Das Bildchensammlerproblem und zufälliges Mischen

Die folgende Ausarbeitung wurde von Mirza Smajic am 04.02.2016 erstellt.

Einleitung

Es sollen in dieser Ausarbeitung zwei verschiedene Probleme vorgestellt und analysiert werden. Zum einen soll das Bildchensammlerproblem untersucht werden und zum anderen wird ein Mischprozess betrachtet, bei der man die oberste Karte eines Kartenstapels zufällig in das Kartendeck hineinmischt und dabei die Anzahl der Mischwiederholungen bestimmen will, bis die Wahrscheinlichkeiten aller möglichen Kartenanordnungen gleichverteilt sind. Wir werden später sehen, dass die beiden Probleme etwas miteinander zu tun haben.

Bildchensammlerproblem

Beim Bildchensammlerproblem gehen wir davon aus, dass es n unterschiedliche Bilder gibt, und wir nacheinander ein zufälliges Bild ziehen und dieses zurück legen. Unser Ziel ist es, jedes Exemplar mindestens einmal zu ziehen. Die Frage die wir hier beantworten ist die, wie häufig wir im Erwartungswert ziehen müssen, bis wir jedes Exemplar mindestens einmal gezogen haben. Im weiteren Verlauf gehen wir davon aus, dass wir nicht Bildchen ziehen, sondern Kugeln.

Beispiel 1. *Wir haben bereits k unterschiedliche Kugeln von insgesammt n unterschiedlichen Kugeln gezogen, wobei $k < n \in \mathbb{N}$ ist. Wir wollen die Wahrscheinlichkeit berechnen, mit der wir nach s Ziehungen genau eine neue Kugel ziehen. Sei X die Zufallsvariable, die die Anzahl der Ziehungen zählt, bis wir eine neue Kugel ziehen. Dann gilt folgende Gleichung:*

$$p(X = s) = \left(\frac{k}{n}\right)^{s-1} \left(1 - \frac{k}{n}\right)$$

Behauptung 1. *Die erwartete Anzahl der Ziehungen für eine neue Kugel erhält man, in dem man über die Anzahl der benötigten Ziehungen aufsummiert, was folgende Formel liefert.*

$$E(neu) = \sum_{s \geq 1} \left(\frac{k}{n}\right)^{s-1} \left(1 - \frac{k}{n}\right) s = \frac{1}{1 - \frac{k}{n}}$$

Beweis.

$$\begin{aligned}
E(neu) &= \sum_{s\geq 1} \left(\frac{k}{n}\right)^{s-1} \left(1-\frac{k}{n}\right) s \\
&= \sum_{s\geq 1} \left(\frac{k}{n}\right)^{s-1} s - \sum_{s\geq 1} \left(\frac{k}{n}\right)^{s} s \\
&= \sum_{s\geq 0} \left(\frac{k}{n}\right)^{s} (s+1) - \sum_{s\geq 0} \left(\frac{k}{n}\right)^{s} s \\
&= \sum_{s\geq 0} \left(\frac{k}{n}\right)^{s} s + \sum_{s\geq 0} \left(\frac{k}{n}\right)^{s} - \sum_{s\geq 0} \left(\frac{k}{n}\right)^{s} s \\
&= \sum_{s\geq 0} \left(\frac{k}{n}\right)^{s} = \frac{1}{1-\frac{k}{n}}
\end{aligned}$$

Wir erhalten eine geometrische Reihe, da $\frac{k}{n} < 1$ ist □

Da der Bildchensammler aber nicht bei k gezogenen Kugeln anfängt, sondern bei 0, müssen wir noch über alle k aufsummieren.

Behauptung 2. *Die erwartete Anzahl von Ziehungen bis wir jede der n verschiedenen Kugeln mindestens 1-Mal gezogen haben ist:*

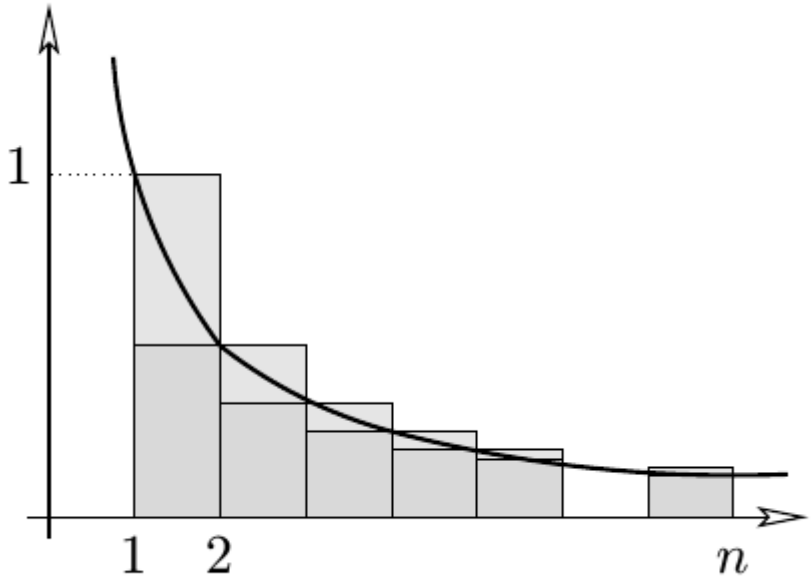
$$n \cdot H_n = n \cdot \sum_{k=1}^{n} k \approx n \cdot ln(n)$$

Beweis.

$$\begin{aligned}
\sum_{k=0}^{n-1} \frac{1}{1-\frac{k}{n}} &= \frac{1}{1} + \frac{1}{1-\frac{1}{n}} + ... + \frac{1}{1-\frac{n-1}{n}} \\
&= \frac{n}{n} + \frac{n}{n-1} + ... + \frac{n}{1} \\
&= n \cdot H_n
\end{aligned}$$

Die Approximation des Ergebnises durch den natürlichen Logarithmus beweisen wir im folgeden Lemma. □

Lemma 1. $H_n \approx ln(n)$



Beweis. Mit der obigen Skizze sind flächengleiche Rechtecke der folgenden veränderten Harmonischen Reihen zu sehen. Mit diesen kann man folgende Ungleichungen erkennen.

$$H_n - 1 = \sum_{k=2}^{n} \frac{1}{k} < \int_1^n \frac{1}{t} dt = ln(n)$$

$$H_n - \frac{1}{n} = \sum_{k=1}^{n-1} \frac{1}{k} > \int_1^n \frac{1}{t} dt = ln(n)$$

$$\rightarrow ln(n) + \frac{1}{n} < H_n < ln(n) + 1$$

woraus die Behauptung folgt. □

Beispiel 2. *Bei 150 Bildchen braucht man im Erwartungswert* $150 \cdot ln(150) \approx 752$ *Ziehungen um jedes Bild mindestens 1-Mal zu ziehen.*

Der folgende Satz zeigt eine Abschätzung nach oben, bei der wir mehr als $n \cdot ln(n)$ Ziehungen brauchen um alle Bilder mindestens 1-Mal zu ziehen.

Satz 1. *Sei* V_n *eine Zufallsvariable, die die Anzahl der Ziehungen zählt bis wir jedes Motiv mindestens 1-Mal gezogen haben. Sei* $m = \lceil nln(n) + cn \rceil$ *die Anzahl der Ziehungen, mit* $n \geq 1, c \geq 0$*. Dann gilt folgende Abschätzung:*

$$p(V_n > m) \leq e^{-c}$$

Beweis. Sei A_i das Ereignis, dass die Kugel i in den ersten m Ziehungen nicht erwischt wird, dann ist:

$$\begin{aligned} p(V_n > m) &= p(\bigcup_i A_i) \leq \sum_i p(A_i) \\ &= n(1 - \frac{1}{n})^m < ne^{\frac{-m}{n}} \\ &\leq ne^{\frac{-nln(n)}{n}} \leq ne^{-ln(n)-c} \\ &= e^{-c} \end{aligned}$$

□

Beispiel 3. *Für $n = 150$ und $c = 2$ erhalten wir $p(V_n > 1052) \leq 0,135$*

Die oberste Karte zufällig hineinmischen

Wir betrachten nun einen Kartenstapel, dessen Startkarten in der Reihenfolge mit $1, ..., n$ bezeichnet werden sollen. Nun betrachten wir einen Mischprozess, indem wir die oberste Karte zufällig in den Kartenstapel hineinmischen. Im folgenden bezeichne σ_n die Menge aller Permutationen der n Karten. Wir wollen solange mischen, bis eine annähernde Gleichverteilung herrscht. Dazu brauchen wir etwas, womit wir Wahrscheinlichkeitsverteilungen messen können. Hierzu definieren wir uns die Variationsdistanz.

Definition 1. *Die Variationsdistanz zwischen zwei Wahrscheinlichkeitsverteilungen Q_1 und Q_2 wird als $||Q_1 - Q_2|| = \frac{1}{2} \sum_{\pi \in \sigma_n} |Q_1(\pi) - Q_2(\pi)|$ erklärt.*

Folgerung 1. *Es gilt $0 \leq ||Q_1 - Q_2|| \leq 1$*

Beispiel 4. *Ohne zu mischen erhalten wir folgende Verteilung $E(id) = 1, E(\pi) = 0$ sonst. Bei der Gleichverteilung gilt $U(\pi) = \frac{1}{n!} \forall \pi \in \sigma_n$*

Beispiel 5. *Seien Q_1 und Q_2 folgende Wahrscheinlichkeitsverteilungen:*

- $Q_1(\pi_1) = \frac{1}{16} \quad Q_2(\pi_1) = \frac{6}{16}$
- $Q_1(\pi_2) = \frac{2}{16} \quad Q_2(\pi_2) = \frac{4}{16}$
- $Q_1(\pi_3) = \frac{5}{16} \quad Q_2(\pi_3) = \frac{2}{16}$
- $Q_1(\pi_4) = \frac{8}{16} \quad Q_2(\pi_4) = \frac{4}{16}$

Dann gilt

$$\begin{aligned} ||Q_1 - Q_2|| &= \frac{1}{2} \left(|\frac{1}{16} - \frac{6}{16}| + |\frac{2}{16} - \frac{4}{16}| + |\frac{5}{16} - \frac{2}{16}| + |\frac{8}{16} - \frac{4}{16}| \right) \\ &= \frac{7}{16} \end{aligned}$$

Satz 2. *Sei $S = \{\pi \in \sigma_n : Q_1(\pi) > Q_2(\pi)\}$, so gilt für die Variationsdistanz*

$$||Q_1 - Q_2|| = |Q_1(S) - Q_2(S)| \quad mit \quad Q_i(S) = \sum_{\pi \in S} Q_i(\pi)$$

Beweis. Sei p_{g1} bzw. p_{k1} die Summe der Wahrscheinlichkeiten der Elementarereignisse, die in Q_1 größer bzw. kleiner sind als in Q_2
Sei p_{g2} bzw. p_{k2} die Summe der Wahrscheinlichkeiten der Elementarereignisse, die in Q_2 größer bzw. kleiner sind als in Q_1

Dann gilt

$$||Q_1 - Q_2|| = \frac{1}{2}\left(|p_{k1} - p_{g2}| + |p_{g1} - p_{k2}|\right) \tag{1}$$

Und aus

- $p_{g1} + p_{k1} = 1$
- $p_{g2} + p_{k2} = 1$

folgt $|p_{g1} - p_{k2}| = |p_{k1} - p_{g2}|$. Wir erhalten damit für die Gleichung (1)

$$\begin{aligned} ||Q_1 - Q_2|| &= \frac{1}{2}\left(|p_{k1} - p_{g2}| + |p_{g1} - p_{k2}|\right) \\ &= \frac{1}{2} \cdot 2\left(|p_{g1} - p_{k2}|\right) \\ &= |p_{g1} - p_{k2}| = |Q_1(S) - Q_2(S)| \end{aligned}$$

□

Von 'zufällig genug gemischt' reden wir dann, wenn die Variationsdistanz zur Gleichverteilung sehr klein ist.

Definition 2. *Top^{*k} beschreibt das k-malige hineinmischen der obersten Karte in das Deck.*

Beispiel 6. *Für die aus Beispiel 4 verwendeten Bezeichnungen erhalten wir folgende Variationsdistanzen*

$$||E - U|| = 1 - \frac{1}{n!} \approx 1$$

$$||Top^{*1} - U|| = 1 - \frac{1}{(n-1)!} \approx 1$$

Definition 3. *Eine Halteregel besagt, nach wie vielen Schritten ein Mischprozess abgebrochen wird, nachdem die daraus resultierenden Permutationen gleichverteilt sind. Sei X_k eine Zufallsvariable die die Reihenfolge der Karten nach k-maligem Mischen annimmt. Eine Halteregel heißt stark gleichverteilt, wenn für alle k folgendes gilt:*

$$p(X_k = \pi | T = k) = \frac{1}{n!} \forall \pi \in \sigma_n$$

Folgerung 2. *Die Halteregel für Top^{*K} ist dann stark gleichverteilt, wenn die ursprüngliche unterste Karte das erste Mal hineingemischt wird.*

Die Begründung hierfür ist sehr trivial. Wenn man Beispielsweise 5 Karten hat, und 3-Mal die oberste Karte zufällig hineinmischt, so muss die vierte karte immernoch vor der fünften Karte sein. Somit sind nich alle Permutationen der 5 Karten möglich. Falls jedoch Karten unter die ursprünglich unterste Karte gelegt werden, so können diese in einer beliebigen Reihenfolge vorliegen. Daher ist die Gleichverteilung erst dann möglich, wenn die ursprünglich unterste Karte, zum ersten Mal hineingemischt wird.

Folgerung 3. *Sei T_i die Zufallsvariable, die zählt, wie oft man mischen muss, bis erstmals i karten unter der Karte n liegen. Wir müssen also die Wahrscheinlichkeitsverteilung von*

$$T_n = T_1 + (T_2 - T_1) + ... + (T_{n-1} - T_{n-2}) + (T - T_{n-1})$$

bestimmen, wobei $T_i - T_{i-1}$ die Zeit ist, bis die oberste Karte an einer der i möglichen Stellen unterhalb von n eingefügt wird. Das ist der selbe Prozess den der Bildchensammler macht, wenn er $(n-i)$ Bilder hat und das $(n-i+1)$te Bild ziehen möchte. Er durchläuft jedoch den Prozess in umgekehrter Reihenfolge.

Folgende Tabelle zeigt die Anzahl der Möglichkeiten für einen Erfolgsschritt beim Bildchensammler, bzw. beim Mischen.

Der k-te Erfolg	Bildchensammler	Mischprozess
1	n	1
2	n-1	2
.	.	.
.	.	.
.	.	.
n	1	n

Folgerung 4. *Somit braucht man $52 \cdot ln(52) \approx 205$ Mischschritte, bis man eine Gleichverteilung des Kartendecks erhält.*

Quellen

Aigner Martin und Günter M. Ziegler - Das Buch der Beweise (2009)