1. 用系统抽样法从 160 个灯泡中抽取一个容量为 20 的样本。若将这 160 个灯泡编号为 $1 \sim 160$ ，若在第 16 个被抽中的个体编号为 126 ，起始抽取编号为多少?

解 抽样间距：$k = \frac{N}{n} = \frac{160}{20} = 8,$

设起始抽样编号为 r, 则: 其余抽样编号为: $r\ r{+}k\ r{+}2k\ \ldots, r{+}15k, \ldots,$

因此第 16 个为：$r + 15k = 126$

所以起始抽样编号为：$r = \frac{126}{15*8} = 6$

2. 为了解某市参加物理竞赛的 1000 名学生的成绩，应采用什么抽样方法比较恰当，简述抽样过程。

解 法一：因为需要了解的是参加物理竞赛的 1000 名学生的成绩，因此总体的容量知道且容量大。因此可选用系统抽样。

设抽取样本的容量为 n，则：

(1) 给这 1000 名学生编号为 1 1000，

(2) 先计算抽样间距：$k = \frac{N}{n} = \frac{1000}{n},$

(3) 在 1 k 的号码中随机抽取一个编号，假设为 r，则其余编号为：$r\ r + k\ r + 2k\ \ldots$

法二：直接采用简单随机采样

3. 某工厂每小时可生产零件 10000 个，每天的生成时间为 12 小时，为了检测产品的合格率，质检员每天需要抽取 1200 个零件进行检测，请设计一个合理的抽样方案。若每天抽取 980 个进行检测呢?

解 因为该工厂每个小时可生产 10000 个零件，每天生成的时间为 12 小时，所以一天总共能生产 120000 个零件。

法一：如果需要抽取 1200 个零件，那么可以采用分层抽样的方法：

每个小时需要抽取：$\frac{1200}{12} = 100$ 个；

然后对每个小时可以采用系统抽样，对每个小时生产的 10000 个零件进行编号，计算抽样间距，$k = \frac{N}{n} = \frac{10000}{100} = 100,$ 在 1 100 间的号码中随机选择一个，然后每隔 100 个抽取一个零件加入样本中。

将每层抽取到的样本合在一起，即为所抽取的样本。(也可直接采用直线等距抽样)

法二：在每小时的 10000 个零件中随机抽取 100 个。

如果需要抽取 980 个零件，可以在每个小时的 10000 个零件中随机抽

取 82 个（或者采用系统采样），一天之后一共有 984 个样本，生成 4 个在 $[1, 984]$ 间的随机整数，将值为 1 的样本删除。

4. 设总体有 14 个个体，并按照 $1 \sim 14$ 进行编号。欲要以系统抽样法抽取容量为 $n = 4$ 的样本，且第一个抽中的编号为 1 ，则其余 3 个样本编号依次为多少?

解　由题意可以，采取系统抽样抽取样本；又因抽样间距 k 不为整数，因此采用圆形等距抽样。

$k = \frac{N}{n} = 3.5$，取与之最近的整数 k=4，因为第一个编号为 1，则其余三个编号为 1+4=5，5+4=9，9+4=13。

5. 已知 A，B，C 三个车间一天内生产的产品分别是 280 件、95 件、125 件，为了掌握各车间产品质量情况，需从中抽取一个样本容量为 100 的样本，应该如何抽取?

解　由题意可知，产品已分为三层，由 A，B，C 车间生产，因此采用分层抽样。

样本容量和总体容量之比为 $100\ 280 + 95 + 125 = 1 : 5$

利用抽样比确定每个车间应该抽取的产品数,依次为:$280/5 = 56\ 95/5 = 19\ 125/5 = 25$,

最后，利用简单随机抽样或系统抽样方法，从三个车间中分别抽取 56，19，25 个产品，合在一起后即为所抽取的样本。

6. 从词库中随机等概率抽取 100 个单词，请写出水库抽样的主要步骤。若被告知词库容量为 10000，请设计合理的抽样方法抽取 100 个单词组成的样本。

解　水库抽样步骤:

(1) 首先将单词流中的前 100 个单词保留下来，构建一个大小为 100 的水库；

(2) 对于单词流中的第 m 个单词（m>100），以 $\frac{100}{m}$ 的概率决定是否由这个单词替换掉水库中的一个单词；

(3) 循环步骤 (2) 直到处理完所有数据。

如果被告知词库的容量为 10000，那么可以使用系统抽样进行抽样（总体中个体数较多）。

7. 某市有甲、乙两个地区，现要进行家庭收入的调查。已知，甲地区有 40000 户居民，乙地区有 60000 户居民。并且，甲乙两地区的居民收入标准差估计分别为 2000 和 4000 。同时，对甲乙两地区每户的平均抽样费用之比为 1 : 3 。若市政府想要抽取容量为 1000 的样本，请分别给出奈曼分配法和经济分配法的样本容量分配方案。

解　奈曼分配法：$n_i = n * \frac{N_i S_i}{\sum_i N_i S_i}$

由公式可知：$n_{\text{甲}} = 1000 * \frac{40000*2000}{40000*2000+60000*4000} = 250$

$n_{\text{乙}} = 1000 - 250 = 750$

经济分配法：$n_i = n * \frac{N_i S_i / C_i}{\sum_{i=1}^{K} N_i S_i / C_i}$

由公式可知：$n_{\text{甲}} = 1000 * \frac{40000*2000/1}{40000*2000/1+60000*4000/3} = 500$

$n_{\text{乙}} = 1000 - 500 = 500$

8. 某校 500 名学生中，有 200 人的血型为 O 型，有 125 人的血型为 A 型，有 125 人的血型为 B 型，有 50 人的血型为 AB 型。为了研究血型与色弱的关系，需从中抽取一个容量为 20 的样本，怎样抽取样本比较合适？请写出具体抽样过程。

解　因为总体由差异明显的不同血型组成，因此采用分层抽样；

法一：样本容量和总体容量之比为 $20\ 500 = 1 : 25$

利用抽样比确定每个血型应该抽取的人数,依次为:$200/25 = 8\ 125/25 = 5\ 125/25 = 5\ 50/25 = 2$;

最后，利用简单随机抽样或系统抽样方法，从四个血型中分别抽取 8，5，5，2 个学生，合在一起后即为所抽取的样本。

法二：因为需要研究血型和色弱的关系，可以采用分层抽样中的等额样本，则：$\frac{20}{4} = 5$;

因此利用简单随机抽样或系统抽样方法，从四个血型中分别抽取 5 个学生，合在一起后即为所抽取的样本。

9. 一个地区共有 5 个乡镇，共 3 万人。其人口比例内 $3 : 2 : 5 : 2 : 3$，从这 3 万人中抽取一个 300 人的样本，分析某种疾病的发病率。已知这种疾病与不同的地理位置及水土有关，则应采取什么样的抽样方法？并写出具体的抽样过程。

解　因为该疾病与不同的地理位置及水土有关，因此采用分层抽样；

法一：每个乡镇的人口占比为：$3:2:5:2:3$；

利用抽样比确定每个乡镇应该抽取的人数，依次为：$300*\frac{3}{3+2+5+2+3}=60$, 同理：$300*\frac{2}{15}=40, 300*\frac{5}{15}=100, 300*\frac{2}{15}=40, 300*\frac{3}{15}=60$;

最后，利用简单随机抽样或系统抽样方法，从五个乡镇中分别抽取 60，40，100，40，60 人，合在一起后即为所抽取的样本。

法二：且需要 5 个乡镇和疾病的关系，因此每个乡镇应该分别抽取 $\frac{300}{5}=60$ 人；

再利用简单随机抽样或系统抽样方法，从 5 个乡镇中分别抽取 60 人，合在一起后即为所抽取的样本。

10. 某村委调查本村各户收入情况所作的抽样阅读并回答问题:

本村人口：1200 人，户数 300，每户平均人口数 4 人

应抽户数：30 户

抽样间隔：$\frac{1200}{30}=40$

确定随机数字：取一张人民币，编码的后两位数为 12

确定第一样本户：编码的后两位数为 12 的户为第一样本户；

确定第二样本户：$12+40=52$，52 号为第二样本户

(1) 该村委采用了何种抽样方法?

(2) 抽样过程中存在哪些问题？并修改。

(3) 何处用到了简单随机抽样?

解 (1) 直线等距抽样

(2) 抽样间隔应该用总户数除以样本户数，为 $\frac{300}{30}=10$, 那么第一样本户的编号应该在 1 10 之间，第二户应该为第一户加 10.

(3) 确定第一样本户时用到了随机采样。

11. 给定一个数组，按照如下方式抽样 k 个元素：考虑第一个元素，其以 $\frac{k}{N}$ 的概率被选中；如果该节点被选中，则从剩下的 $(N-1)$ 个元素中选出 $(k-1)$ 个元素；如果没有被选中，则从剩下的 $(N-1)$ 个元素中选出 k 个元素；$\cdots$，依次这样做下去，直到获取到 k 个元素为止。请问这是一个均匀抽样方法吗（每个元素被抽中的概率相等）?

解 这不是一个均匀抽样方法:

假设已经遍历了 m 个元素 ($0 \le m \le N$), 其中共抽取了 i 个样本 ($0 \le i \le k$), 则对于第 $m+1$ 个元素，其被选择的概率为 $p = \frac{k-i}{N-m}$，

若为均匀抽样，则有：$\frac{k-i}{N-m} = \frac{k}{N}$, 但该等式并不恒成立。因此这不是一个均匀抽样。

12. 给定一个文件，在不知道文件总行数的情况下，如何从文件中随机的抽取若干行?

解　在不知道总行数的情况下，可以使用水库抽样，选定需要抽取的行数 n，前 n 行直接放入水库，第 k 行（k>n）以概率 $\frac{n}{k}$ 的概率替换水库中的记录。

13. Google 曾经有一道非常经典的面试题：给定一个长度为 N 的链表，其中 N 很大，而且不知道确切知道 N 的大小。你的任务是从这 N 个元素中随机取出 k 个元素，当只能遍历这个链表一次，而且必须保证取出的元素恰好是 k 个，且它们是完全随机的（每个元素被抽中的概率相等)?

解　采样水库抽样就可实现题目中的要求：

(1) 首先将长度为 N 的链表中的前 k 个元素保留下来，构建一个大小为 k 的水库；

(2) 对于链表中的第 m 个元素（m>k)，以 $\frac{k}{m}$ 的概率决定是否由这个元素替换掉水库中的一个元素；

(3) 循环步骤 (2) 直到链表末尾。

14. 假设有一个大小为 N 的数组，其中未知 N 的值，依次扫描该数组的每个元素，为每个元素赋予一个随机数，然后使用 Top-k（譬如最大 k 个数）得到需要的 k 个元素。请问该方法是一种等概率抽样方法吗?

解　该方法是一个等概率抽样方法；对于前 k 个元素，会被选入 Top-k，概率为 $\frac{k}{k} = 1$;

设随机数范围为 $[a, b]$, 则 $Top-k$ 的范围为 $[a + \frac{k}{m}(b-a), b]$, 则随机数在 $Top-k$ 区间的概率为：

$$\frac{b - (a + \frac{k}{m}(b-a))}{b-a} = \frac{k}{m}$$

对于第 m 个元素 (m>k)，对它赋予随机数，那么它有 $\frac{k}{m}$ 的概率被被选入前 k 个数，每个元素被抽取的概率相等。

15. 在水库抽样算法中，当第 i 个元素到达时，以 $\frac{1}{i}$ 的概率替换水库中选定的某个元素，直到最后一个元素到达为止，请证明每个元素被选中的概率相等，即为 $\frac{1}{n}$。

解　使用数学归纳法证明如下:

(1) 当 $n=1$ 时，水库中的任意一条记录被抽到的概率为 $\frac{1}{1}=1$;

(2) 假设当 $n=m(m>1)$ 时，水库中任意一条记录被抽取到的概率为 $\frac{1}{m}$;

(3) 假设记录 t 是水库中的记录，在上一轮抽样中，它以 $\frac{1}{m}$ 的概率保留在水库中。

当 n=m+1 时, 元素 t 在下面这种情况下会继续保存在水库中: 第 m+1 条记录不会替代水库中的记录;

概率为: $\frac{1}{m}(1-\frac{1}{m+1})=\frac{1}{m+1}$;

因此，根据归纳假设，对于长度为 n 的数据，水库抽样能保证每条记录以 $\frac{1}{n}$ 的概率保留在水库中。