Testdaten mit Reservoir (Sampling)

Author: Philipp G. Freimann (BBW (Berufsbildungsschule Winterthur) https://www.bbw.ch)

Beim Aussuchen von Testdaten aus einer Datenmenge kann man sehr einfach vorgehen. Wir suchen aus einer gegebenen Menge mit n Elementen t Testdaten ($t \leq n$) folgendermaßen: Wir mischen die gegebene Menge und wählen danach die ersten t Elemente aus der gemischten Menge aus.

Ist jedoch n (die gegebene Menge aller Datensätze) so **groß**, dass die Daten nicht im Hauptspeicher Platz finden, oder ist n überhaupt **nicht bekannt**, da die gegebene Datenmenge nur Datensatz um Datensatz eintrifft, so bietet sich folgende **Methode mit Reservoir** an:

Erstellen Sie ein Array mit t Plätzen. Das ist das Reservoir, das zunächst gefüllt werden muss; am Ende enthält das Reservoir die gesuchte Testdatenmenge.

Die ersten t Daten werden einfach ins Reservoir geschrieben (ist $n = t$, so sind wir fertig).

Für jeden weiteren k-ten Datensatz ($t < k \leq n$) wird eine Zufallsposition p zwischen 1 und k (bzw. zwischen 0 und $k-1$, falls das Reservoir ab 0 indexiert ist) erzeugt. Ist $p \leq t$ (bzw. $\leq t - 1$), so wird der Datensatz an Stelle p ins Reservoir geschrieben. Ist p jedoch größer als t (bzw. $t - 1$), so wird der Datensatz ignoriert.

Erklärung: Ist $t = n$, so ist die Sache trivial. Jeder weitere Datensatz ($k = n + 1$) wird nur mit Wahrscheinlichkeit (t/k) ins Reservoir einsortiert. Der Beweis kann leicht über die vollständige Induktion geführt werden. Beweisen Sie zunächst (Induktionsanfang) den Fall $n = t + 1$.