Häufigkeitsmomente

Wir betrachten Datenströme $x_1, x_2, x_3, \ldots$, deren Elemente x_i einem Universum U angehören (z.B.: $U = \{0, \ldots, m-1\}$ für eine feste nat. Zahl m).

Zum Zeitpunkt $n \geq 1$ ist die Häufigkeit (oder Frequenz) eines Elements $u \in U$ im bisher gesehenen Strom $x_1, \ldots, x_n$ der Wert

$$f_u := |\{ i \in \{1, \ldots, n\} : x_i = u \}|.$$

Definition:

Für jede Zahl $k \in \mathbb{N}$ ist das

k-te Häufigkeitsmoment des Stroms $x_1, \ldots, x_n$

definiert als die Zahl

$$F_k := \sum_{u \in U} (f_u)^k.$$

Insbesondere ist

$F_0 = \sum_{u \in U} (f_u)^0$ die Anzahl der verschiedenen Elemente, die im Strom vorkommen (hier wird $0^0 := 0$ und $a^0 := 1$ für alle $a \neq 0$ gesetzt)

$F_1 = \sum_{u \in U} (f_u)^1 = n$ die Länge des bisher gesehenen Stroms

$F_2 = \sum_{u \in U} (f_u)^2$ die "self-join-size" des bisher gesehenen Stroms.

Beispiel:

Für den Datenstrom

3, 1, 3, 3, 2, 1, 5, 2

gilt: $f_1 = 2$, $f_2 = 2$, $f_3 = 3$, $f_4 = 0$, $f_5 = 1$

$F_0 = 4$ (Anzahl verschiedener Elemente)

$F_1 = 8$

$F_2 = f_1^2 + f_2^2 + f_3^2 + f_5^2$

$= 2^2 + 2^2 + 3^2 + 1^2$

$= 8 + 9 + 1$

$= 18$

Die Häufigkeitsmomente liefern wichtige statistische Informationen über die Verteilung der Elemente aus U, die im Datenstrom vorkommen. Insbes. werden die Werte F_0 (und F_2?) z.B. bei der Optimierung von Datenbankanfragen genutzt.

Die Häufigkeitsmomente kann man natürlich leicht bestimmen, indem man für jedes $u \in U$ einen Zähler abspeichert, der besagt, wie oft u im bisher gesehenen Teil des Stroms bereits vorkam. Bei enorm großen U und enorm großen Datenströmen ist dies aber zu aufwändig. Daher wollen wir (unter Verwendung von deutlich weniger Speicherplatz) eine Approximation der Häufigkeitsmomente berechnen.

Der Einfachheit halber sei im Folgenden stets

$$U = \{0, 1, \ldots, m-1\}$$

für eine natürliche Zahl m.

Das 0-te Häufigkeitsmoment:

Wie viele verschiedene Elemente enthält der Strom?

Intuition:

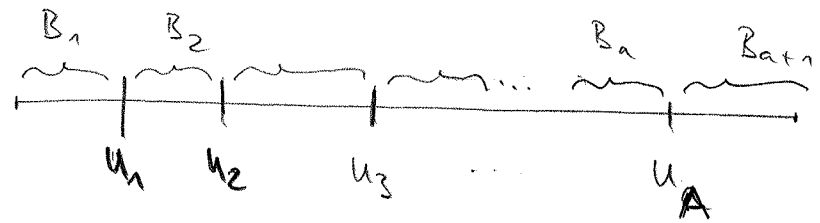
Betrachte zunächst den Fall, in dem die im Datenstrom vorkommenden Elemente zufällig, gleichverteilt aus $U = \{0, \dots, m-1\}$ gewählt wurden.

Sei min das kleinste Element aus U, das im Strom vorkommt.

Anzahl verschiedener Elemente im Strom	durchschnittlicher Wert von min
1	$\approx \frac{m}{2}$
2	$\approx \frac{m}{3}$
A	$\approx \frac{m}{A+1}$

Skizze:

$U = \{0, \dots, m-1\}$:



Verschiedene Elemente im Strom: $u_1 < u_2 < \dots < u_A$

$\Rightarrow$ U wird in $A+1$ verschiedene Blöcke aufgeteilt, jeder Block hat erwartete Größe $\approx \frac{|U|}{A+1} = \frac{m}{A+1}$

Also:

In einem zufällig gewählten Strom mit genau A verschiedenen Elementen (d.h. $F_0 = A$) vermuten wir für das kleinste Element min, das im Strom vorkommt, dass gilt:

$$min \approx \frac{m}{A+1}$$

Somit: $A \approx \frac{m}{min} - 1$ bzw $A \approx \frac{m}{min}$.

Ein erster Algorithmus zum Absätzen von F_0:

- Das Universum $U = \{0, \dots, m-1\}$ ist bekannt
- Beim Lesen des Stroms speichere das kleinste im Strom gesehene Element min
- Gib den Wert $\frac{m}{min} - 1$ aus und behaupte, dass dieser Wert eine Abschätzung für die Anzahl $A = F_0$ der im Strom vorkommenden verschiedenen Elemente ist.

Problem:

Wenn die verschiedenen Datenstrom-Elemente zufällig, gleichverteilt sind, mag zumindest die Intuition stimmen. Aber tatsächliche Datenströme sind in der Regel nicht "zufällig und gleichverteilt".

Lösung:

Nimm eine hinreichend große Zahl M und eine geeignete Familie H von Hash-Funktionen von $U = \{0, 1, \dots, m-1\}$ nach $V := \{0, 1, \dots, M-1\}$.

Wähle zufällig gleichverteilt ein $h \in H$.

An Stelle des Stroms $x_1, x_2, \dots, x_n$

betrachte den Strom $h(x_1), h(x_2), \dots, h(x_n)$,

berechne das kleinste Element min dieses Stroms

und gib als Schätzer für F_0 den Wert

$$F_0^* := \frac{M}{\min'} \quad \text{aus, für} \quad \min' := \begin{cases} 1 & \text{falls } \min = 0 \\ \min & \text{sonst.} \end{cases}$$

Fragen:

(1) Welches M und welches H sollen wir wählen?

(2) Wie nah liegt der Schätzwert F_0^* am tatsächlichen Wert F_0 dran?

Zu (1):

- Wähle eine Zahl $M > 9m$.
- Wähle eine streng 2-universelle Familie H von Hash-Funktionen von U nach $V := \{0, \ldots, M-1\}$.

(Zur Erinnerung:
Wir haben bereits gesehen, dass wenn M eine Primzahl ist, die Menge $H = \{h_{a,b} : a, b \in V\}$ streng 2-universell ist, wobei
$h_{a,b}(x) := a \cdot x + b \pmod M$ ist
$(\forall\, a \cdot + \in U)$.)

Diese Wahl von M und H stellt sicher, dass wir beweisen können, dass Frage (2) wie folgt beantwortet werden kann:

zu (2): Es gilt:

$$Pr\left(\frac{F_0}{9} \leq F_0^* \leq 9 \cdot F_0\right) \geq \frac{2}{3} \qquad (\triangle)$$

d.h. mit Wk. mindestens $\frac{2}{3}$ liegt die Abschätzung F_0^*, die wir ausgeben, zwischen

$\frac{1}{9} \cdot F_0$ und $9 \cdot F_0$,

wobei F_0 die tatsächliche Anzahl der verschiedenen im Strom vorkommenden Elemente ist.

Beweis von $(\triangle)$:

Behauptung 1: $Pr\left(F_0^* > 9 \cdot F_0\right) \leq \frac{2}{9}$

Beweis:

$$Pr\left(\underbrace{F_0^*}_{= \frac{M}{min}\ (\text{bzw } M, \text{falls } min = 0)} > 9 \cdot F_0\right) = Pr\left(min < \frac{M}{9 \cdot F_0}\right)$$

$\frac{M}{min} > 9 \cdot F_0 \ (\Leftrightarrow) \ \frac{M}{9 \cdot F_0} > min$

Sei $A := F_0$ und seien $u_1, \ldots, u_A$ die verschiedenen im Strom $x_1, \ldots, x_n$ vorkommenden Elemente. Dann ist min das kleinste Element in $h(u_1), \ldots, h(u_A)$, wobei h unsere zufällig gewählte Hash-Funktion ist.

Es gilt $\quad A := F_0$

$$\Pr\left(\min < \frac{M}{g \cdot F_0}\right) = \Pr\left(\exists i \in \{1,\dots,A\} : h(u_i) < \frac{M}{g \cdot A}\right)$$

$$\leq \sum_{i=1}^{A} \Pr\left(h(u_i) < \frac{M}{g \cdot A}\right)$$

Aus "H ist streng 2-universell" folgt: dass für zufällig gewähltes h und festes u_i der Wert $h(u_i)$ wie zufällig aus $\{0,1,\dots,M-1\}$ gewählt ist. D.h.: $\Pr(h(u_i)=v) \leq \frac{1}{M}$ für jedes $v \in \{0,\dots,M-1\}$

Und: $h(u_i) < \frac{M}{gA}$ (=) $h(u_i) \in \{0,1,\dots,\lceil \frac{M}{gA} \rceil - 1\}$

$$\leq A \cdot \frac{\lceil \frac{M}{gA} \rceil}{M}$$

$$\leq A \cdot \frac{\frac{M}{gA} + 1}{M}$$

$$= A \cdot \left(\frac{\frac{M}{gA}}{M} + \frac{1}{M}\right) = A \cdot \left(\frac{1}{gA} + \frac{1}{M}\right) = \frac{1}{g} + \frac{A}{M} = \frac{1}{g} + \frac{F_0}{M}.$$

Da jedes im Strom vorkommende Element aus $U = \{0,\dots,m-1\}$ ist, können im Strom nur maximal m verschiedene Elemente vorkommen, d.h.: $F_0 \leq m$.

Daher ist $\frac{F_0}{M} \leq \frac{m}{M} < \frac{1}{g}$, da $M > gm$ ist.

Insgesamt ist also $\Pr\left(\min < \frac{M}{gF_0}\right) \leq \frac{1}{g} + \frac{1}{g} = \frac{2}{g}$

□ Beh 1

Behauptung 2: $\Pr\left(F_0^* < \frac{F_0}{g}\right) \leq \frac{1}{g}$

Beweis: Wegen $F_0^* = \frac{M}{\min}$ und $A := F_0$ gilt:

$$F_0^* < \frac{F_0}{g} \quad (=) \quad \frac{M}{\min} < \frac{A}{g} \quad (=) \quad \min > \frac{gM}{A}.$$

Also: $\Pr\left(F_0^* < \frac{F_0}{g}\right) = \Pr\left(\min > \frac{gM}{A}\right) = \Pr\left(\forall i \in \{1,\dots,A\} : h(u_i) > \frac{gM}{A}\right)$

wobei $u_1,\dots,u_A$ sämtliche verschiedene im Strom vorkommende Elemente sind.

Für $i \in \{1, \dots, A\}$ sei Y_i die Zufallsvariable mit

$$Y_i = \begin{cases} 0 & \text{falls } h(u_i) > \frac{gM}{A} \\ 1 & \text{sonst, d.h. } h(u_i) \leq \frac{gM}{A} \end{cases}$$

und sei $Y := \sum_{i=1}^{A} Y_i$. Es gilt: $Y = 0 \iff \forall i \in \{1, \dots, A\}: h(u_i) > \frac{gM}{A}$

Ziel: Zeige, dass $Pr(Y=0) < \frac{1}{g}$ ist.

Dazu nutzen wir die Tschebyscheff-Ungleichung, die Folgendes besagt:

Tschebyscheff-Ungleichung (über einem endlichen Wahrscheinlichkeitsraum)

Für jede Zufallsvariable X mit Erwartungswert $E(X)$ und Varianz $Var(X)$ gilt für jede reelle Zahl $r > 0$:

$$Pr(|X - E(X)| \geq r) \leq \frac{Var(X)}{r^2}.$$

Zur Erinnerung:

Erwartungswert $E(X) := \sum_i i \cdot Pr(X = i)$

Varianz $Var(X) := E((X - E(X))^2)$

$= E(X^2) - (E(X))^2$ (↖ Linearität des Erwartungswerts)

↑ ein Maß dafür, wie weit die tatsächlichen Werte von X um den Erwartungswert $E(X)$ verstreut sind

Für uns: Je kleiner die Varianz, desto besser!

Es gilt:

- Für jedes $i \in \{1, \dots, A\}$ ist

$$E(Y_i) = \Pr(Y_i = 1) \geq \frac{g}{A}$$

↑ bei der Wahl des Werts $h(u_i)$ gibt es $\geq \frac{gM}{A}$ "gute Werte" bei insgesamt M Werten, nämlich: $0, 1, \dots, \lceil \frac{gM}{A} \rceil - 1$

- $$E(Y) = E\left(\sum_{i=1}^{A} Y_i\right) = \sum_{i=1}^{A} E(Y_i) \geq A \cdot \frac{g}{A}$$

D.h.: $\boxed{E(Y) \geq g}$ (1)

- Da H streng 2-universell ist,

gilt: $$\mathrm{Var}\left(\sum_{i=1}^{A} Y_i\right) = \sum_{i=1}^{A} \mathrm{Var}(Y_{ij})$$

(Beweis: Übung!).

Somit ist $\boxed{\mathrm{Var}(Y) = A \cdot \mathrm{Var}(Y_1)}$

- $\mathrm{Var}(Y_1) = \underbrace{E(Y_1^2)}_{= E(Y_1), \text{ da}} - \underbrace{(E(Y_1))^2}_{\geq 0} \leq E(Y_1)$

Y_1 nur den Wert 0 oder den Wert 1 annehmen kann

- Somit ist $\mathrm{Var}(Y_1) \leq E(Y_1)$ und

$$\mathrm{Var}(Y) = A \cdot \mathrm{Var}(Y_1) \leq A \cdot E(Y_1) = E(Y).$$

Also $\boxed{\mathrm{Var}(Y) \leq E(Y)}$ (2)

- Tschebyscheff liefert für $X := Y$ und $r := E(Y)$:

$$\Pr\left(|Y - E(Y)| \geq E(Y)\right) \leq \frac{\mathrm{Var}(Y)}{E(Y)^2}$$

$$\overset{②}{\leq} \frac{E(Y)}{E(Y)^2}$$

$$= \frac{1}{E(Y)}$$

$$\overset{①}{\leq} \frac{1}{9}$$

- Insgesamt gilt:

$$\Pr\left(F_0^* < \frac{F_0}{9}\right) = \Pr(Y = 0)$$

$$\leq \Pr\left(|Y - E(Y)| \geq E(Y)\right)$$

Für $Y = 0$ ist $|Y - E(Y)| = E(Y)$

$$\leq \frac{1}{9}$$

□ Beh 2

Zusammenfassung:

$$\Pr\left(F_0^* < \frac{F_0}{9} \text{ oder } F_0^* > 9 \cdot F_0\right)$$

$$\leq \Pr\left(F_0^* < \frac{F_0}{9}\right) + \Pr\left(F_0^* > 9 \cdot F_0\right)$$

$$\overset{\text{Beh 2+1}}{\leq} \frac{1}{9} + \frac{2}{9} = \frac{3}{9} = \frac{1}{3}$$

Somit: $\Pr\left(\frac{F_0}{9} \leq F_0^* \leq 9 \cdot F_0\right) \geq 1 - \frac{1}{3} = \frac{2}{3}$.

□ Beweis von ⚠

Insgesamt haben wir gesehen, dass Folgendes gilt:

Satz (Abschätzung von F_0)
Für $m \geq 1$ sei $U := \{0, \dots, m-1\}$
Das folgende Verfahren liefert bei Eingabe eines Stroms $x_1, \dots, x_n$ von Elementen aus U einen Schätzwert F_0^*, für den gilt:

$$\Pr\left(\frac{F_0}{9} \leq F_0^* \leq 9 \cdot F_0\right) \geq \frac{2}{3}$$

und benutzt dazu $O(\log m)$ Speicherbits.

Verfahren:

- Wähle eine Primzahl M mit $9m < M \leq 18m$ (eine solche Primzahl gibt es gemäß Bertrand'schem Postulat).
- Wähle zufällig, gleichverteilt und unabhängig voneinander zwei Zahlen $a, b \in \{0, 1, \dots, M-1\}$ und nutze im Folgenden die Hash-Funktion
  $h: U \longrightarrow \{0, 1, \dots, M-1\}$ mit $h(x) := ax + b \pmod M$
  $(\forall a, x \in U)$
- Beim Lesen des ersten Strom-Elements x_1 initialisiere die Variable min durch $\min := h(x_1)$
- Für jedes $i \geq 1$ tue beim Lesen des $(i+1)$-ten Strom-Elements x_{i+1} Folgendes:
  - berechne $h(x_{i+1})$
  - falls $h(x_{i+1}) < \min$, so setze $\min := h(x_{i+1})$
- Nach dem Lesen des n-ten Strom-Elements x_n gib den Wert $F_0^* := \frac{M}{\min'}$ aus, wobei $\min' := \begin{cases} 1, \text{falls } \min = 0 \\ \min, \text{ sonst.} \end{cases}$

(Quellennachweis: Die Darstellung auf den Seiten 37–46 orientiert sich an dem Abschnitt "Number of distinct elements in a data stream" im Buch "Foundations of Data Science" von Blum, Hopcroft, Kannan (Version: May 14, 2015).)

Anmerkungen zur Abschätzung von F_0:

Mit dem auf Seite 46 vorgestellten Verfahren bekommen wir eine Abschätzung F_0^*, für die gilt:

$$Pr\left(\frac{1}{9}\cdot F_0 \leq F_0^* \leq 9\cdot F_0\right) \geq \frac{2}{3} = 1-\frac{1}{3}$$

Für jede feste Zahl δ mit $0<\delta<1$ erhalten wir durch Anwenden des Median-Tricks (siehe Seite 50) ein Verfahren, das einen Schätzwert $\hat{F}_0$ für F_0 liefert, so dass gilt:

$$\boxed{Pr\left(\frac{1}{9}\cdot F_0 \leq \hat{F}_0 \leq 9\cdot F_0\right) \geq 1-\delta}$$

Dieses Verfahren nutzt Speicherplatz $O(s\cdot \log m)$ für $s = \lceil c\cdot \ln\left(\frac{1}{\delta}\right)\rceil$ mit

$$c = \frac{6}{\left(1-2\cdot\frac{1}{3}\right)^2} = \frac{6}{\left(\frac{1}{3}\right)^2} = \frac{6}{\frac{1}{9}} = 6\cdot 9 = 54.$$

Gut an dem Verfahren ist:

- hohe Erfolgswahrscheinlichkeit: $1-\delta$
- geringer Speicherplatzverbrauch: $O\left(54\cdot \ln\left(\frac{1}{\delta}\right)\cdot \log m\right)$

Schlecht ist:

- Der Schätzwert $\hat{F}_0$ kann noch recht weit vom tatsächlichen Wert F_0 wegliegen.
  An Stelle von "$\frac{1}{9}\cdot F_0 \leq \hat{F}_0 \leq 9\cdot F_0$" hätten wir lieber:
  $(1-\varepsilon)\cdot F_0 \leq \hat{F}_0 \leq (1+\varepsilon)\cdot F_0$ (also: $\hat{F}_0 \in (1\pm\varepsilon)\cdot F_0$).

Lösung:

An Stelle des auf Seite 46 beschriebenen Verfahrens ("Speichere in min den kleinsten Wert $h(x_i)$ für $x_1, \dots, x_n$ und gib den Schätzwert $F_0^* := \frac{M}{\min}$ aus") gehe wie folgt vor:

- Wähle $M := m^3$
- Wähle zufällig, gleichverteilt und unabhängig voneinander zwei Zahlen $a, b \in \{0, 1, \dots, M-1\}$ und nutze im Folgenden die Hash-Funktion
  $h: \mathcal{U} \longrightarrow \{0, 1, \dots, M-1\}$ mit $h(x) = ax + b \pmod M$ (f.a. $x \in \mathcal{U}$)
- Wähle $t := \lceil \frac{96}{\varepsilon^2} \rceil$ und speichere während des Lesens des Stroms $x_1, x_2, \dots$ die t kleinsten paarweise verschiedenen Werte in $h(x_1), h(x_2), \dots$
- Nach dem Lesen des n-ten Strom-Elements x_n betrachte den t-kleinsten gespeicherten Wert $\min_t$
  z.B. für $t=4$ wäre das der viert-kleinste Wert
  und gib den Wert $\tilde{F}_0 := \frac{tM}{\min_t}$ aus.

Speicherverbrauch: $O(\frac{1}{\varepsilon^2} \cdot \log m)$ Bits

Satz (Bar-Yossef et al, 2002)
Für den mit diesem Verfahren berechneten Schätzwert $\tilde{F}_0$ gilt:

$$P_r\left((1-\varepsilon) \cdot F_0 \leq \tilde{F}_0 \leq (1+\varepsilon) F_0 \right) \geq \frac{2}{3} - \frac{1}{m}$$

Hier ohne Beweis — die Details sind ähnlich wie beim Beweis des Satzes auf Seite 46, aber etwas schwieriger. Siehe Bar-Yossef, Jayram, Kumar, Sivakumar, Trevisan "Counting Distinct Elements in a Data Stream", RANDOM 2002.

Durch Anwenden des Median-Tricks (siehe Seite 50) erhalten wir dann ein Verfahren, das einen Schätzwert $\hat{F}_0$ liefert, so dass gilt:

$$\Pr\left(\hat{F}_0 \in (1\pm\varepsilon)\cdot F_0\right) \geq 1-\delta$$

und dabei $O\left(\frac{1}{\varepsilon^2}\cdot \ln\left(\frac{1}{\delta}\right)\cdot \log m\right)$ Speicherbits nutzt:

Für eine geeignete Zahl $s = \Theta\left(\ln\left(\frac{1}{\delta}\right)\right)$ führe unabhängig voneinander s-mal das auf Seite 48 vorgestellte Verfahren durch und gib als $\hat{F}_0$ den Median der s berechneten Schätzwerte aus.

Der Median-Trick

... zum Erhöhen der Erfolgswahrscheinlichkeit auf $1-\delta$, für $0 < \delta < 1$.

Wenn wir ein Verfahren $\mathbb{V}$ haben, das einen Schätzwert F^* für einen Zielwert F berechnet, so dass gilt

$$\Pr\left(aF \leq F^* \leq bF \right) = 1-p,$$

für Werte a, b, p mit $0 \leq a \leq b$ und $0 \leq p < \frac{1}{2}$,

so liefert folgendes Verfahren $\hat{\mathbb{V}}$ einen Schätzwert $\hat{F}$ für F, so dass gilt

$$\Pr\left(aF \leq \hat{F} \leq bF \right) \geq 1-\delta.$$

Verfahren $\hat{\mathbb{V}}$:

- Wähle ein beliebiges $c \geq \frac{6}{(1-2p)^2}$ und ein beliebiges $s \geq c \cdot \ln\left(\frac{1}{\delta}\right)$.
- Führe unabhängig voneinander s-mal das Verfahren $\mathbb{V}$ durch, um s Schätzwerte $F^*_1, F^*_2, \ldots, F^*_s$ zu erhalten
- Berechne den Median $\hat{F}$ der Werte $F^*_1, \ldots, F^*_s$ und gib diesen aus.

Bsp: $\delta = \frac{1}{1.000.000} \Rightarrow \frac{1}{\delta} = 10^6 \Rightarrow \ln\left(\frac{1}{\delta}\right) \approx 13{,}81\ldots$
$p = \frac{1}{3} \Rightarrow (1-2p)^2 = \left(1-\frac{2}{3}\right)^2 = \frac{1}{9} \Rightarrow \frac{6}{(1-2p)^2} = 6 \cdot 9 = 54$
$\Rightarrow$ Wähle $s = 750$

Um zu beweisen, dass tatsächlich

$$\Pr\left(aF \leq \hat{F} \leq bF \right) \geq 1-\delta$$

gilt, wenden wir die Chernoff-Schranke an:

Die Chernoff-Schranke

Satz B.20 (Chernoff-Schranke).
Sei p eine reelle Zahl mit $0 < p < 1$. Wir werfen eine Münze, die mit Wahrscheinlichkeit p auf „Kopf" und mit Wahrscheinlichkeit $1-p$ auf „Zahl" landet. Nach dem i-ten Münzwurf setzen wir

$$X_i := \begin{cases} 1 & \text{falls die Münze beim } i\text{-ten Wurf auf „Kopf" gefallen ist} \\ 0 & \text{sonst.} \end{cases}$$

Wir machen insgesamt $s \in \mathbb{N}_{\geqslant 1}$ Münzwürfe und setzen

$$X := \sum_{i=1}^{s} X_i,$$

d.h. X gibt an, bei wie vielen der s Münzwürfe die Münze auf „Kopf" gefallen ist. Dann gilt:

(a) Der Erwartungswert von X ist $\mathrm{E}(X) = p \cdot s$.

(b) Für jede reelle Zahl ε mit $0 < \varepsilon < 1$ ist

$$\mathrm{P}\Big(X < (1-\varepsilon)\cdot\mathrm{E}(X)\Big) \quad < \quad e^{-\frac{1}{2}\cdot\varepsilon^2\cdot ps}.$$

(c) Für jede reelle Zahl $\beta > 0$ ist

$$\mathrm{P}\Big(X > (1+\beta)\cdot\mathrm{E}(X)\Big) \quad < \quad e^{-\frac{1}{3}\cdot\min(\beta,\beta^2)\cdot ps}.$$

(d) Für jede reelle Zahl ε mit $0 < \varepsilon < 1$ ist

$$\mathrm{P}\Big(X \notin (1\pm\varepsilon)\cdot\mathrm{E}(X)\Big) \quad < \quad 2\cdot e^{-\frac{1}{3}\cdot\varepsilon^2\cdot ps}.$$

Hierbei verwenden wir „$X \notin (1\pm\varepsilon)\cdot\mathrm{E}(X)$" als Kurzschreibweise für

„ $X \;<\; (1-\varepsilon)\cdot\mathrm{E}(X) \quad$ oder $\quad X \;>\; (1+\varepsilon)\cdot\mathrm{E}(X)$ "

Wir betrachten nun folgendes Szenario:
Die bei der Chernoff-Schranke betrachtete Münze fällt beim i-ten Wurf auf "Kopf" (mit Wk p), falls nicht $aT \le F_i^* \le bT$ gilt.

Zur Anwendung der Chernoff-Schranke setzen wir

$$X_i := \begin{cases} 1 & \text{falls nicht } aT \le F_i^* \le bT \text{ gilt} \\ 0 & \text{sonst} \end{cases}$$

und $X := \sum_{i=1}^{s} X_i$. D.h. X gibt an, bei wie vielen der s Durchläufe des Verfahrens V der Schätzwert zu weit vom Fixwert F weg ist.

Gemäß Chernoff-Schranke gilt

(1) $E(X) = p \cdot s$, und

(3) für jede Zahl $\beta > 0$ gilt:

$$Pr\left(X > (1+\beta)E(X)\right) < e^{-\frac{1}{3}\cdot \min(\beta, \beta^2)\cdot p \cdot s} \quad (*)$$

- Wir wollen zeigen, dass $Pr\left(\text{es gilt nicht } aT \le \hat{F} \le bT\right) \le \delta$ ist.
Dazu beachte, dass $\hat{F}$ der Median von $F_1^*, \dots, F_s^*$ ist.
Daher gilt Folgendes:

$aT \le \hat{F} \le bT$ gilt nicht

$\Rightarrow$ für mehr als die Hälfte der Werte $F_1^*, \dots, F_s^*$ ist $F_i^* < aT$
oder
für mehr als die Hälfte der Werte $F_1^*, \dots, F_s^*$ ist $F_i^* > bT$

$\Rightarrow X > \frac{s}{2}$

Somit gilt:

$$\Pr\left(\text{es gilt } \underline{\text{nicht}}\ \alpha T \le \hat{T} \le bT\right) \le \Pr\left(X > \frac{s}{2}\right)$$

Es reicht also, zu zeigen, dass $\Pr\left(X > \frac{s}{2}\right) \le \delta$ ist.

- Wegen $E(X) = p \cdot s$ ist $s = \frac{E(X)}{p}$. Daher ist

$$X > \frac{s}{2} \iff X > \frac{1}{2p} \cdot E(X)$$

- Um $\circledast$ anwenden, wählen wir β so, dass

$$1+\beta = \frac{1}{2p} \quad \text{ist.}$$

D.h. wir wählen $\beta := \frac{1}{2p} - 1$.

Beachte: $\beta > 0$, denn:

$$\frac{1}{2p} - 1 > 0 \iff \frac{1}{2p} > 1 \iff 1 > 2p \iff p < \frac{1}{2},$$ was laut Voraussetzung erfüllt ist.

Gemäß $\circledast$ gilt also:

$$\Pr\left(X > \frac{s}{2}\right) = \Pr\left(X > (1+\beta) \cdot E(X)\right) < e^{-\frac{1}{3} \cdot \min(\beta, \beta^2) \cdot p \cdot s}$$

Es reicht also, zu zeigen, dass $e^{-\frac{1}{3} \cdot \min(\beta, \beta^2) \cdot p \cdot s} \le \delta$ ist.

- Es gilt:

$$e^{-\frac{1}{3} \cdot \min(\beta, \beta^2) \cdot p \cdot s} \le \delta$$

$$\iff -\frac{1}{3} \cdot \min(\beta, \beta^2) \cdot p \cdot s \le \ln(\delta)$$

$$\iff \frac{1}{3} \cdot \min(\beta, \beta^2) \cdot p \cdot s \ge \underbrace{-\ln(\delta)}_{= \ln(\frac{1}{\delta})}$$

$$\iff s \ge \frac{3}{p \cdot \min(\beta, \beta^2)} \cdot \ln\left(\frac{1}{\delta}\right)$$

- Was ist $\frac{3}{p \cdot \min(\beta, \beta^2)}$?

Es gilt:

* $\min(\beta, \beta^2) = \begin{cases} \beta^2, & \text{falls } 0 < \beta < 1 \\ \beta, & \text{falls } \beta \geq 1 \end{cases}$

* $\beta \overset{\text{Def}}{=} \frac{1}{2p} - 1 > 0$, da $p < \frac{1}{2}$ laut Voraussetzung

* $\beta \geq 1 \iff \frac{1}{2p} - 1 \geq 1 \iff \frac{1}{2p} \geq 2 \iff 1 \geq 4p$
  $\iff p \leq \frac{1}{4}$

Somit gilt:

* Falls $p \leq \frac{1}{4}$, so ist
  $\min(\beta, \beta^2) = \beta = \frac{1}{2p} - 1 = \frac{1-2p}{2p}$, und daher
  $p \cdot \min(\beta, \beta^2) = \frac{1-2p}{2}$, und
  $\frac{3}{p \cdot \min(\beta, \beta^2)} = \frac{3 \cdot 2}{1-2p} = \frac{6}{1-2p} \overset{0 \leq 1-2p \leq 1, \text{ da } 0 \leq p \leq \frac{1}{2}}{\leq} \frac{6}{(1-2p)^2} \leq c$

* Falls $p > \frac{1}{4}$, so ist $\beta \geq 1$ und daher
  $\min(\beta, \beta^2) = \beta^2 = \left(\frac{1}{2p} - 1\right)^2 = \left(\frac{1-2p}{2p}\right)^2 = \frac{(1-2p)^2}{4p^2}$,
  und daher
  $p \cdot \min(\beta, \beta^2) = \frac{(1-2p)^2}{4p}$, und
  $\frac{3}{p \cdot \min(\beta, \beta^2)} = \frac{3 \cdot 4p}{(1-2p)^2} = \frac{12p}{(1-2p)^2} \overset{0 \leq p \leq \frac{1}{2}}{\leq} \frac{6}{(1-2p)^2} \leq c$

- Durch Wählen von $s \geq c \cdot \ln\left(\frac{1}{\delta}\right)$ erreichen wir also, dass $s \geq \frac{3}{p \cdot \min(\beta, \beta^2)} \cdot \ln\left(\frac{1}{\delta}\right)$ ist und erhalten somit, dass $e^{-\frac{1}{3} \cdot \min(\beta, \beta^2) \cdot p \cdot s} \leq \delta$ ist. Dies beendet den Beweis dafür, dass der Median-Trick funktioniert. □

Der BJKST - Algorithmus zur Abschätzung von F_0

In der Arbeit "Counting distinct elements in a data stream" bei RANDOM 2002 haben die Autoren Bar-Yossef, Jayram, Kumar, Sivakumar und Trevisan (kurz: BJKST) den folgenden Algorithmus zur Abschätzung von F_0 vorgestellt, der in vielerlei Hinsicht optimal (bzgl Speicherplatzverbrauch und Approximationsgüte) ist.

Zur Darstellung des Algorithmus nutzen wir folgenden Begriff:

Für eine natürliche Zahl n setzen wir

$$\text{Nullen}(n) := \max\{ i : n \text{ ist ein Vielfaches von } 2^i \}$$

D.h.: Nullen(n) ist die Anzahl der Nullen am Ende der Binärdarstellung von n

Bsp:

n	Binärdarstellung von n	Nullen(n)
0	0	1
1	1	0
2	10	1
3	11	0
4	100	2
5	101	0
6	110	1
7	111	0
8	1000	3
16	10000	4

Satz (BJKST-Algorithmus)

Für $m \geq 1$ sei $U := \{0,1,\dots,m-1\}$.
Sei ε mit $0 < \varepsilon < 1$, so dass $\frac{1}{\varepsilon^2} = o(m)$.
Der folgende so genannte BJKST-Algorithmus liefert bei Eingabe eines Stroms $x_1,\dots,x_n$ von Elementen aus U einen Schätzwert F_0^*, für den gilt

$$\Pr\left(F_0^* \in (1 \pm \varepsilon) F_0 \right) \geq \frac{2}{3}$$

und benutzt dazu $O\left(\log m + \frac{1}{\varepsilon^2} \cdot \left(\log\left(\frac{1}{\varepsilon}\right) + \log\log m\right)\right)$ Speicherbits.

BJKST-Algorithmus:

- Wähle zufällig, gleichverteilt und unabhängig voneinander zwei Zahlen $a, b \in \{0,1,\dots,m-1\}$ und nutze im Folgenden die Hash-Funktion $h := h_{a,b}$, d.h.
  $h: U \to U$ mit $h(x) := ax + b \pmod m$ f.a. $x \in U$.
- Setze $M := \left\lceil \frac{4}{\varepsilon^4} \cdot (\log m)^2 \right\rceil$, wähle zufällig, gleichverteilt und unabhängig voneinander zwei Zahlen $a', b' \in \{0,1,\dots,M-1\}$ und nutze im Folgenden die Hash-Funktion $g := g_{a',b'}$ mit
  $g: U \to \{0,1,\dots,M-1\}$ mit $g(x) := a'x + b' \pmod M$ f.a. $x \in U$.
- Setze $N := 0$ und $S := \emptyset$
- Beim Lesen eines Strom-Elements x_i tue Folgendes:
  Falls Nullen $(h(x_i)) \geq N$, dann
  - Füge das Tupel $(g(x_i), \text{Nullen}(h(x_i)))$ in S ein
  - Solange $|S| \geq \frac{600}{\varepsilon^2}$ setze $N := N+1$ und entferne aus S alle Tupel (α, β) mit $\beta < N$.
- Nach dem Lesen des n-ten Strom-Elements gib $F_0^* := |S| \cdot 2^N$ aus.

Wir werden hier keinen formalen Beweis des Satzes angeben, wollen uns aber eine Intuition dazu erarbeiten, wie der Algorithmus funktioniert.

1) Speicherplatzbedarf:

- $\leq 6 \cdot \log m$ Bits zum Speichern von h, g, M, N
- Gemäß Algorithmus ist stets $|S| \leq \frac{60}{\varepsilon^2} + 1$.

  Zum Speichern eines einzelnen Tupels $(g(x_i), \text{Nullen}(h(x_i)))$ in S reichen

  $$\underbrace{\log M}_{\log\left(\frac{4}{\varepsilon^4}\right)} + \log\log m \quad \text{Bits},$$

  also $O\left(\log\left(\frac{1}{\varepsilon}\right) + \log\log m\right)$ Bits.

  Zum Speichern der gesamten Menge S werden also $O\left(\frac{1}{\varepsilon^2} \cdot \left(\log\left(\frac{1}{\varepsilon}\right) + \log\log m\right)\right)$ Bits genutzt.

2) Arbeitsweise:

- Die Hash-Funktion g nutzen wir, um Speicherplatz zu sparen. Man kann nachrechnen, dass g mit hoher Wahrscheinlichkeit kollisionsfrei ist, d.h. $|S| = |S'|$ für

  $S' := \{(x_i, \text{Nullen}(h(x_i))) : (g(x_i), \text{Nullen}(h(x_i))) \in S\}$

  (Beachte dann, dass $|S| \leq \frac{60}{\varepsilon^2} + 1$ und dass insgesamt g auf maximal $\left(\frac{60}{\varepsilon^2} + 1\right) \cdot \log m$ Elemente angewendet wird.)

Die Zahl $M := \frac{4}{\varepsilon^2}(\log m)^2$ ist so gewählt, dass die Wahrscheinlichkeit hoch genug und gleichzeitig der Speicherplatzbedarf möglichst gering ist).

Im Folgenden gehen wir davon aus, dass $|S| = |S'|$ ist.

Warum liefert $|S'| \cdot 2^N$ einen guten Schätzwert für F_0?

* Die Hash-Funktion h wird – wie bei unserem ersten Algorithmus zur Abschätzung von F_0 – genutzt, um die verschiedenen Datenstrom-Elemente möglichst zufällig auf Werte aus $\{0, 1, \ldots, m-1\}$ zu verteilen, d.h. auf Bit-Strings der Länge $\log m$ (indem wir die Binärrepräsentation der Werte betrachten).

* Die Zahl N ist so gewählt, dass S' aus genau denjenigen $(x_i, \text{Nullen}(h(x_i)))$ besteht, für die $\text{Nullen}(h(x_i)) \geq N$ ist. Beachte: S' ist eine Menge und keine Multimenge, d.h. für $x_i = x_j$ mit $i \neq j$ enthält S' das Tupel $(x_i, \text{Nullen}(h(x_i)))$ nur einmal, nicht zweimal.

* Wenn der Strom aus $A := F_0$ verschiedenen Elementen besteht und h diese Elemente zufällig auf Bitstrings verteilt, ist zu erwarten, dass bei etwa $\frac{A}{2^N}$ dieser Bitstrings die letzten N Bits gleich 0 sind. D.h: $|S'| \approx \frac{A}{2^N}$, und somit ist $A \approx |S'| \cdot 2^N$ – und das ist der Schätzwert, den der BJKST-Algorithmus ausgibt.

//

Unter Verwendung des Median-Tricks können wir natürlich wieder die Erfolgswahrscheinlichkeit des BJKST-Algorithmus von $\frac{2}{3}$ auf $1-\delta$ verstärken, indem wir parallel voneinander und unabhängig voneinander $s = O(\ln(\frac{1}{\delta}))$ mal den BJKST-Algorithmus durchführen und den Median der berechneten Schätzwerte ausgeben.