Das Geburtstagsphänomen[1]

Die Wahrscheinlichkeit eines Treffers

Beispielhafte Fragen

- Wie groß ist die Wahrscheinlichkeit, dass von $r = 23$ zufällig in einem Raum befindlichen Leuten einer am 1. April Geburtstag hat?
- Wie groß ist die Wahrscheinlichkeit, unter r unabhängig zufällig gewählten Zeichenketten (über einem Alphabet aus n Zeichen) der Länge t eine bestimmte vorgegebene von den insgesamt möglichen $N = n^t$ zu treffen?
- Wir groß ist die Wahrscheinlichkeit, bei r Zügen aus einer Urne mit N verschieden markierten Kugeln (mit Zurücklegen) eine bestimmte Kugel zu erwischen?

Wahrscheinlichkeitsberechnung

- Es gibt N mögliche Ereignisse ($N \approx 365$ im Falle der Geburtstage).
- Jedes dieser Ereignisse tritt mit der Wahrscheinlichkeit $\frac{1}{N}$ ein, sein Gegenteil mit der Wahrscheinlichkeit $q = 1 - \frac{1}{N}$.
- Bei zwei unabhängigen Versuchen ist die Wahrscheinlichkeit für

keinen Treffer:	q^2,
mindestens einen Treffer:	$1 - q^2$,

  bei r unabhängigen Versuchen ist die Wahrscheinlichkeit für

keinen Treffer:	q^r,
mindestens einen Treffer:	$1 - q^r$

Satz 1 (i) *Die Wahrscheinlichkeit, bei r unabhängigen Ereignissen aus einer Menge von N möglichen ein bestimmtes vorgegebenes zu beobachten, ist*

$$1 - \left(1 - \frac{1}{N}\right)^r .$$

(ii) *Ist $N \geq 2$, so ist diese Wahrscheinlichkeit $\geq$ einem vorgegebenen Wert p, wenn*

$$r \geq \frac{\ln(1-p)}{\ln(1-\frac{1}{N})}$$

(iii) *... oder wenn*

$$r \geq N \cdot |\ln(1-p)|.$$

[1] Klaus Pommerening, Kryptologie; letzte Änderung: 30. April 2002

Beweis. Die Formel in (ii) folgt über die äquivalenten Zwischenumformungen

$$\begin{aligned} 1 - \left[1 - \frac{1}{N}\right]^r &\geq p, \\ 1 - p &\geq \left[1 - \frac{1}{N}\right]^r, \\ \ln(1-p) &\geq r \cdot \ln(1 - \frac{1}{N}), \end{aligned}$$

weil $\ln(1 - \frac{1}{N})$ negativ ist.

(iii) folgt, weil $\ln(1 - x) \leq -x$ für $0 < x < 1$, also $\ln(1 - \frac{1}{N}) \leq \frac{1}{N}$. Ist also $r \geq N \cdot |\ln(1-p)|$, so ist erst recht die Voraussetzung von (ii) erfüllt. $\diamond$

Anwendungen

Geburtstage 1: Für $N \approx 365.22$, $r = 23$, ist die Wahrscheinlichkeit eines Treffers $p \approx 1 - 0.99726^{23} \approx 0.0611$.

Geburtstage 2: Wieviele Leute müssen im Raum sein, damit die Wahrscheinlichkeit mindestens $\frac{1}{2}$ ist? Nach Aussage (ii) im Satz ist die Mindestzahl für $p = \frac{1}{2}$

$$\frac{\ln(0.5)}{\ln(0.99726)} \approx \frac{0.6931}{0.002742} \approx 252.8,$$

also 253.

Zeichenketten: Die Wahrscheinlichkeit, in einer zufälligen Zeichenfolge der Länge 1000 über dem Alphabet {A,...,Z} eine vorgegebene Kette der Länge 4 - etwa „DUNJ“ - anzutreffen ist in erster Näherung (die die durch die Überlappung gegebene Abhängigkeit ignoriert) ungefähr

$$1 - (1 - \frac{1}{456976})^{1000} \approx 1 - (0.999999781)^{1000} \approx 1 - 0.9978 = 0.0022,$$

also etwa 2 Promille.

Die Wahrscheinlichkeit eines Zusammentreffens (Kollision)

Beispielhafte Fragen

- Wie groß ist die Wahrscheinlichkeit, dass von $r = 23$ zufällig in einem Raum befindlichen Leuten mindestens zwei am gleichen Tag Geburtstag haben? (Egal an welchem!)

- Wie groß ist die Wahrscheinlichkeit, unter r unabhängig zufällig gewählten Zeichenketten der Länge t mindestens zwei übereinstimmen?
- Wie groß ist die Wahrscheinlichkeit, bei r Zügen aus einer Urne mit N verschieden markierten Kugeln (mit Zurücklegen) eine Kugel mindestens zweimal zu erwischen? (Egal welche!)

Wahrscheinlichkeitsberechnung

- Die Wahrscheinlichkeit, dass das erste Ereignis eine Wiederholung ist, ist 0.
- Die Wahrscheinlichkeit, dass das erste Ereignis *keine* Wiederholung ist, ist also $1 = \frac{N}{N}$.
- Die Wahrscheinlichkeit, dass das zweite Ereignis keine Wiederholung ist, ist $\frac{(N-1)}{N}$.
- Die Wahrscheinlichkeit, dass **dann auch** das dritte Ereignis keine Wiederholung ist, ist $\frac{(N-2)}{N}$. (Soviele Auswahlmöglichkeiten gibt es dann noch, die keine Wiederholung verursachen.)
- Allgemein gilt: Ist bisher noch keine Wiederholung aufgetreten, so ist die Wahrscheinlichkeit, dass auch im r-ten Versuch keine Wiederholung auftritt, $\frac{(N-r+1)}{N}$.

Daraus folgt:

Satz 2 *Die Wahrscheinlichkeit, bei r unabhängigen Ereignissen aus einer Menge von N möglichen eine Wiederholung („Kollision“) zu beobachten, ist*

$$C(N,r) = 1 - P(N,r)$$

mit

$$P(N,r) = \frac{N \cdot (N-1) \cdots (N-r+1)}{N^r} = [1 - \frac{1}{N}] \cdots [1 - \frac{(r-1)}{N}] \leq e^{-\frac{r(r-1)}{2N}}.$$

Beweis. Die letzte Ungleichung ist noch zu beweisen. Sie folgt aus der Ungleichung $1 - x \leq e^{-x}$ für $x \in \mathbb{R}$, also

$$P(N,r) \leq e^{-\frac{1}{N}} \cdots e^{-\frac{r-1}{N}} \leq e^{-\frac{r(r-1)}{2N}},$$

wie behauptet. $\diamond$

Anwendungen

Geburtstage: Für $N \approx 365.22$, $r = 23$, ist $P(N, r) \approx 0.493$, die Wahrscheinlichkeit eines Zusammentreffens also ≈ 0.507.

Sind 23 Leute in einem Raum, ist die Wahrscheinlichkeit, dass zwei davon den gleichen Geburtstag haben, größer als $\frac{1}{2}$.

Dieses auf den ersten Blick verblüffende Ergebnis wird als „Geburtstagsphänomen“ oder gar als „Geburtstagsparadox“ bezeichnet.

Zeichenketten 1: Die Wahrscheinlichkeit p_t, in einer zufälligen Zeichenfolge der Länge 1000 über dem Alphabet {A,...,Z} eine wiederholte Kette der Länge t anzutreffen ist ungefähr

t	1	2	3	4	5	6	7
p_t	1.00	1.00	1.00	0.665	0.041	0.0016	0.000062

D. h., eine wiederholte Viererkette ist noch recht wahrscheinlich, eine wiederholte Fünferkette schon ziemlich unwahrscheinlich.

Asymptotisches Verhalten

Für die Zahl der Kollisionen $C(N, r)$ erhält man leicht eine obere Schranke:

- Die Wahrscheinlichkeit, dass das i-te Ereignis eine Kollision ist, ist $\leq \frac{i-1}{N}$ – denn bisher sind nur $i-1$ Ereignisse überhaupt aufgetreten.
- Die Wahrscheinlichkeit, dass spätestens beim r-ten Ereignis eine Kollision aufgetreten ist, ist also

$$C(N, r) \leq \frac{0}{N} + \cdots + \frac{i-1}{N} + \cdots + \frac{r-1}{N} = \frac{r(r-1)}{2N}$$

Damit ist gezeigt:

Satz 3 (i) *Für die Kollisionswahrscheinlichkeit* $C(N, r)$ *gilt*

$$1 - e^{-\frac{r(r-1)}{2N}} \leq C(N, r) \leq \frac{r(r-1)}{2N}.$$

(ii) *Ist* $r \leq \sqrt{2N}$*, so gilt sogar*

$$(1 - \frac{1}{e}) \cdot \frac{r(r-1)}{2N} \leq C(N, r) \leq \frac{r(r-1)}{2N}.$$

oder, abgeschwächt,

$$0.3 \cdot \frac{r(r-1)}{N} \leq C(N, r) \leq 0.5 \cdot \frac{r(r-1)}{N}.$$

Beweis. Zu beweisen ist noch die untere Schranke in (ii). Sie folgt aus der Ungleichung $1 - e^{-x} \geq (1 - \frac{1}{e})x$ für $0 \leq x \leq 1$. Denn $f(x) = 1 - e^{-x}$ ist konkav, $g(x) = (1 - \frac{1}{e})x$ ist linear, und $f(0) = g(0)$, $f(1) = g(1)$ $\diamond$

Daraus ergibt sich als weitere Anwendung:

Zeichenketten 2: Bei zufälligen Texten der Länge $r = n^{t/2}$ (über einem Alphabet aus n Buchstaben) ist die Wahrscheinlichkeit für das Auftreten eines wiederholten t-Gramms $\leq \frac{r^2}{2N} = \frac{n^t}{2n^t} = \frac{1}{2}$.

Bis zu einer Länge von $n^{t/2}$ ist in zufälligen Zeichenketten eine t-Gramm-Wiederholung eher unwahrscheinlich.

Oder umgekehrt ausgedrückt: *Bei zufälligen Zeichenketten der Länge r ist für*

$$t \geq 2 \cdot \frac{^2\log(r)}{^2\log(n)}$$

eine t-Gramm-Wiederholung eher unwahrscheinlich.

Bei $n = 26$ liegt diese Grenze bei $0.425 \cdot {}^2\log(r)$ In der Tabelle oben mit $r = 1000$ entspricht das dem Wert ≈ 4.2.