温兆和 10205501432. 数据科学算法作业1.

1. 解. 抽样间距为: 160/20 = 8
设起始编号为x. 则 x + 15 × 8 = 126.
解得. x = 6.

4. 解: 14/4 = 3.5 故抽样间距为4.
第1个: 1. 第2个: 1+4 = 5. 第3个: 5+4 = 9 第4个: 9+4 = 13

5. 解: 应使用分层抽样法. 由样本/总体 = 100/500 = 1/5.
则A车间以系统抽样抽取 280 × 1/5 = 56 样本.
B车间以系统抽样抽取 95 × 1/5 = 19 样本.
C车间以系统抽样抽取 125 × 1/5 = 25 样本.

8解: 由于要研究血型与色弱的关系, 最好抽取的样本中每种血型人数相等. 且若采用等比分配法, 会造成AB型血样本过少. 故最好采用分层抽样法+等额分配法. 从四种血型中各随机抽取 20/4 = 5 样本.

13. 解: 应该使用水库抽样法. 维护一个大小为k的数组, 先把链表中遍历到的前k个元素全部放入数组中. 此后, 对于被遍历到的第m个元素 (m > k). 在1~m中随机等概率地生成一个随机数x. 若x ≤ k. 就把数组中的第x个元素替换成当前被遍历到的元素; 否则就把当前遍历到的元素丢弃. 如此循环往复. 直到链表中的所有元素都被遍历过一次.

14. 证：这种抽样算法等价于：对每个元素赋予一个随机数后，对这批总体使用水库排序算法。前k个元素直接进入数组；当m>k时对遍历到的第m个元素，若其随机数大于该数组中元素的随机数的最小值，就把这个带有"k个元素中的最小值"的元素从数组中替换掉。否则，它就被丢弃。现在研究当m>k时，遍历到的第m个元素的值大于数组中元素的概率是否为 $\frac{k}{m}$.

对第k+1个元素：前k个元素按随机数从小到大排，分隔出(k+1)个缝：

① □ ② □ ... □ (k+1).

元素 I_{k+1} 要留在样本中，只能落入缝① ~ 缝(k+1)，$p=\frac{k}{k+1}$

其余元素要留在样本中：则不能成为最小值，$p=\frac{k}{k+1}$

假设，前m个元素想留在样本，$p=\frac{k}{m}$，对元素 I_{m+1}：

I_{m+1} 欲留在样本中：则要落在前m个元素之间最右边的k个缝里，$p=\frac{k}{m+1}$.

其余元素欲留在样本：①、它是前k大元素中最小者，且未被替换 ②、其已留在样本中且非最小者.

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

由数学归纳法知，到进第m个元素时，前m个元素留在样本中概率均为 $\frac{k}{m}$.

故：该算法为等概率抽样.