Das Geburtstags-Problem
Einige empirische Daten und einige Approximationen +

von Thomas R. Knapp
University of Rochester

Übersetzt von Ingeborg Strauß

"Das Geburtstags-Problem" (Wie groß ist die Wahrscheinlichkeit, daß bei n zufällig ausgewählten Personen wenigstens zwei den - abgesehen vom Jahr - selben Geburtstag haben?) fasziniert Lehrer und Studenten der Statistik seit vielen Jahren vor allem deshalb, weil die Wahrscheinlichkeit selbst für relativ kleines n hoch ist (z.B. ca. 0,50 für n = 23, ca. 0,99 für n = 60). Die gängige mathematische Lösung dieses Problems beruht auf der Annahme, daß kein Tag des Jahres als Geburtstag bevorzugt oder benachteiligt ist. Schaltjahre und Mehrfachgeburten werden ignoriert.

Bloom (1973), Munford (1976) und andere wiesen nach, daß diese Annahme einer Gleichverteilung einer Überprüfung nicht standhält: die Wahrscheinlichkeit eines 'birthday-sharing' ist sogar größer. (Wenn jeder Mensch am 1. Januar geboren wäre, würde die Wahrscheinlichkeit für jedes n gleich 1 sein.) Bissell (1980) nennt diese Prämisse der Gleichverteilung gewagter als andere Vereinfachungen, die wir häufig zur Lösung eines realen Problems benutzen.

Wie stark verletzt denn nun die Annahme der Gleichverteilung die tatsächlichen Gegebenheiten? Für die 28-Jahres-Periode von 1941-1968 liefern die statistischen Daten des im Staat New York liegenden Monroe County (Rochester und Umgebung) eine Teilantwort. (Man benötigt die Daten von 28 Jahren, um Schaltjahre und den Wochenend-Effekt - weniger Geburten an Samstagen und Sonntagen verglichen mit den Werktagen - angemessen berücksichtigen sowie kurzzeitige Klumpungen und Ausdünnungen nivellieren zu können.) Die Prozentzahlen der Geburten für jeden Tag eines Jahres sind (bezogen auf obengenannten Zeit- und geographischen Raum) in Tabelle 1 aufgelistet.

Obgleich die meisten Angaben dem auf Gleichverteilung beruhenden Zahlenwert $\frac{1}{366}\cdot 100 = 0{,}2732$ ziemlich nahe liegen, gibt es einige recht interessante Abweichungen. Der kleinste Prozentsatz ist natürlich der vom 29. Februar, aber das ist ein Kunstprodukt des Schaltjahr-Phänomens. Für die anderen 365 Tage des Jahres reicht die Spannweite von 0,2222 für den 24. Dezember bis zu 0,3131 für den 1. Oktober: etwa 40% mehr Geburten zum letzten Termin verglichen mit dem erstgenannten.

+Originaltitel in 'TEACHING STATISTICS' (1982) Heft 1, Band 4
'The Birthday Problem: Some Empirical Data and Some Approximations'

TABELLE 1 Prozentzahlen der Geburten in der Monroe County für jeden Tag der Jahre 1941-1968

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
1	0·2320	0·2595	0·2755	0·2726	0·2764	0·2862	0·2924	0·2586	0·2773	0·3131	0·2557	0·2592
2	0·2347	0·2675	0·2776	0·2805	0·2622	0·2891	0·2921	0·2714	0·2773	0·3039	0·2921	0·2625
3	0·2412	0·2575	0·2942	0·2791	0·2678	0·2669	0·2711	0·2729	0·2947	0·2977	0·2717	0·2681
4	0·2421	0·2563	0·2888	0·2820	0·2746	0·2652	0·2646	0·2805	0·2856	0·2882	0·2687	0·2761
5	0·2563	0·2581	0·2717	0·2684	0·2776	0·2847	0·2563	0·2817	0·2711	0·2829	0·2746	0·2743
6	0·2545	0·2521	0·2678	0·2752	0·2847	0·2622	0·2708	0·2817	0·2652	0·2894	0·2859	0·2640
7	0·2604	0·2726	0·2829	0·2663	0·2690	0·2956	0·2773	0·2847	0·2740	0·2678	0·2669	0·2693
8	0·2444	0·2649	0·2874	0·2669	0·2702	0·2746	0·2826	0·2811	0·2740	0·2791	0·2788	0·2770
9	0·2512	0·2480	0·2640	0·2530	0·2876	0·2755	0·2814	0·2859	0·2779	0·2791	0·2794	0·2601
10	0·2560	0·2595	0·2746	0·2808	0·2817	0·2909	0·2711	0·2746	0·2897	0·2696	0·2708	0·2731
11	0·2631	0·2640	0·2681	0·2672	0·2663	0·2675	0·2791	0·2906	0·2947	0·2767	0·2788	0·2776
12	0·2607	0·2634	0·2734	0·2820	0·2770	0·2832	0·3066	0·2666	0·2761	0·2817	0·2909	0·2773
13	0·2575	0·2521	0·2734	0·2714	0·2749	0·2720	0·2767	0·2687	0·2734	0·2693	0·2758	0·2578
14	0·2687	0·2885	0·2758	0·2743	0·2604	0·2687	0·2918	0·2903	0·2891	0·2758	0·2791	0·2711
15	0·2601	0·2655	0·2743	0·2669	0·3004	0·2811	0·2619	0·2782	0·3084	0·2640	0·2702	0·2791
16	0·2702	0·2613	0·2737	0·2779	0·2655	0·2847	0·2933	0·3039	0·3007	0·2906	0·2548	0·2856
17	0·2471	0·2640	0·2669	0·2681	0·2841	0·2776	0·2906	0·2737	0·3019	0·2717	0·2586	0·2817
18	0·2400	0·2631	0·2572	0·2714	0·2788	0·2705	0·2767	0·2687	0·2930	0·2696	0·2708	0·2708
19	0·2613	0·2811	0·2628	0·2551	0·2726	0·2690	0·2705	0·2805	0·2959	0·2764	0·2554	0·2669
20	0·2501	0·2690	0·2705	0·2660	0·2897	0·2814	0·2906	0·2731	0·2882	0·2749	0·2779	0·2794
21	0·2773	0·2696	0·2660	0·2681	0·2761	0·2800	0·2868	0·2770	0·3051	0·2758	0·2734	0·2575
22	0·2433	0·2687	0·2720	0·2711	0·2808	0·2687	0·2865	0·2811	0·2832	0·2669	0·2619	0·2486
23	0·2554	0·2779	0·2805	0·2723	0·2791	0·2868	0·2823	0·2776	0·3093	0·2604	0·2560	0·2406
24	0·2569	0·2797	0·2705	0·2734	0·2921	0·2814	0·2669	0·2956	0·2989	0·2800	0·2628	0·2222
25	0·2622	0·2841	0·2622	0·2841	0·2791	0·2776	0·2921	0·2900	0·2874	0·2563	0·2717	0·2240
26	0·2589	0·2690	0·2865	0·2601	0·2900	0·2856	0·2802	0·2962	0·3090	0·2702	0·2566	0·2433
27	0·2655	0·2800	0·2548	0·2660	0·2950	0·2726	0·2956	0·2746	0·2936	0·2726	0·2785	0·2616
28	0·2720	0·2604	0·2805	0·2660	0·2634	0·2705	0·2808	0·2965	0·2879	0·2675	0·2958	0·2850
29	0·2578	0·0675	0·2773	0·2729	0·2832	0·2708	0·2767	0·2770	0·2977	0·2785	0·2835	0·3013
30	0·2507		0·2746	0·2652	0·2604	0·2536	0·2868	0·2767	0·2885	0·2956	0·2607	0·2720
31	0·2592		0·2622		0·2702		0·2874	0·2823		0·2637		0·2915

Die größeren Prozentzahlen gehören zu Tagen der Monate Juli bis Oktober, wobei der September die stärksten Geburtsraten aufweist. Die kleinen Prozentwerte finden sich im Dezember und Januar. Natürlich zeigen sich Variationen auch innerhalb der Monate, doch ist für den Januar auffällig, daß nur der 21. über dem Mittelwert aller Daten liegt, während der September nur an zwei Tagen (dem 5. und 6.) unter diesen Durchschnitt rutscht. Im Juni ballen sich die meisten Angaben eng um das Jahresmittel.

Es sei p_i der Anteil der Geburten am Tage i (i = 1,2,...,366). Die Wahrscheinlichkeit, daß von n Personen, die zufällig aus dieser Population (bestehend aus 337914 Personen) ausgewählt werden, mindestens zwei am gleichen Tag Geburtstag haben, ergibt sich aus dem folgenden Ausdruck:

$$P_n = 1 - \sum_{\substack{i,j,k,\ldots=1 \\ i\neq j\neq k\neq\ldots}}^{366} p_i p_j p_k \cdots \text{, n Faktoren.} \qquad (1)$$

Die Summe besteht aus $n!\binom{366}{n} = \frac{366!}{(366-n)!}$ Termen, einer - selbst für so kleine n wie 5 oder 10 - außerordentlich großen Anzahl, was rasch die Grenzen auch der schnellsten und größten Computer übersteigt. Deshalb müssen Approximations-Techniken zu Hilfe genommen werden, will man die P_n berechnen. Man setze $p_i = c + \varepsilon_i$ mit $c = \frac{1}{366} = 0{,}002732$. Dann kann Gleichung (1) so geschrieben werden:

$$P_n = 1 - \sum_{\substack{i,j,k,\ldots=1 \\ i\neq j\neq k\neq\ldots}}^{366} (c+\varepsilon_i)(c+\varepsilon_j)(c+\varepsilon_k)\cdots . \qquad (2)$$

Für n = 3 ergibt sich beispielsweise

$$P_3 = 1 - \sum_{\substack{i,j,k=1 \\ i\neq j\neq k}}^{366} \left[c^3 + c^2\varepsilon_i + c^2\varepsilon_j + c^2\varepsilon_k + c\varepsilon_i\varepsilon_j + c\varepsilon_i\varepsilon_k + c\varepsilon_j\varepsilon_k + \varepsilon_i\varepsilon_j\varepsilon_k \right]. \qquad (3)$$

Man wende das Distributivgesetz auf das Summationszeichen an und beachte noch, daß $\sum \varepsilon_i = \sum \varepsilon_j = \sum \varepsilon_k = 0$ und $\sum \varepsilon_i\varepsilon_j = \sum \varepsilon_i\varepsilon_k = \sum \varepsilon_j\varepsilon_k$ ist.

Diese Vereinfachungen führen zu

$$P_3 = 1 - \left[3!\binom{366}{3}c^3 + 366\cdot 3c\cdot \sum_{\substack{i,j=1 \\ i\neq j}}^{366} \varepsilon_i\varepsilon_j + \sum_{\substack{i,j,k=1 \\ i\neq j\neq k}}^{366} \varepsilon_i\varepsilon_j\varepsilon_k \right]. \qquad (4)$$

Aus den Werten der Tabelle 1 folgt die Bandbreite der ε_i: von -0,002057 (für den 29. Februar) bis 0,000399 (für den 1. Oktober). Das heißt, da die Produkt-Terme $\varepsilon_i\varepsilon_j\varepsilon_k$ einen sehr kleinen Betrag haben,

$$P_3 = 1 - \left[3!\binom{366}{3}c^3 + 366\cdot 3c\cdot \sum_{\substack{i,j=1 \\ i\neq j}}^{366} \varepsilon_i\varepsilon_j \right]. \qquad (5)$$

Ignoriert man auch im allgemeinen Fall die ε-Produkte mit drei und mehr Faktoren, lautet die Beziehung

$$P_n = 1 - \left[n!\binom{366}{n}c^n + (n-2)!\binom{366}{n-2}\binom{n}{2}c^{n-2}\cdot \sum_{\substack{i,j=1 \\ i\neq j}}^{366} \varepsilon_i\varepsilon_j \right] \qquad (6)$$

was auch in folgender Form geschrieben werden kann:

$$P_n = 1 - \frac{365!}{(366-n)!366^{n-1}} - \frac{n(n-1)365!}{2(368-n)!366^{n-3}}\cdot \sum_{\substack{i,j=1 \\ i\neq j}}^{366} \varepsilon_i\varepsilon_j. \qquad (7)$$

Umgeformt erhält man [2]

$$P_n = 1 - \frac{365!}{(366-n)!\cdot 366^{n-1}}\left(1 - \frac{n(n-1)}{2}\cdot\frac{366^2}{(368-n)(367-n)}\cdot \sum_{\substack{i,j=1 \\ i\neq j}}^{366} \varepsilon_i\varepsilon_j\right) \qquad (8)$$

Bei nur leichter Abweichung von der Gleichverteilung ist $\sum \varepsilon_i\varepsilon_j$ negativ mit einem relativ kleinen Betrag. Aus Gleichung (8) ist ersichtlich, daß der Term $\frac{n(n-1)\cdot 365!}{2\cdot(368-n)!\cdot 366^{n-3}}\cdot \sum_{\substack{i,j=1 \\ i\neq j}}^{366} \varepsilon_i\varepsilon_j$ nur für recht großes n zu Buche schlägt. Da aber beide Terme von (7) für großes n nahe bei 0 liegen, wird man nur für stark ungleichgewichtige Verteilungen erwarten, daß die empirisch gewonnenen P_n wesentlich größer als die a priori-P_n ausfallen.

Die Tabelle 1 liefert $\sum_{\substack{i,j=1 \\ i\neq j}}^{366} \varepsilon_i\varepsilon_j = -0{,}0000112427$. Tabelle 2 listet die durch Gleichung (7) approximativ erhaltenen Werte der P_n für n = 10,20,30, 40,50 und 60 zusammen mit den korrespondierenden a priori-P_n auf. Letztere sind dem Buch von Kemeny, Snell und Thompson (1957) entnommen, die auf der Basis eines 365-Tage-Jahres und mit Gleichverteilung und Unabhängigkeit der Daten gerechnet haben.

TABELLE 2

Ein Vergleich der empirischen und der a priori-Werte für P_n

n	P_n (unter Benutzung von (7))	P_n (a priori)
10	0,1171	0,117 (tatsächlich 0,1166)
20	0,4119	0,411
30	0,7070	0,706
40	0,8917	0,891
50	0,9706	0,970
60	0,9942	0,994

Die aus der Erfahrung gewonnenen P_n liegen, wie erwartet, alle höher als die zugehörigen a priori-P_n. Der Unterschied ist jedoch nicht groß, da die Abweichung der aus Tabelle 1 stammenden p_i von der Gleichverteilung gering ist. Dagegen ist in den Südstaaten wie z.B. Alabama ein stärkeres Ungleichgewicht zu konstatieren (vgl. Rosenberg 1963). Die empirischen und die a priori-Werte klaffen viel stärker auseinander, weil im Spätsommer (August und September) nahezu 25% mehr Geburten registriert werden als zu Frühlingsende (April und Mai), ein Verteilungsmodell, das konträr zu dem Saison-Verteilungsmodell der meisten europäischen Länder ist.

Der Autor würde gern andere über 28 Jahre sich erstreckende Geburtstagsdatensammlungen, sofern solche existieren, untersuchen, um festzustellen, welchen Grad der Übereinstimmung die nach Gleichung (7) berechneten P_n dieser Stichproben mit den a priori-P_n zeigen.

Literatur:

Bissell, S.F. (1980), "Breakdowns and Birthdays", Teaching Statistics, 2,15-18

Bloom, D.M. (1973), "A Birthday Problem", American Mathematical Monthly, 80, 1141-1142

Kemeny, J.G., Snell, J.L., and Thompson, G.L. (1957), Introduction to Finite Mathematics, Englewood Cliffs, N.J.: Prentice-Hall, Inc. [3]

Munford, A.G. (1977), "A Note on the Uniformity Assumption in the Birthday Problem", The American Statistician, 31,119

Rosenberg, H.M. (1966), Seasonal Variation of Births, United States 1933-1963, Series 21, No 9, Washington, D.C.: U.S. Department of Health, Education, and Welfare, Public Health Service, National Center for Health Statistics

Vital Statistics of the United States: Natality (various years), Washington, D.C.: U.S. Department of Health, Education, and Welfare, Public Health Service, National Center for Health Statistics.

Mein Dank gilt Donald Peoples und den Mitarbeitern der Birth and Death Records division des Monroe County Department of Health für ihre Unterstützung beim Erheben der Daten sowie William Arcuri, Michael Davidson und Avadis Tevanian für ihre Hilfe und Beratung hinsichtlich der rechnerischen Aspekte dieser Studie.

Anmerkungen der Übersetzerin:

[1] Druckfehler im Original: $\binom{366}{3}$ statt $\binom{366}{n}$

[2] Gleichung (8) nicht im Original. Die darauf folgenden Ausführungen dieses Absatzes sind der hoffentlich besseren Verständlichkeit wegen frei übersetzt. Mein Dank gilt Herrn Prof. Fillbrunn für seine Hilfe bei der Gestaltung dieser Passage.

[3] Deutsche Ausgabe: J. G. Kemeny, J. L. Snell, G. L. Thompson, Einführung in die endliche Mathematik, Fachverlag für Wirtschaftstheorie und Ökonometrie, Ludwigshafen am Rhein, 1963 (besonders S. 129 ff.)