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; 18. Oktober 1999, letzte Änderung: 15. Mai 2005

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, unter 1000 zufällig gewählten Zeichenketten der Länge 4 über dem Alphabet {A,...,Z} eine vorgegebene Kette – etwa „DUNJ“ – anzutreffen, ist ungefähr

$$1 - (1 - \frac{1}{26^4})^{1000} \approx 1 - (0.999999781)^{1000} \approx 1 - 0.9978 = 0.0022,$$

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

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

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: Aus r Zeichenketten über dem Alphabet {A,...,Z} werden r Stück der Länge t zufällig und unabhängig ausgewählt; es gibt also $N = 26^t$ mögliche Ereignisse. Die Wahrscheinlichkeit, dass mindestens

zwei dieser Zeichenketten übereinstimmen, ist $C(26^t, r)$. Für $r = 100$, 300, 1000, 5000 seien diese Wahrscheinlichkeiten mit p_t, q_t, r_t, s_t bezeichnet. Direkte Berechnung nach Satz 2 – mit Hilfe eines kleinen Programms – ergibt die folgende Tabelle (in der Einträge $< 10^{-4}$ weggelassen wurden):

$t \rightarrow$	1	2	3	4	5	6	7	$r \downarrow$
p_t	1.000	**1.000**	**0.246**	0.011	0.00042			100
q_t	1.000	1.000	**0.923**	**0.094**	0.0038	0.00015		300
r_t	1.000	1.000	1.000	**0.665**	**0.041**	0.0016		1000
s_t	1.000	1.000	1.000	1.000	**0.651**	**0.040**	0.0016	5000

Das bedeutet z. B., dass für $r = 1000$ die Chance, zwei identische Ketten der Länge 4 zu finden, größer als 60% ist; aber zwei identische Ketten der Länge 5 zu finden, ist sehr unwahrscheinlich ($< 5\%$ Wahrscheinlichkeit). In jeder Zeile liegt die Grenze „50% Wahrscheinlichkeit“ zwischen den beiden fettgedruckten Einträgen.

Schranken für die Anzahl der Kollisionen

Die Formel in Satz 2 ist für eine manuelle Berechnung ziemlich ungeeignet; außerdem gibt sie keine Vorstellung von der Größe der Wahrscheinlichkeit. Glücklicherweise kann man mit Hilfe von etwas elementarer Analysis bequeme Schranken herleiten, die auch das asymptotische Verhalten deutlich machen. Zunächst erhält man leicht eine obere Schranke für die Zahl der Kollisionen $C(N, r)$:

- 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 sind schon die rechten Ungleichungen im folgenden Satz gezeigt:

Satz 3 (i) *Die Kollisionswahrscheinlichkeit $C(N, r)$ ist beschränkt durch*

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

(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,

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

(iii) *Ist* $r \leq \sqrt{N}$*, so* $C(N,r) < \frac{1}{2}$.

(iv) *Ist* $r \geq 1 + \sqrt{2 \ln 2} \cdot \sqrt{N}$*, so* $C(N,r) > \frac{1}{2}$.

Beweis. Die linke Schranke in (i) folgt aus der Ungleichung $1 - x \leq e^{-x}$ für $x \in \mathbb{R}$, also

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

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

Die untere Schranke in (ii) 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 $f(0) = g(0)$, $f(1) = g(1)$.

In (iii) vereinfacht sich die obere Schranke aus (ii) zu $C(N,r) < \frac{r^2}{2N} \leq \frac{N}{2N} = \frac{1}{2}$.

(iv) folgt sofort aus der linken Seite von (i). $\diamond$

Die Aussagen (iii) und (iv) fasst man meist zusammen zu der Faustregel:

> *Etwa ab* $r > \sqrt{N}$ *ist die Kollisionswahrscheinlichkeit* $C(N,r)$ *größer als* $\frac{1}{2}$.

Als Spezialfall von (iii) folgt sofort:

Satz 4 *Für* $r \leq n^{t/2}$ *Zeichenketten der Länge* t *über einem Alphabet der Länge* n *ist die Kollisionswahrscheinlichkeit kleiner als* $\frac{1}{2}$.

Die Wahrscheinlichkeit einer zufälligen Wiederholung

Diese Ergebnisse werden jetzt auf die Teilketten einer zufälligen Zeichenkette (über einem n-buchstabigen Alphabet) angewendet; dabei bedeutet „zufällig“, dass jedes Zeichen der Kette unabhängig mit Wahrscheinlichkeit jeweils $\frac{1}{n}$ gewählt wird. Von jetzt an wird die mathematische Exaktheit durch die **vereinfachende Annahme** abgeschwächt, dass die Teilketten stochastisch unabhängig sind; da sich die Teilketten überlappen, ist dies offenbar nicht korrekt, es wirkt sich aber, wie man an Simulationsrechnungen sehen kann, nicht wesentlich aus. Außerdem wird vernachlässigt, dass eine Zeichenkette der Länge r nur $r-t+1$ Teilketten der Länge t hat. Dann ist die Wahrscheinlichkeit für eine Wiederholung der Länge t (ungefähr) $C(n^t, r)$, und für den Fall $n = 26$ kann die obige Tabelle zu Rate gezogen werden.

Aus Satz 4 kann man daher folgern: Für eine zufällige Zeichenkette der Länge $r < n^{t/2}$ (über einem n-buchstabigen Alphabet) ist die Wahrscheinlichkeit, dass eine Wiederholung der Länge t vorkommt, $< \frac{1}{2}$. Das bedeutet: *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

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

eine t-Gramm-Wiederholung eher unwahrscheinlich.

Für $n = 26$ liegt diese Grenze bei $0.425 \cdot {}^2\log(r) \approx 1.413 \cdot \log(r)$. Das bedeutet zum Beispiel:

- Für Texte der Länge 100 sind zufällige Wiederholungen der Länge 3 oder mehr ziemlich unwahrscheinlich; die Tabelle ergibt genauer, dass die Wahrscheinlichkeit $< 25\%$ ist.
- Für Texte der Länge 300 sind zufällige Wiederholungen der Länge 4 oder mehr ziemlich unwahrscheinlich (nach der Tabelle ist die Wahrscheinlichkeit $< 10\%$), aber es gibt sehr wahrscheinlich wenigstens eine zufällige Wiederholung der Länge 3 (Tabelle: $> 90\%$).

Und so weiter – nach der obigen Faustformel (A), der Tabelle oder Satz 3.

Wenn der Kryptoanalytiker eine Periodenanalyse nach Kasiski durchführt, untersucht er allerdings ja gar keinen zufälligen Text. Um die vorstehenden Ergebnisse anwenden zu können, hat er eine **weitere vereinfachende Annahme** nötig: Ein polyalphabetischer Geheimtext verhält sich genügend zufällig mit Ausnahme der Effekte, die durch die Periode bedingt sind. Findet er nun eine Wiederholung der Länge t, und t ist wenigstens so groß wie in der Faustformel (A), so kann er ziemlich sicher sein, dass er eine „echte“, d. h. durch die Periodizität der Chiffre hervorgerufene Wiederholung gefunden hat und die Periode somit ein Teiler des Abstands ist. Je kleiner t ist, desto mehr ist er darauf gefasst, dass er einige der gefundenen Wiederholungen verwerfen muss; findet er aber eine „lange“ Wiederholung, so kann er mit an Sicherheit grenzender Wahrscheinlichkeit annehmen, dass diese echt ist.