1024550143S 杨茜雅 DaSE算法 1^{st}

习题2 P28-29

1、 1~160 灯泡

$k=\frac{N}{n}=\frac{160}{20}=8$ 直线等距抽样

设第一次抽的是第r个个件，以r为起初位置

r	r+8	r+8×2	…	r+8×15=126
1	2	3		16

r=6

∴起始抽取编号为6

4、 $k=\frac{N}{n}=\frac{14}{4}=3.5$

∴采用圆形等距抽样 $k=\lfloor 3.5\rfloor=3$

r=1 为起始位置

∴	r+3×1	r+3×2	r+3×3
	4	7	10

∴其余3个样本编号为4、7、10

5、采用分层抽样

样本容量与总体容量之比 100:500=1:5

A: $\frac{280}{5}=56$ B: $\frac{95}{5}=19$ C: $\frac{125}{5}=25$

A抽取56件 B抽取19件 C抽取25件

8、分层抽样 按比例分配法

20:500=1:25 O: $\frac{200}{25}=8$ A: $\frac{125}{25}=5$ B: $\frac{125}{25}=5$

AB: $\frac{50}{25}=2$ ∴O型8人，A型5人，B型5人，AB型2人

13、 ___ ___ ___ ___ ___ ___ ... N

采用水库抽样算法

该链表可被看作一个数据规模大小未知的数据流。

只能遍历链表一次 → 即数据只能被访问一次

每个元素被抽中的概率相等

用水库抽样算法

①先将链表中的前k个元素保留下来，构建一个大小为k的水库

1	2	3	4	······	k-1	k

1 2 3 ··—·· ···k··· ··· N

②对于第m条记录（m>k），以 $\frac{k}{m}$ 的概率决定是否由这条记录替换水库中的一条记录

③ 循环这步

④ 直到遍历链表

当读到第i个元素时，产生1~i的均匀随机数r，if $r \le k$，则当前元素替换水库中第r个元素，发生概率为 $\frac{k}{i}$

证明：等概率且为 $\frac{k}{N}$　① $N=k$ 时　$prob=\frac{k}{k}=1$

② $N=m\ (>k)$ 时，假设 $prob=\frac{k}{m}$

③ 当 $N=m+1\ (>k)$ 时　$prob=P$(替换不发生)$+P$(替换但不替换t，t是一直保留在水库中的元素)

P(替换不发生)$=1-\frac{k}{m+1}$　　k个数中选到t

P(替换但不替换t)$=\frac{k}{m+1}\times(1-P$(替换t)$)=\frac{k}{m+1}\times(1-\frac{1}{k})$

∴ P(t一直保留)$=\frac{k}{m}\times\left(\frac{k}{m+1}\times\frac{k-1}{k}+\frac{m+1-k}{m+1}\right)=\frac{k}{m+1}$

14、 __ __ $\underset{k}{\_\_}$ | __ __ __ __

假设每个元素被赋予的随机数 是 ，无重复的，则是等概率抽样。

证明：

① 当 $N=k$ 时，则选取最大的 k 个数时（Top-k），则会选到所有数，$prob=1=\frac{k}{N}$

② 当 $N=m\ (m>k)$ 时，假设，水库中任意一条记录被抽取到的概率为 $\frac{k}{m}$

③ 假设记录 t 是水库中的记录，在上一轮抽样中，它以 $\frac{k}{m}$ 的概率保存在水库中，当 $N=m+1$ 时，t 留下的情况有二

a、第 $m+1$ 条记录被赋予的随机数不在 top-k 个中

$m \quad m+1$ 　 $prob=1-\frac{k}{m+1}$

b、第 $m+1$ 条记录被赋予的随机数在 top-k 个中但不替代 t

$\frac{k}{m+1}\left(1-\frac{1}{k}\right)$

$$\therefore prob=\frac{k}{m}\times\left(\frac{m+1-k}{m+1}+\frac{k}{m+1}\times\frac{k-1}{k}\right)$$

$$=\frac{k}{m}\times\frac{m}{m+1}=\frac{k}{m+1}$$

为等概率抽样

1	2	3	…	t	t+1	…	m	m+1
1	5	3	6	12	11	---	random	

第 top k

$m+1$ 个数

$(m+1)$ 被赋予的随机数不在 top-k 个之间的 prob $1-\frac{k}{m+1}$

在 top-k 个之间的 prob $\frac{k}{m+1}$ 且正好取代 t，则证明 t 当初被赋予的随机数是 top-k 中最小的 $prob=\frac{1}{k}$

$$\therefore \frac{k}{m+1}\left(1-\frac{1}{k}\right)$$