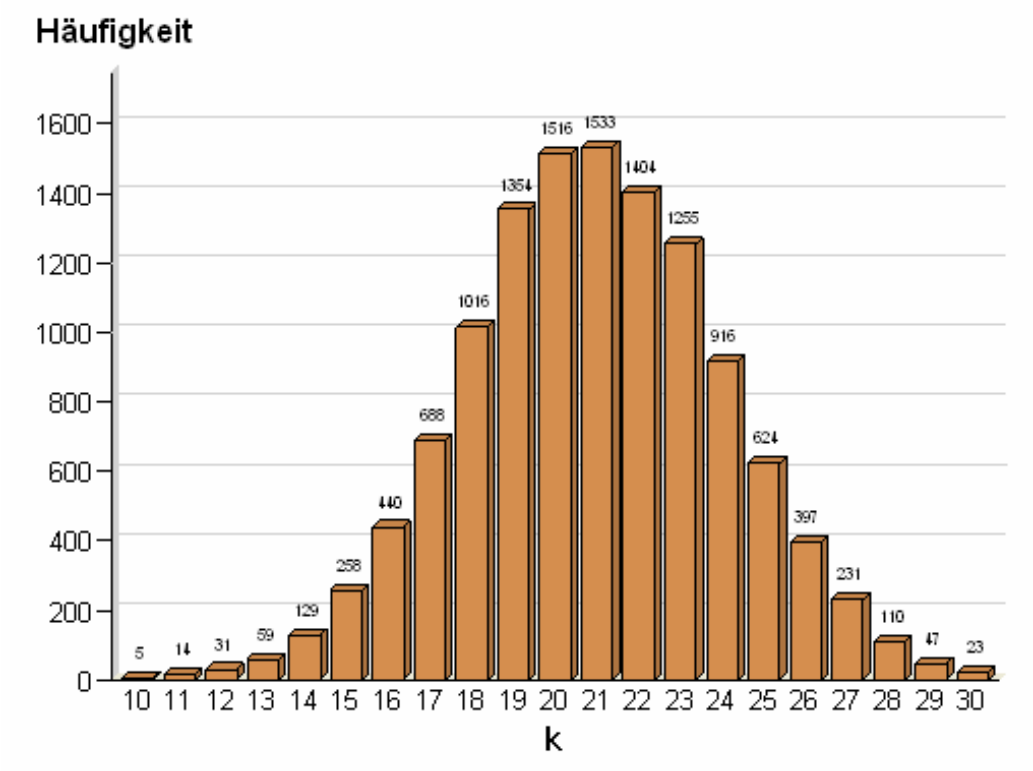


Abbildung 4.3: Bedingte Verteilung der Anzahl an Runs mit $n_1 = 18$, $n_2 = 22$, $E(K \mid N_1 = 24, N_2 = 16) = 20.80$ und $V(K \mid N_1 = 24, N_2 = 17) = 9.54$

Tabelle 4.1: Gute Übereinstimmung bei 10 000 Simulationen von bedingten Erwartungswerten $E(K|N_1, N_2)$ und Mittelwerten m_k sowie von bedingten Varianzen $V(K|N_1, N_2)$ und empirische Varianzen v_k , wenn der Umfang für die Schätzung groß genug ist (Schattierung)

n_2	m_k	$E(K \mid N_1 = n_1, N_2 = n_2)$	v_k	$V(K \mid N_1 = n_1, N_2 = n_2)$	Umfang
28	18.00	17.80	.	6.80	1
27	17.66	18.55	8.60	7.44	12
26	19.54	19.20	7.24	8.02	102
25	19.68	19.75	8.23	8.53	505
24	20.28	20.20	9.04	8.96	1943
23	20.58	20.55	9.37	9.29	5695
22	20.85	20.80	9.56	9.54	12050
21	20.94	20.95	9.46	9.69	18740
20	20.97	21.00	9.81	9.74	21838
19	20.96	20.95	9.70	9.69	18961
18	20.79	20.80	9.55	9.54	12003
17	20.52	20.55	9.20	9.29	5618
16	20.18	20.20	9.01	8.96	1992
15	19.90	19.75	7.73	8.53	457
14	18.98	19.20	10.01	8.02	74
13	18.55	18.55	4.52	7.44	9

Die Übereinstimmung von exakter bedingter Wahrscheinlichkeit $P(K = k \mid N_1 = n_1, N_2 = n_2)$ und Verteilung $F(k) = P(K \le k \mid N_1 = n_1, N_2 = n_2)$ für $n_1 = 22$ und $n_2 = 18$ mit der von Wald und Wolfowitz gegebenen approximativen Normalverteilung $N\left(\frac{2n_1}{1+\alpha}, \frac{4\alpha n_1}{(1+\alpha)^3}\right)$ für $n_1 = \alpha n_2$ auf der einen Seite und den simulierten relativen Häufigkeiten andererseits gibt Tabelle 4.2 an.

Tabelle 4.2: Überprüfung der Übereinstimmung von exakter bedingter Wahrscheinlichkeit $P(K = k \mid N_1 = n_1, N_2 = n_2)$ für $n_1 = 22$ und $n_2 = 18$ mit der von Wald und Wolfowitz gegebenen approximativen Normalverteilung

k	p exakt	p asymp.	Rel. H.	F exakt	F asymp.	F empir.
7	0.00000	0.00001		0.00000	0.00001	
8	0.00002	0.00003	0.0002	0.00002	0.00004	0.0002
9	0.00006	0.00010		0.00008	0.00015	0.0002
10	0.00025	0.00033	0.0001	0.00033	0.00050	0.0002
11	0.00075	0.00095	0.0010	0.00109	0.00149	0.0012
12	0.00222	0.00245	0.0022	0.00331	0.00401	0.0035
13	0.00518	0.00572	0.0057	0.00849	0.00986	0.0092
14	0.01185	0.01204	0.0112	0.02034	0.02209	0.0204
15	0.02200	0.02291	0.0234	0.04234	0.04523	0.0438
16	0.03989	0.03934	0.0391	0.08223	0.08480	0.0829
17	0.05984	0.06100	0.0598	0.14207	0.14592	0.1427
18	0.08726	0.08542	0.0856	0.22933	0.23127	0.2284
19	0.10665	0.10802	0.1050	0.33598	0.33898	0.3333
20	0.12604	0.12334	0.1272	0.46202	0.46183	0.4606
21	0.12604	0.12717	0.1282	0.58807	0.58846	0.5888
22	0.12100	0.11841	0.1201	0.70907	0.70644	0.7088
23	0.09900	0.09955	0.0991	0.80807	0.80578	0.8080
24	0.07700	0.07558	0.0750	0.88507	0.88137	0.8829
25	0.05133	0.05181	0.0547	0.93641	0.93336	0.9377
26	0.03208	0.03208	0.0337	0.96849	0.96567	0.9713
27	0.01728	0.01793	0.0149	0.98577	0.98383	0.9863
28	0.00854	0.00905	0.0075	0.99431	0.99304	0.9938
29	0.00366	0.00413	0.0044	0.99797	0.99727	0.9982
30	0.00139	0.00170	0.0012	0.99937	0.99903	0.9994
31	0.00046	0.00063	0.0003	0.99983	0.99968	0.9998
32	0.00013	0.00021	0.0001	0.99996	0.99991	0.9998
33	0.00003	0.00006	0.0002	0.99999	0.99998	1.0000
34	0.00001	0.00002		1.00000	0.99999	

Ist darüber hinaus vom Zufallszahlengenerator bekannt, dass er gleich verteilte Zufallszahlen X aus dem Intervall von 0 bis 1 liefert, der Erwartungswert mithin E(X) = 0.5 ist, so kann die unbedingte Wahrscheinlichkeitsverteilung $P(K=k)$ für die zufällige Anzahl K der Runs angegeben werden. Es sind

$$P(K=k)=\frac{\binom{n_1-1}{\frac{k}{2}-1}\cdot\binom{n_2-1}{\frac{k}{2}-1}}{\binom{n}{n_1}\cdot 2^{n-1}}$$

für gerades k und

$$P(K=k)=\frac{\binom{n_1-1}{\frac{k-1}{2}}\binom{n_2-1}{\frac{k-3}{2}}+\binom{n_1-1}{\frac{k-3}{2}}\binom{n_2-1}{\frac{k-1}{2}}}{\binom{n}{n_1}\cdot 2^{n}}$$

für ungerades k sowie

$$E(K)=\frac{n+1}{2} \quad \text{und} \qquad V(K)=\frac{n-1}{4}$$

der Erwartungswert und die Varianz. Für große Sequenzlängen n ist die standardisierte Zufallsgröße

$$Z=\frac{K-\frac{n+1}{2}}{\sqrt{\frac{n-1}{4}}}$$

asymptotisch N(0, 1)-verteilt. Von Wishart und Hirschfeld [5] ist unter allgemeineren Voraussetzungen bewiesen worden, dass für Iterationen der Länge n, wobei p die Wahrscheinlichkeit für die 0 und q = 1- p die Wahrscheinlichkeit für die 1, die Verteilung von K asymptotisch beschrieben wird durch

$$Z_1=\frac{K-2npq}{2\cdot\sqrt{npq(1-3pq)}}.$$

Im Spezialfall p = q = ½ , bei einem Zufallszahlengenerator, der gleichverteilte Zufallszahlen zwischen 0 und 1 liefert, geht Z1 über in

$$Z_1=\frac{K-2\cdot n\cdot \tfrac{1}{2}\cdot\tfrac{1}{2}}{2\cdot\sqrt{n\cdot\tfrac{1}{2}\cdot\tfrac{1}{2}\cdot(1-3\cdot\tfrac{1}{2}\cdot\tfrac{1}{2})}}=\frac{K-\frac{n}{2}}{\sqrt{\frac{n}{4}}}\ .$$

Man erkennt sofort, dass Z und Z_1 zwar asymptotisch gleich sind, welche Zufallsgröße aber die tatsächlichen Verhältnisse bei kleinem n besser widerspiegelt, bleibt unklar. Klarheit kann bei konkretem n nur ein Simulationsexperiment bringen.

Beispiel 4.2:
Für n = 40 sind 10000 Simulationen gelaufen und die relativen Häufigkeiten für die Run-Anzahlen K bestimmt worden. In Abbildung 4.4 sind die empirischen Verläufe mit dem Symbol „*“ wiedergegeben, die gestrichelte Linie gibt die asymptotische Näherung nach Wishart und Hirschfeld [5], die durchgezogene Linie diejenige asymptotische Näherung wieder, die auf dem Erwartungswert und der Varianz von K beruht. Man erkennt, dass die letztgenannten asymptotischen Methoden im untersuchten Falle besser an die simulierten Werte passen.

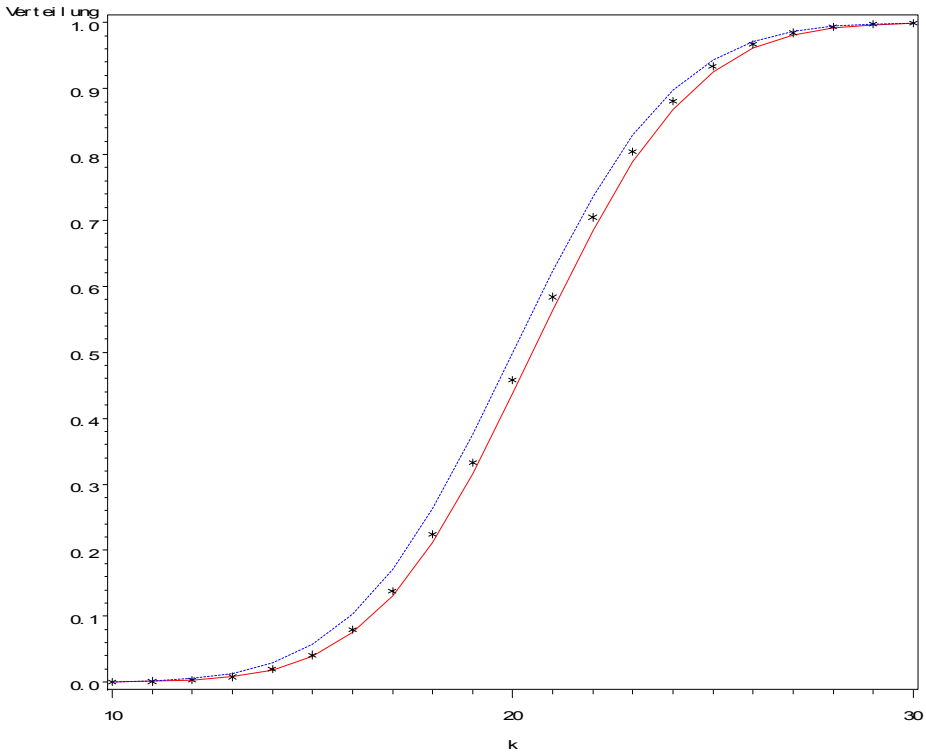


Abbildung 4.4: Run-Anzahlen *K* bei Sequenzlängen von $n = 40$, empirische Verteilungsfunktion und Verteilungsfunktionen der beiden asymptotischen Normalverteilungen mit $\mu_1 = \frac{n}{2}$ und $\sigma_1 = \sqrt{n/4}$ (nach WISHARD und HIRSCHFELD, gestrichelt), sowie mit $\mu_1 = \frac{n+1}{2}$ und $\sigma = \sqrt{(n-1)/4}$ (nach Normierung, volle Linie)

Literatur

[1] Knuth, D.E. (1981): The art of computer programming, Vol. 2 Seminumerical Algorithms, 2. ed. Addison-Wesley, Reading, Mass.

[2] Marsaglia, George: Diehard Battery of Tests of Randomness, http://stat.fsu.edu/pub/diehard/

[3] SAS® Institute Inc. (2004). SAS/STAT 9.1 User's Guide. Cary, NC: SAS® Institute Inc.

[4] SAS Institute Inc. (2004): SAS® 9.1 Macro Language: Reference. Cary, NC: SAS Institute Inc.

[5] Wishart,J., Hirschfeld, H.D.: A theorem concerning the distribution of joins between line segments, J. of the London Math. Soc., 11(1936)227

[6] Wald, A., Wolfowitz, J.: On the test whether two samples are from the same population, Ann. of Math. Statistics 11(1940)147

[7] Fisz, M. (1973): Wahrscheinlichkeitsrechnung und mathematische Statistik, 7. Aufl., Dt. Verl. d. Wiss.