Assignments

빅데이터분석

천세진

Transaction Monitoring Services

- Transaction Monitoring Service는 실시간으로 고객이 수행한 거래에 대해서 모니터링하는 시스템이며, 카드발급회사 및 재정관리 기관에서 사용된다.
- 사기 탐지, 비정상적인 거래(예, 높은 금액) 등을 위해 거래는 실시간으로 분석되거나, 해당 스트림을 대표하는 요약본을 지속적으로 생성해야한다.

거래금액의 데이터 스트림

가능한 분석작업

Data

- 두 가지 형태의 데이터가 제공된다
  - 1000개의 연습용 데이터셋: 숫자로 구성 (trial_amount.csv)
  - 1M (백만개)의 테스트용 데이터셋: 시험 채점용 (test_amount.csv)

- 각 숫자는 거래금액을 의미한다
  - 예로 50은 50$

Questions

- Transaction Monitoring Service과 앞서 언급한 데이터가
  주어졌을때, 5가지 문제 중 4문제를 선정하여 코드를 구현하시오
  - 분석작업#1: Reservoir Sampling 구현
  - 분석작업#2: 고유한 거래 금액의 개수 추정 (Approximation)
  - 분석작업#3: 거래금액의 중간값(Median) 추정(Approximation)
  - 분석작업#4: 거래금액에서 자주 등장하는 금액 추정 (Approximation)
  - 분석작업#5: 계층적 샘플링

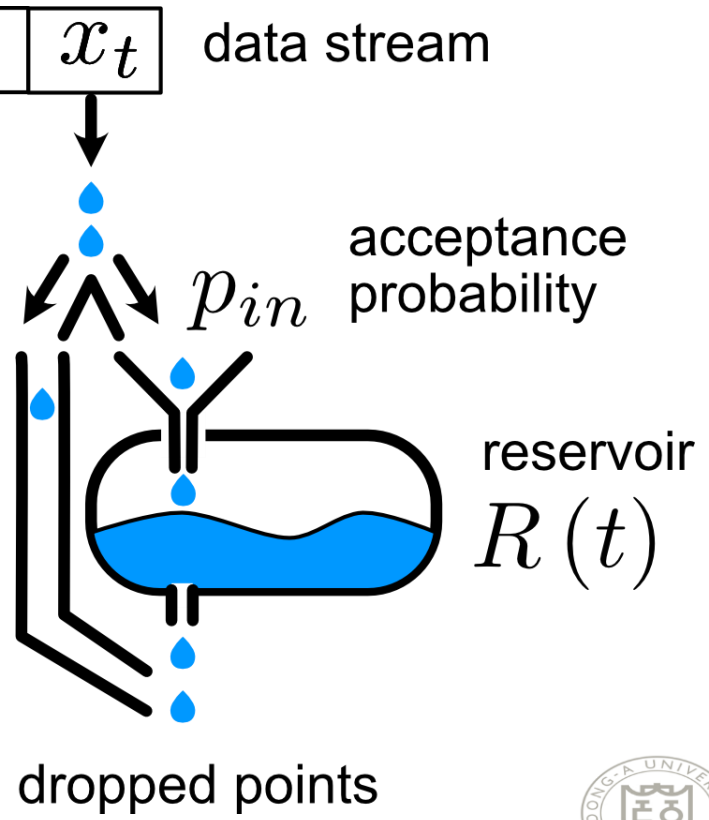
- 마감일: ~6월 15일까지 (600점, 문제당 150점)

Template Code & How to submit

- 문제에 대한 Template Code 가 주어진다
- 각 분석작업에 대응되는 코드는 주어진 템플릿 내에 #TODO# 내에서 이루어져야한다
  - def task1AReservoirSampling
  - def task2BDistinctAmount
  - def task3CMedian
  - def task4DMostFreqAmount
- 코드 작성 후에 파일명을 변경하여 가상대학에 업로드
  - BDA4_lastname_id.py
  - 예시) BDA4_Chun_202200001.py
- 코드 내 학생이름 및 학번은 필히 기입하여 제출
  - 코드에 대한 주석 필수

분석작업#1: Reservoir Sampling

- 거래금액의 데이터스트림으로부터 Naïve reservoir sampling 방법을 적용하여라
  - Reservoir의 크기는 100으로 수행할 것
  - 최종 Reservoir 내 값은 무작위임



분석작업#2: 고유한 거래금액 (Distinct transaction amounts)의 갯수

- 거래금액의 데이터스트림으로부터 고유한 거래금액의 개수를 추정(Approximation)해야한다
  - Flajolet-Martin 알고리즘을 사용해야함
  - 적절한 수의 Hash function을 결정하여라
    - 메모리에는 100개의 요소까지 관리가능하기 때문

분석작업#3: 거래금액의 중간값을 추정

- 거래금액의 데이터스트림으로부터 거래금액의 중간값(Median)을 추정(Approximation)해야한다
- 중간값은 거래금액의 50%가 m보다 작고 나머지 50%는 m보다 크다는 것을 의미한다
- 주어진 데이터는, 파레토 유형1 분포를 따른다고 가정한다.
  - 수식 1은 CDF(Cumulative distribution function): 임의의 값이 x보다 작을 확률에 대해 계산, 최소값은 1

$$P(X < x) = 1 - \left(\frac{1}{x}\right)^{\alpha}$$

<수식 1: CDF>

표준 분포 norm에서 임의의 값이 1.96보다 작을 확률은
norm.cdf(1.96) => 0.975 # 97%수준
1 - norm.cdf(1.96)은 1.96보다 클 확률로 계산

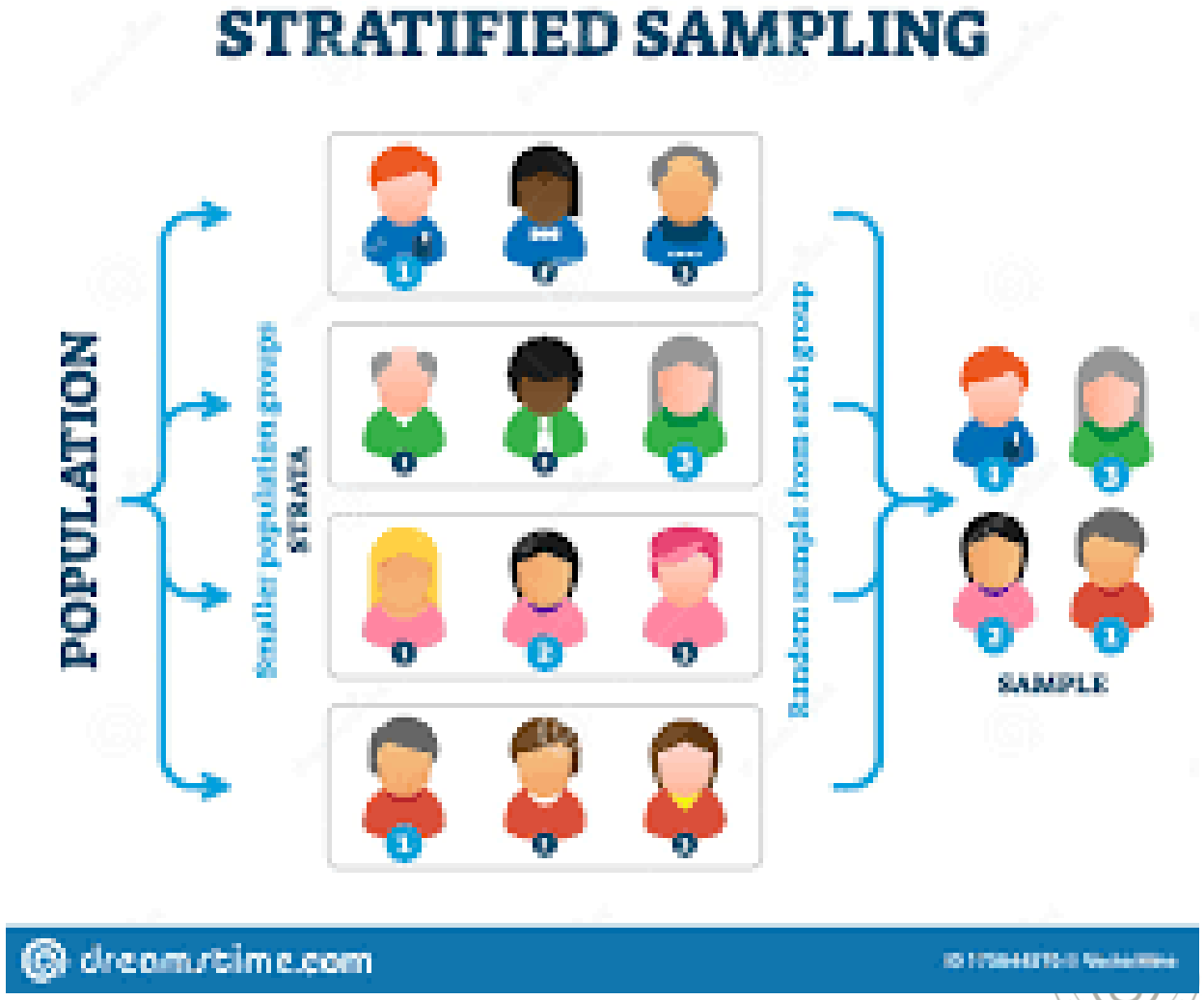
** 파레토 추정에 대한 추가 코드 참고

분석작업#4: 자주 등장하는 거래금액의 값(최빈값)을 추정

- 거래금액의 데이터스트림으로부터 자주 등장하는 거래금액을 찾으시오
  - 일반적으로, 파레토 분포(Pareto distribution)에서, 최빈값은 가장 작은 가능성을 값는 값이다. 하지만, 실세계에서는 이론적인 분포를 따르는 경우는 매우 적다. 따라서, 이번 데이터에 대해 해당 분포에 대한 확률을 적용하는 것은 어렵다.
  - 이번 분석작업은 여러분의 최빈값을 위한 여러분만의 추정 방법을 제시하는 것이 목표이다.
    - 답을 정확하게 찾는 것도 중요하지만, 제한된 메모리 내에서 추정하기 위한 다양한 아이디어를 제시하기를 바란다.
    - 가능한 Single-Pass내 이루어지도록 한다.

BONUS 분석작업: 계층적 샘플링(Stratified Sampling)

- Strata(Stratum의 복수)는 데이터의 작은 그룹을 표현할때 사용한다.
- Stratified sampling은 주어진 데이터로부터 목적/특징에 따른 그룹으로 분리하게 위해 사용되며, Stratified Random Sampling(SRS)이라고도 불리운다.
  - 또한, 분리된 그룹은 모집단의 분포를 따른다(Follow)는 점이 매우 중요하다



BONUS 분석작업: 계층적 샘플링(Stratified Sampling)

- 거래금액의 데이터셋을 k개의 구간으로 나누었다고 가정하자.
- 나누어진 모집단 데이터(Original data)로부터 Train과 Test 데이터셋을 생성하여라
  - 두 데이터셋의 서로 분포가 거의 동일하여야 한다.
- 본 작업은 Single-pass로 꼭 구현할 필요는 없습니다.

- 참고
  - http://www.gisdeveloper.co.kr/?p=9891
  - https://github.com/vikotse/Reservoir-Sampling