Mathe Ausarbeitung

über

Geburtstagsproblem Teil II

für den Kurs

Angewandte Informatik

an der

DHBW Mosbach

von

Tim Hönnige, Matthias Jooß und Liam Friedrich

Fällig am	31.03.2021
Kurs	INF19A
Dozent	David Weniger

Zusammenfassung

Sie sitzen im Wartezimmer. Die Darmspiegelung, die sie bis jetzt aufgeschoben haben steht nun an. Sie beschließen sich mit einem Matheproblem abzulenken. Das Wartezimmer füllt sich langsam mit Menschen und sie überlegen wie Wahrscheinlich es ist, dass einer dieser Personen am selben Tag wie sie Geburtstag hat. Sie knobeln eine Weile und kommen schließlich zu dem Schluss, dass das Wartezimmer garnicht genug Raum bietet, um so viele Personen zu beherbergen, dass die Wahrseinlichkeit einer Geburtstagsüberschneidung auch nur bei 50% liegt. Sie geben auf und googeln und tatsächlich die Wahrscheinlichkeit dass ein Person im Raum mit ihnen Geburtstag hat ist mit der 23 Person die gerade das Wartezimmer betrat auf 23 gestiegen. Die Mathematik dahinter finden sie hier.

Inhaltsverzeichnis

1 Einführung

Das Geburtstagsproblem generiert eine Warschleinlichket das unter k zufällig gewählten Personen mindestens zwei am selben Tag geburtstag haben. Es ist eine Abwandlung des Paradoxons der ersten Kollision. [1] Bei k=23 ist die warscheinlickteit das zwei personen bereits am gleichen tag geburtsagt haben bei über 50%. Das ist erstaunlich stochastisch weniger bewanderte personen fürden stehts subjektiv zu einer sehr viel geringeren warscheinlichkeit kommen. In der folgenden ausarbeotung wrrden wir auf die näherung für das paradoxon eingehen sowie quanitle und den erwartungswert.

das an irgendeinem Tag im Jahr irgendeine der k personen

zuerst bestimmen wir n. Das ist die warscheinlichkeit das eine person an einem bestimmten Tag der jahren geburtstag hat. Das gewählte Jahr besitzt 365 Tage (Auf Schaltjahre wird nicht nachgegangen). Außerdem gehen wir von gleich gewichteten Tagen aus.

Daraus ergibt sich:

$$n = \frac{1}{365} \tag{1.1}$$

Um auszurechnen, wie viele Personen sich in einem Raum befinden müssen, so dass die Wahrscheinlichkeit dass mindestens zwei Personen am selben Tag Geburtstag haben bei 50% oder meht liegt verwenden wir das Gegenereignis. Wir berechnen die Wahrscheinlichkeit, dass alle Personen im Raum an verschiedenen Tagen Geburtstag haben und nähern uns den von oben den 50%

Für 2 Personen:

$$\frac{365}{365} * \frac{364}{365} = 0,997 \tag{1.2}$$

Die erste Person kann aus 365 Tagen wählen ohne dass es zu einer Kollision kommt, für die 2. Person bleiben 364 Tage

Für 3 Personen:

$$\frac{365}{365} * \frac{364}{365} * \frac{363}{365} = 0,991 \tag{1.3}$$

Dies wird weitergeführt, bis die Wahrscheinlichkeit für das Gegenereignis bei etwa 50% liegt, somit liegt dann auch das Ereignis, dass mindestens Zwei Personen am selben Tag Geburtstag haben bei etwa 50%. Dieser Fall Tritt bei einer Personen Zahl von 23 ein.

$$\frac{365}{365} * \frac{364}{365} * \frac{363}{365} * ... * \frac{343}{365} = 0,493 \tag{1.4}$$

Die Wahrscheinlichkeit für keine Kollision liegt bei 0,493, somit liegt die Wahrscheinlichkeit für eine Kollision bei 0,507

$$1 - 0,492 = 0,507 \tag{1.5}$$

2 Herleitungen der verwendeten Formeln

Zur Modellierung des Geburtstagsproblems betrachten wir die Zufallsvariable:

$X_n :=$ Zeitpunkt der ersten Kollision bei n Personen mit rein zufällig gewählten Geburtstagen (2.1)

Da zumindest zwei Personen vorhanden sein müssen damit es zu einer Kollision kommt ist der minimale Wert 2. Höchstens sind es $n+1$ Personen. Somit nimmt X_n die Werte $2, 3, ..., (n+1)$ an und es gilt:

$$\mathbb{P}(X_n \geq k+1) = \frac{n * (n-1) * (n-2) * ... * (n-k+1)}{n^k} \tag{2.2}$$

für jedes $k = 1, 2, ..., n+1$. Durch die Annahme der gleichen Verteilung der Zufallsereignisse (Laplace-Modell), ergibt der Zähler von (Equation 2.2) die Anzahl der günstigen Fälle an.

Aus Equation 2.2 folgt durch Verwendung des Gegenereignisses:

$$\mathbb{P}(X_n \geq k) = 1 - \prod_{j=1}^{k-1} (1 - \frac{j}{n}) \tag{2.3}$$

$$\mathbb{P}(Xn \leq 1) = 0 \tag{2.4}$$

Da bei einer einzigen Person $k = 1$ keine Kollision auftreten kann ist die Wahrscheinlichkeit für dieses Ereignis 0 (Equation 2.4). Deshalb ist der Wertebereich für k mit $k = [2; n+1]$ angegeben.

In der Abbildung xxx ist die Wahrscheinlichkeit $P(X_n \leq k)$ durch eine Funktion von k mit dem Parameter $n = 365$ dargestellt. Unterschiedliche Zahlenwerte für n sind in Tabelle xxx aufgeführt. Für das Ereignis $X_n \leq 23$ ist die Wahrscheinlichkeit bereits höher als 50%.

Auf den ersten Blick scheint überraschen das bei $8,5 * 10^{58}$ möglichen Kombinationen (365^{23}) der Geburtstagsverteilung bei 23 Personen. Die Wahrscheinlichkeit eines doppelten Geburtstags schon über 50% liegt.

Die Erklärung hierfür ist das wir auf irgendeine und nicht auf eine bestimmte Kollision warten. Es soll im folgenden gezeigt werden das die erste Kollision bei zufälliger Besetzung von n Tagen von der Größenordnung $\sqrt{n}$ ist.

((((((Warum ist die Größenordnung sqrt(n) hier einfügen))))))

Henze Seite 70 Satz 10.1

Warum verwendung der ungleichung Mittelwertsatz beweis

$$1 - x <= e^{-x} \tag{2.5}$$

2.1 Vereinfachung der Wahrscheinlichkeitsverteilung

Unter Verwendung der Ungleichung

$$1 - x \leq e^{-x} (x \in \mathbb{R}) \tag{2.6}$$

ist es uns möglich die unter Equation 2.3 angegebene Funktion so weit zu vereinfachen, das kein Produkt- oder Summenzeichen mehr vorhanden ist, dadurch sind weitere Berechnungen einfacher zu realisieren.

$$\mathbb{P}(X_n \leq k) \approx 1 - \prod_{j=1}^{k-1} 1 - \frac{j}{n} \geq 1 - exp(-\sum_{j=1}^{k-1} \frac{j}{n}) \tag{2.7}$$

$$\approx 1 - exp(-\frac{k(k+1)}{2n}) \tag{2.8}$$

Bei der ersten Umformung in Equation 2.7 wird das Produkt zu einer Summe im Exponenten von e, aufgrund der allgemeinen Potenzgesetze $a^r * a^s = a^{r+s}$. In der nächsten Umformung wird die im ersten Schritt geschaffene Summe mithilfe der Gausschen Summenformel ersetzt.

Somit ergibt sich zur Berechnung der Wahrscheinlichkeit die allgemeine Formel:

$$\mathbb{P}(X \leq k) \approx 1 - exp(-\frac{k(k+1)}{2n}) \tag{2.9}$$

2.2 Wahrscheinlichkeitsdichte

In der Stochastik beschreibt die Wahrscheinlichkeitsdichte ist eine spezielle reellwertige Funktion zur Konstruktion von Wahrscheinlichkeitsverteilungen.

Im Unterschied zu Wahrscheinlichkeiten kann die Wahrscheinlichkeitsdichte auch größere Werte als 1 annehmen. Dabei wird nicht der Funktionswert sondern die Fläche unterm Funktionsgraphen berechnet, also das Integral.

Die Dichte kann mit zwei verschiedenen Herangehensweisen konstruiert werden: Durch eine Funktion die aus der Wahrscheinlichkeitsverteilung generiert wird, oder durch die Ableitung der Wahrscheinlichkeitsverteilung. Es unterscheidet sich nur die Herangehensweise.

Im Weiteren wird nur noch auf den Fall eingegangen, die in dem die Dichte aus der Wahrscheinlichkeitsverteilung abgeleitet wird.

Allgemein ist die Wahrscheinlichkeitsdichte dann folgendermaßen definiert:

$$\mathbb{P}(]-\infty, a]) = \int_{-\infty}^{a} f(x)\,\mathrm{d}x \tag{2.10}$$

bzw.

$$\mathbb{P}(X \leq a) = \int_{-\infty}^{a} f(x)\,\mathrm{d}x \tag{2.11}$$

Für das Geburtstagsproblem ist die Wahrscheinlichkeitsverteilung aus Formel Equation 2.9 als gegeben anzusehen. Es lässt sich aus der Definition für die Dichte folgende Gleichung ableiten:

$$\mathbb{P}(X \leq k) \approx \int_{-\infty}^{k} f(k)dk \tag{2.12}$$

$$\implies 1 - exp(-\frac{k(k+1)}{2n}) \approx \int_{-\infty}^{k} f(k)dk \tag{2.13}$$

Durch ableiten der Gleichung ergibt sich eine Definition für die Funktion f(k):

$$\implies 1 - exp(-\frac{k(k+1)}{2n}) \approx \int_{-\infty}^{k} f(k)dk \quad | \quad \frac{d}{dk} \tag{2.14}$$

$$\implies \frac{(k-1)}{n} * exp(-\frac{k(k+1)}{2n}) \approx f(k) \tag{2.15}$$

Somit ist die Dichte für das Geburtstagsproblem folgendermaßen definiert:

$$\mathbb{P}(X = k) \approx \frac{(k-1)}{n} * exp(-\frac{k(k+1)}{2n}) \tag{2.16}$$

2.3 Erwartungswert

Der Erwartungswert für das Geburtstagsproblem lässt sich mithilfe der Weibull-Verteilung konstruieren. In dieser Ausarbeitung wird nicht näher auf die Weibull-Verteilung oder die verwendete Gamma-Funktion eingegangen, diese werden als vorausgesetzt angesehen.

Für die um 1 nach rechts verschobene Weibull-Verteilung werden folgende Parameter verwendet. $k = 2$ dadurch ergibt sich eine Rayleigh-Verteilung und $\lambda = \frac{1}{\sqrt{2n}}$. $\mathcal{T}$ bezeichnet die Gammafunktion.

Somit ergibt sich die Funktion:

$$\mathbb{E}(X-1) = \mathbb{E}(X) - 1 \approx \sqrt{2n} * \mathcal{T}(1 + \frac{1}{2}) = \sqrt{2n} * \frac{\sqrt{\pi}}{2} = \sqrt{\frac{\pi * n}{2}} \tag{2.17}$$

$$\mathbb{E}(X) \approx \sqrt{\frac{\pi * n}{2}} + 1 \tag{2.18}$$

für den Erwartungswert.

2.4 Varianz

Die Varianz wird so wie der Erwartungswert auch mithilfe der Weibull-Verteilung bestimmt. Die Parameter sind hierbei die selben wie schon für den Erwartungswert verwendet wurden.

$$Var(X) \approx \frac{1}{\lambda^2} * [\mathcal{T}(1+\frac{2}{k}) - \mathcal{T}^2(1+\frac{1}{k})] \tag{2.19}$$

$$\approx 2n * [\mathcal{T}(1+\frac{2}{2}) - \mathcal{T}^2(1+\frac{1}{2})] \tag{2.20}$$

$$\approx 2n * [1-\frac{\pi}{4}] \tag{2.21}$$

Die Definition der Varianz der Weibull-Verteilung ist in Equation 2.19 gegeben.

2.5 Quantile

Ein Quantil ist ein Lagemaß in der Statistik. Den meisten ist der Median bekannt, dabei handelt es sich um das 50% oder $\frac{1}{2}$ Quantil. Es lassen sich aber auch beliebige Quantile zwischen 0 und 1 bestimmen. Allgemein sind Quantile Schwellenwerte. Werden die gegebenen Daten nach ihrer Wertigkeit sortiert, ist ein bestimmter Anteil kleiner als das Quantil.

Gegeben sei eine beliebige Zufallsvariable X. Dann ist x_p das p-Quantil von X, wenn gilt:

$$\mathbb{P}(X \leq x_p) \geq p \tag{2.22}$$

und

$$\mathbb{P}(x_p \leq X) \geq 1-p \tag{2.23}$$

Im Folgenden wird beschrieben wie aus dieser Definition eine Funktion konstruiert werden kann mit der sich die Quantile für das Geburtstagsparadoxon bestimmen lassen.

Für das $\frac{1}{2}$ Quantil (Median):

$$\begin{aligned}
&\mathbb{P}(X \leq k) = 1 - \mathbb{P}(X > k) \geq 1 - exp(-\frac{k(k-1)}{n}) \stackrel{!}{=} \frac{1}{2} \\
&\Leftrightarrow ln(\frac{1}{2}) = -\frac{k(k-1)}{2n} \\
&\Leftrightarrow -2n * ln(2) = -k(k-1) \\
&\Leftrightarrow 2n * ln(2) = k(k-1) \\
&\Leftrightarrow 2n * ln(2) = k^2 - k \\
&\Leftrightarrow -k^2 + k + 2n * ln(2) = 0 \\
&\Leftrightarrow k^2 - k - 2n * ln(2) = 0 \\
&\Leftrightarrow k = \frac{1}{2} \pm \sqrt{\frac{1}{4} - 2n * ln(2)}
\end{aligned}$$

Daraus lässt sich dann folgendes Ableiten:

$$Q_{\frac{1}{2}}(X) \leq \left(\frac{1}{2} + \sqrt{\frac{1}{4} - 2nln(2)}\right) \leq \left(1 + \sqrt{2nln(2)}\right) \tag{2.24}$$

Das Quantil befindet sich somit in den Grenzen der Quadratischen Funktion $k = \frac{1}{2} \pm \sqrt{\frac{1}{4} - 2nln(2)}$.

Das spezielle $\frac{1}{2}$-Quantil lässt sich auch allgmein bestimmen, sodass der Schwellwert p ein Parameter der Funktion ist:

$$\begin{aligned}
&1 - exp(-\frac{k(k-1)}{n}) \stackrel{!}{=} p \\
&\Leftrightarrow ln(p) = -\frac{k(k-1)}{2n} \\
&\Leftrightarrow 2n * ln(p) = -k(k-1) \\
&\Leftrightarrow 2n * ln(p) = k(k-1) \\
&\Leftrightarrow 2n * ln(p) = k^2 - k \\
&\Leftrightarrow -k^2 + k + 2n * ln(p) = 0 \\
&\Leftrightarrow k^2 - k - 2n * ln(p) = 0 \\
&\Leftrightarrow k = \frac{1}{2} \pm \sqrt{\frac{1}{4} + 2n * ln(p)}
\end{aligned}$$

Daraus lässt sich wie bei der speziellen Lösung, folgende Aussage ableiten:

$$Q_p(X) \leq \left(\frac{1}{2} + \sqrt{\frac{1}{4} - 2nln(p)}\right) \leq \left(1 + \sqrt{2nln(p)}\right) \tag{2.25}$$

Somit lassen sich die Quantile durch folgende Funktionen approximieren:

$$Q_p(X) \approx \frac{1}{2} + \sqrt{\frac{1}{4} - 2nln(p)} \tag{2.26}$$

$$Q_p(X) \approx 1 + \sqrt{2nln(p)} \tag{2.27}$$