计算题A

1. 某淡水鱼养殖户想了解一下鱼塘中虹鳟鱼的数量，拟采用抽样的方法进行估计，该户先捕捞到200条虹鳟鱼，做上记号后放回鱼塘，数星期后进行再一次的捕捞，共捕捞到100条虹鳟鱼，发现其中有32条鱼带有记号。试估计该鱼塘大约有多少条虹鳟鱼。
   解： 已知n1=200，n=100, r=32

$$\hat{N} = \frac{n_1 n}{r} = \frac{200 \times 100}{32} = 625$$

$$v(\hat{N}) = \frac{n_1^2 n(n-r)}{r^3} = \frac{200^2 \times 100(100-32)}{32^3} = 8300.78$$

$$\alpha = 0.05, t = 1.96$$

$$\Delta = t\sqrt{v(\hat{N})} = 1.96 \times \sqrt{8300.78} \approx 1.96 \times 91.1 \approx 179$$

因此总体单元数N的95%置信区间为625$\pm$179，即446~804

2. 为估计市区人均居住面积，按与各区人数呈比例的概率从12个区中抽了4个区，经调查的数据如下：

样本区号	区居住面积（ ）	人口数
1	2835326	604746
2	1670996	456035
3	1835226	470981
4	2895058	585257

试对市区人均居住面积作点估计和置信度为95%的区间估计。
解：已知N=12，n=4,

$$\bar{y_1} = \frac{2835326}{604746} = 4.688$$

$$\bar{y_2} = \frac{1670996}{456035} = 3.664$$

$$\bar{y_3} = \frac{1835226}{470981} = 3.896$$

$$\bar{y_4} = \frac{2895058}{585257} = 4.946$$

人均居住面积的点估计为

$$\hat{\bar{Y}}_{pps} = \frac{1}{4}\sum_{i=1}^{4} \bar{Y_i} = 4.299$$

方差为

$$V(\hat{\bar{Y}}_{PPS}) = \frac{1}{4}\frac{\sum_{i=1}^{4}(\bar{Y_i} - \hat{\bar{Y}}_{PPS})^2}{n-1} \approx 0.0947$$

在95%的置信度下，t=1.96

$$\Delta = 1.96\sqrt{0.0947} = 0.603$$

因此人均居住面积的95%置信区间为4.299 ± 0.603,即3.69~4.89

3. 要调查学生对某课程的兴趣问题，将问题陈述为“我对该课程感兴趣”和“我对该课程不感兴趣”，对此问题采用沃纳模型处理，预先设定P=4/5,在接受调查并作出明确回答的320人中（假定被调查者如实回答问题），结果统计出回答“是”的人数为156人，请估计对该课程感兴趣学生比例的置信区间。
   解：采用沃纳模型，已知P=0.8, 样本量n=320，回答是的人数n1=156, 则有

$$\hat{\pi_A} = \frac{1}{2P-1}[\frac{n_1}{n} - (1-P)] = 0.477$$

$$v(\hat{\pi_A}) = \frac{\hat{\pi_A}(1-\hat{\pi_A})}{n} + \frac{P(1-P)}{n(2P-1)^2} = 0.0021$$

$$\Delta = 1.96\sqrt{0.0021} = 0.089$$

因此对该课程感兴趣学生比例的95%置信区间为0.477 ± 0.089

4. 某社会心理学研究机构要了解有过偷盗行为的人数比例，考虑到该问题的特殊性，准备借助敏感性调查的方法开展，采用西蒙斯模型处理，调查所用的两张卡片中问题分别为：
   卡片A：我曾经有过偷盗行为；
   卡片B：我的身份证号码尾号为奇数。
   卡片A、B在卡片总数中的比例各为1/2，且已知身份证号码尾号为奇数的人数比例为1/2。一共调查了500人，结果回答“是”的人数为256，试估计人群中有过偷盗行为的人数比例和估计误差。
   解： 设具有卡片A特征的比例为πA,具有卡片B特征的比例为πB，已知样本量n=500, 回答为是的人数为n1=256，P=0.5, πB=0.5则有

$$\hat{\pi_A} = \frac{\frac{n1}{n} - (1-P)\pi_B}{P} = 0.524$$

$$v(\hat{\pi_A}) = \frac{1}{(n-1)P^2}\frac{n_1}{n}(1-\frac{n_1}{n}) = 0.00199$$

$$\Delta = 1.96\sqrt{0.00199} = 0.087$$

因此人群中有过偷到行为的人数比例的95%置信区间为0.524 ± 0.087

5. 从一个大的总体调查二种疾病的发病率，一种疾病的发病率约为50%，另一种疾病的发病率约为1%，
   - 若二者都要求估计标准误为0.05，应各调查多少人?
   - 若二者都要求达到相同的差异系数（vcp）为0.05，各应调查多少人?
     解： 已知两种病的比例估计分别为

$$P_1 = 0.5, P_2 = 0.01$$

     一、 由绝对误差定义$\Delta = t\sqrt{v(p)}$ 相对误差$r = \Delta/p$ 得到

$$\Delta = 2 \times 0.05 = 0.1$$

$$r = 0.1/0.01 = 10$$

$$n_1 = \frac{t^2}{\Delta^2}PQ = \frac{4 \times 0.5 \times 0.5}{0.1^2} = 100$$

$$n_2 = \frac{t^2}{\Delta^2}PQ = \frac{4 \times 0.01 \times 0.99}{0.1^2} \approx 4$$

二、由变异系数定义$C = \frac{S}{P}$得到

$$S_1 = C_1 P_1 = 0.05 \times 0.5 = 0.025$$

$$S_2 = C_2 P_2 = 0.05 \times 0.01 = 0.0005$$

由绝对误差计算公式

$$\Delta = t\sqrt{v(p)}$$

得

$$\Delta_1 = 0.05, \Delta_2 = 0.001$$

根据样本量计算公式

$$n_0 = \frac{t^2}{d^2}PQ$$

得

$$n_1 = \frac{4}{0.05^2} \times 0.05^2 = 400$$

$$n_2 = \frac{4}{0.001^2} \times 0.01 \times 0.99 = 39600$$

总体单元数很大，n接近于n0

6. 某镇在2000户家庭中随机抽选36户家庭调查生活费用支出，以 表示食物支出费用， 表示总支出费用，得恩格尔系数（食物支出在总支出中所占的比例），r=y/x=41.7% ，y与x的样本变异系数分别是cy=0.09,cx=0.085，y与x的相关系数$\hat{\rho}$=0.79，给定置信度95%，求恩格尔系数的区间估计.
解：已知N=2000, n=36, f=n/N=0.018
恩格尔系数点估计

$$\hat{R} = \bar{y}/\bar{x} = 41.7$$

样本变异系数

$$C = \frac{\sqrt{v(\bar{\theta})}}{\bar{\theta}}$$

$$v(\bar{\theta}) = C^2\bar{\theta}^2$$

由样本均值方差的估计

$$v(\bar{\theta}) = \frac{1-f}{n}s_\theta^2$$

得到

$$s_\theta = \sqrt{\frac{nv(\bar{\theta})}{1-f}}$$

根据样本相关系数

$$\hat{\rho} = \frac{s_{yx}}{s_x s_y}$$

可得

$$s_{yx} = \hat{\rho} s_x s_y = \hat{\rho} \sqrt{\frac{nC_x^2 \bar{x}^2}{1-f}} \sqrt{\frac{nC_y^2 \bar{y}^2}{1-f}} = \hat{\rho} \frac{nC_x \bar{x} C_y \bar{y}}{1-f}$$

根据V(R)估计公式

$$v(\hat{R}) \approx \frac{1-f}{n\bar{X}^2}(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy})$$

将以上代入 重新得到

$$v(\hat{R}) \approx \hat{R^2}(C_y^2 + C_x^2 - 2\hat{\rho} C_x C_y) = 0.000563$$

$$\Delta = 1.96 \times \sqrt{0.000563} \approx 0.0465$$

因此恩格尔系数的95%置信区间为41.7%±4.65%

7. 某企业有工人132人，技术人员92人，管理人员27人。现欲通过抽样调查估计去年全年平均每人请假天数，拟采用分层抽样。若已知工人请假天数的总体方差为36，技术人员的方差为25，管理人员的方差为9，设样本量为30，试用内曼分配确定各层的样本量。
   解: 已知N=132+92+27=251, n=30, N1=132, N2=92, N3=27 得到各层层权

$$w_1 = 132/251 = 0.52, w_2 = 92/251 = 0.37, w_3 = 27/251 = 0.11$$

$$w_1 s_1 = 0.52 \times 36 = 18.72$$

$$w_2 s_2 = 0.37 \times 25 = 9.25$$

$$w_3 s_3 = 0.11 \times 9 = 0.99$$

$$\sum_{h=1}^{3} W_h S_h = 18.72 + 9.25 + 0.99 = 28.96$$

根据奈曼分配样本量计算公式

$$n1 = n\frac{W_1 S_1}{\sum_{h=1}^{3} W_h S_h} = 30 \times \frac{18.72}{28.96} = 19.39$$

$$n2 = n\frac{W_2 S_2}{\sum_{h=1}^{3} W_h S_h} = 30 \times \frac{9.25}{28.96} = 9.58$$

$$n3 = n\frac{W_3 S_3}{\sum_{h=1}^{3} W_h S_h} = 30 \times \frac{0.99}{28.96} = 1.03$$

即在工人层抽19个，技术人员抽10个，管理人员抽1个

8. 某淡水鱼养殖户想了解一下鱼塘中虹鳟鱼的数量，拟采用抽样的方法进行估计，该户先捕捞到200条虹鳟鱼，做上记号后放回鱼塘，数星期后进行再一次的捕捞，要求捕捞到带有记号的鱼30条，结果捕捞到120条后才实现目标，试估计该鱼塘大约有多少条虹鳟鱼。
   解：已知n1=200，n=120, r=30

$$\hat{N}=\frac{n_1 n}{r}=\frac{200\times 100}{32}=625$$

$$v(\hat{N})=\frac{n_1^2 n(n-r)}{r^2(r+1)}=\frac{200^2\times 120(120-30)}{30^2(30+1)}=15483.87$$

$$\alpha=0.05, t=1.96$$

$$\Delta=t\sqrt{v(\hat{N})}=1.96\times\sqrt{15483.87}\approx 1.96\times 124.43\approx 243.89$$

因此总体单元数N的95%置信区间为625$\pm$243.89

9. 欲估计某校初一学生每周用于英语课程时间占学习总时间的比重，该校初一年级共有学生102名，随机抽取了10学生，记录了他们每周英语学习时间（Y）和总学习时间（X），并得出如下数据：$\bar{x}=40$，$\bar{y}=14$，$s_x=4.78$，$s_y=3.09$，$s_xy=13$。试估计该校初一学生每周用于英语课程时间占学习总时间的比重及其标准差。
   解：已知总体总量N=102，样本量n=10, $\bar{x}=40$，$\bar{y}=14$，$s_x=4.78$，$s_y=3.09$，$s_xy=13$ 得到

$$\hat{R}=\frac{\bar{y}}{\bar{x}}=14/40=0.35$$

$$v(\hat{R})\approx\frac{1-f}{n\bar{X}^2}(s_y^2+\hat{R}^2 s_x^2-2\hat{R}s_{xy})\approx 0.000183$$

$$\sqrt{v(\hat{R})}\approx 0.0135$$

计算B

1．某住宅区调查居民的用水情况，该区共有 N=1000户，调查了 n=100户，得 $\bar{y}$=12.5吨， s^2=1252，有40户用水超过了规定的标准。 要求计算：

1. 该住宅区总的用水量及95%的置信区间；
2. 若要求估计的相对误差不超过10%，应抽多少户作为样本?
3. 以95%的可靠性估计超过用水标准的户数；
4. 若认为估计用水超标户的置信区间过宽，要求缩短一半应抽多少户作为样本?
   解:

- 由已知得f=n/N=0.1, 总用水量的点估计：

$$Y=N\bar{Y}=1000\times 12.5=12500$$

$$v(Y)=N^2V(\bar{y})=\frac{N^2(1-f)}{n}s^2=11268000$$

$$\Delta=1.96\times\sqrt{11268000}\approx 6579.297$$

因此该住宅区总用水量95%置信区间的估计为12500$\pm$6579.297

- 已知$r\leq 0.1$, 因为总体均值和方差未知，利用案例中给出的样本量为100户的简单随机样本的均值和方差替代，得到

$$n_0=\frac{t^2s^2}{r^2\bar{y}^2}=1.96^2 1252^2/0.1^2 12.5^2\approx 3078.2$$

  在无放回条件下的样本量

$$n=\frac{n_0}{1+\frac{n_0}{N}}\approx 755$$

- 该住宅区超过用水标准的用户比例的无偏估计

$$P=40\%$$

比例估计的方差为

$$v(p)=\frac{1-f}{n-1}p(1-p)=2.1818\times 10^-3$$

$$\Delta=1.96\times\sqrt{2.1818\times 10^-3}=9.16\%$$

则该小区超过用水标准的比例在95%的置信区间下的估计为 $40\%\pm 9.16$
超过用水标准的户数为$100\times(40\pm 9.16)\%$

- 估计用水超过标准的户数的置信区间缩短一半，就意味着估计估计误差缩短一半，则

$$\Delta'=\frac{1}{2}\Delta=4.508\%$$

$$n_0=\frac{t^2}{\Delta'^2}PQ=460.99\approx 461$$

在无放回条件下的样本量

$$n=\frac{n_0}{1+\frac{n_0-1}{N}}\approx 315.753$$

2. 从某农村的200户中随机等概率（无放回）抽取50户，发现其中8户有自行车，这8户人数分别为3，5，3，4，7，4，4，5人。根据这一资料要求：

    1. 估计该村具有自行车的户数及其估计精度；
    2. 估计该村具有自行车的总人数及其估计精度。

    解：已知N=200, n=50, f=n/N=50/200=0.25

    - 拥有自行车户数的比例为

    $$P=8/50=0.16$$

    比例估计的方差为

    $$v(p)=\frac{1-f}{n-1}p(1-p)=2.057\times 10^-3$$

    $$\Delta=1.96\sqrt{2.057\times 10^-3}=0.0889$$

    所有拥有自行车户数的比例在95%置信区间下的估计为0.16 ± 0.0889 拥有自行车的户数的区间估计为$200\times(0.16\pm 0.0889)$

    - 由已知可得样本每户人数平均值

    $$\bar{y}_D=\frac{3+5+3+4+7+4+4+5}{8}=4.375$$

    样本均值的方差

$$s_D^2=\frac{\sum_{i=1}^{8}(y_i-\bar{y}_D)^2}{n_D-1}=1.6964$$

拥有自行车的总人数的点估计

$$\hat{Y}=4.375\times 200\times 8/50=140$$

总人数估计量的方差为

$$v(\hat{Y}_D)=\frac{N(N-n)}{n}s_d^2=\frac{N(N-n)}{n}(\frac{n_D-1}{n-1}s_D^2+\frac{n}{n-1}p_Dq_D\bar{y}_D^2)\approx 1720.4057$$

$$\Delta=1.96\sqrt{1720.4057}\approx 81.3$$

因此，该村拥有自行车人数的置信区间为140 ± 81.3

3. 某城市共有1000家餐馆，分为小中大三层，现预估计在餐馆就餐的人数，采用抽样调查，根据以往资料

层	N_h	S_n^2
中	300	2500
小	600	400
大	100	1000

1. 若欲估计就餐总人数的误差不超过4000人，可靠性为95%，采用内曼分配应抽多少家餐馆作为样本（假设每层每户的调查费用相等）；
2. 若不按比例抽样在数据上比较复杂，其费用相当于调查50家餐馆，因此从效益上看改为按比例抽样是否值得?

解:

- 已知N=1000, N1=300, N2=600, N3=100 得到各层层权

$$W_1=300/1000=0.3, W_2=600/1000=0.6, W_3=100/1000=0.1$$

$$S_1=\sqrt{2500}, S_2=\sqrt{400}, S_3=\sqrt{1000}$$

$$W_1S_1=0.3\times\sqrt{2500}=15$$

$$W_2S_2=0.6\times\sqrt{400}=12$$

$$W_3S_3=0.1\times\sqrt{1000}=3.16$$

$$\sum_{h=1}^{3}W_hS_h=15+12+3.16=30.16$$

根据条件可算出平均每家餐馆就餐人数的允许误差为 $\Delta=4000/1000=4$

$$V(\bar{y})=\Delta^2/t^2=4.1649$$

根据奈曼分配的样本量公式

$$n=\frac{(\sum_{h=1}^{3}W_hS_h)^2}{V+\frac{\sum_{h=1}^{3}W_hS_h^2}{N}}=173.1256\approx 174$$

所以，假设每层每户的调查费用相等的条件下，若欲估计就餐总人数的误差不超过4000 人，可靠性为95%，采用最优分配应抽174 家餐馆作为样本。

- 根据要求，样本的分配方式为比例分配，则有放回抽样条件下的样本量为：

$$n_0 = \frac{\sum_{h=1}^{3} W_h S_h^2}{V} = 261.709$$

无放回修正后的样本量为

$$n = \frac{n_0}{1 + \frac{n_0}{n}} = 207.42 \approx 208$$

即，在相同条件下，按照比例分配抽样调查时所需的样本是208 家。但是，案例中要求若不按照比例分配样本，则在调查数据方面比较复杂，其费用相当于调查50 家餐馆。那么，奈曼分配要达到与比例分配相同的精度，费用相当于调查174 + 50= 223 家餐馆，大于比例分配需要调查的208家，因此从经济效益上看，改为按比例分配样本是较为值得的。

4. 为了解某小区住户的平均月支出（单位：元），在7000户家庭中按不放回简单随机抽样抽出200户进行调查，并得到样本均值$\bar{y} = 1800$，样本方差$s^2 = 640000$。 （1）试估计该小区住户的平均月支出，并给出95%置信度下的区间估计。 （2）若要求估计的相对误差不超过10%，则需抽出多少户家庭进行调查?
解：

- 已知$N = 7000, n = 200, f = 200/7000 = 0.0286$

$$v(\bar{y}) = \frac{1-f}{n} s^2 = 3107.2$$

$$\Delta = 1.96\sqrt{3107.2} = 109.25$$

因此该小区平均月支出在95%置信区间下的估计为1800 ± 109.25

- 根据要求$r \leq 10\%$，因为总体均值 和方差 未知，利用案例中给出的样本量为200的简单随机样本的均值和方差替代，因为该样本的均值和方差均是总体均值和方差的无偏估计。有放回抽样条件下所需要的样本量为：

$$n_0 = \frac{t^2 s^2}{r^2 \bar{y}^2} = 75.883$$

在无放回抽样下样本量

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = 76$$

5. 有下列数据

层	W_h	$\bar{y_h}$	s_h	P_h
1	0.35	3.1	2	0.54
2	0.55	3.9	3.3	0.39
3	0.1	7.8	11.3	0.24

设n=1000

1. 采用按比例分层抽样的方法估计$\bar{Y}$和P并计算其标准误；
2. 采用内曼分配的方法估计$\bar{Y}$和P并计算标准误；
3. 将按比例分配和内曼分配与简单随机抽样相比能提高效率多少

   解：在本案例中，假设总体单元众多，则忽略有限总体修正系数以及含有总体单元倒数$\frac{1}{N}$的项

   - $$\bar{y}_{prop}=\sum_{h=1}^{3}W_h\bar{y}_h=4.01$$

     $$V(\bar{y}_{prop})=\frac{1}{n}\sum_{h=1}^{3}W_hs_h^2=0.02$$

     $$p_{prop}=\frac{1}{n}\sum_{h=1}^{3}p_h=0.39$$

     $$V(p_{prop})=\frac{1}{n}\sum_{h=1}^{3}W_hP_hQ_h=2.36\times10^{-4}$$

   - 按奈曼分层的抽样方法，各层样本量

     $$n_h=n\frac{W_hS_h}{\sum_{h=1}^{3}W_hS_h}W_hS_h$$

     , 有

     $$n_1\approx192,n_2\approx498,n_3\approx310$$

于是，奈曼分配抽样的参数估计

$$\bar{y}_{min}=\sum_{h=1}^{3}\frac{n_h}{n}\bar{y}_n=4.96$$

$$V(\bar{y}_{min})=\frac{1}{n}(\sum_{h=1}^{3}W_hs_h)^2\approx0.031$$

$$P_{min}=\sum_{h=1}^{3}\frac{n_h}{n}p_n=0.3732$$

$$V(p_{min})=\frac{1}{n}\sum_{h=1}^{3}W_h^2s_h^2=\frac{1}{n}(\sum_{h=1}^{3}W_h\sqrt{p_h(1-P_h)})^2\approx2.22\times10^{-4}$$

- a. 对于比例分配而言，同样样本量的简单随机抽样对总体均值估计的方差：

$$V(\bar{y}_{srs})=\frac{1}{n}[\sum_{h=1}^{3}W_hs_h^2-\sum_{h=1}^{3}\frac{W_hs_h^2}{n_h}+\sum_{h=1}^{3}W_h\bar{y}_h^2-\bar{y}_{prop}^2+v(\bar{y}_{prop})]\approx0.022$$

估計效率：

$$deff=\frac{V(\bar{y}_{prop})}{V(\bar{y}_srs)}=\frac{0.020}{0.022}\approx0.91$$

同样样本量的简单随机抽样对总体比例估计的方差：

$$V(p_{srs})=\frac{1}{n}[\sum_{h=1}^{3}W_hp_hq_h-\sum_{h=1}^{3}\frac{W_hp_hq_h}{n_h}+\sum_{h=1}^{3}W_hp_h^2-p_{prop}^2+v(p_{prop})]\approx2.75\times10^{-4}$$

估計效率：

$$deff = \frac{V(p_{prop})}{V(p_{srs})} = \frac{2.36 \times 10^{-4}}{2.75 \times 10^{-4}} \approx 0.86$$

b. 对于奈曼分配而言，同样样本量的简单随机抽样对总体均值估计的方差：

$$V(\bar{y}_{srs}) = \frac{1}{n}[\sum_{h=1}^{3} W_h s_h^2 - \sum_{h=1}^{3} \frac{W_h s_h^2}{n_h} + \sum_{h=1}^{3} W_h \bar{y}_h^2 - \bar{y}_{min}^2 + v(\bar{y}_{min})] \approx 0.049$$

估計效率：

$$deff = \frac{V(\bar{y}_{min})}{V(\bar{y}_{srs})} = \frac{0.031}{0.049} \approx 0.63$$

同样样本量的简单随机抽样对总体比例估计的方差：

$$V(p_{srs}) = \frac{1}{n}[\sum_{h=1}^{3} W_h p_h q_h - \sum_{h=1}^{3} \frac{W_h p_h q_h}{n_h} + \sum_{h=1}^{3} W_h p_h^2 - p_{min}^2 + v(p_{min})] \approx 2.33 \times 10^{-4}$$

估計效率：

$$deff = \frac{V(p_{min})}{V(p_{srs})} = \frac{2.22 \times 10^{-4}}{2.33 \times 10^{-4}} \approx 0.95$$

6. 有下列數據

层	N_h	S_h	$\bar{Y}_h$
1	60	2	3
2	30	4	5
3	10	15	12

现令$n = 40$，要求

1. 样本在各层中进行的按比例分配估计量方差；
2. 样本在各层中进行的内曼分配估计量方差；
3. 计算内曼分配较按比例分配的得益；

解： 已知N=100,

$$W_1 = 0.6$$

$$W_2 = 0.3$$

$$W_3 = 0.1$$

1. 按比例分配，则有：

$$n_1 = W_1 n = 24$$

$$n_2 = W_2 n = 12$$

$$n_3 = W_3 n = 4$$

2. 内曼分配，则有

$$n_1 = n\frac{W_1S_1}{\sum_{h=1}^{3}W_hS_h} = 12$$

$$n_2 = n\frac{W_2S_2}{\sum_{h=1}^{3}W_hS_h} = 12$$

$$n_3 = n\frac{W_3S_3}{\sum_{h=1}^{3}W_hS_h} = 16$$

但是，由于N3=10<16, 因此取n3=10.待分配的样本量n-n3=30在第一二层重新分配：

$$n_1 = n\frac{W_1S_1}{\sum_{h=1}^{2}W_hS_h} = 15$$

$$n_2 = 30 - n_1 = 15$$

即：样本在各层中进行的最优分配，使得三层的样本量分别为15、15、10。

3. 按比例分配的方差：

$$V(\bar{y}_{prop}) = \frac{1-f}{n}\sum_{h=1}^{3}W_hS_h^2 = 0.4455$$

按照奈曼分配的方差：

$$V(\bar{y}_{min}) = \frac{1}{n}(\sum_{h=1}^{2}W_hS_h)^2 - \frac{1}{N}\sum_{h=1}^{2}W_hS_h^2 \approx 0.12$$

因此，按照最优分配（奈曼分配）较按比例分配的进行抽样的得益：

$$V(\bar{y}_{prop}) - V(\bar{y}_{min}) = 0.4455 - 0.12 = 0.36$$

7. 将总体分为三层，采用分层随机抽样方法每层抽取6个，得到如下资料：

层	各单元标志值
1	2 4 5 5 6 8
2	8 8 10 14 14 18
3	16 16 16 18 22 26

1. 按比例分配抽样方式计算$\bar{y}_{st}$及$v(\bar{y}_{st})$；
2. 计算$Deff$因子；
3. 若达到以上同样的精度采用简单随机抽样的样本量应为多少?

解：

1. 由已知可计算出下表的数值

层	W_h	$y\bar{h}$	s_h^2	s
1	1/3	5	4	2
2	1/3	12	16	4
3	1/3	19	17.2	4.1473

已知 L=3，N=18,

$$\bar{y}_{prop} = \sum_{h=1}^{3} W_h \bar{y}_h = \frac{1}{n}\sum_{i=1}^{18} y_i = 12$$

$$V(\bar{y}_{prop}) = \frac{1-f}{n}\sum_{h=1}^{3} W_h S_h^2 = 0.6888$$

2. 简单随机抽样的方差为：1-f忽略不计

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}) = 45.52941$$

$$V(\bar{y}_{srs}) = \frac{S^2}{n} = 2.53$$

$$Deff = \frac{0.6888}{2.53} = 0.272$$

3. 按要求若达到以上分层抽样的精度采用有放回简单随机抽样的样本量n0 应为：

$$n_0 = \frac{n}{Deff} = 66$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = 14.13$$

计算C

1. 调查某条街的居民居住条件，从该街道的100个居民小组随机抽取了8个居民小组，取得以下数据

样本居民小组	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
居民数	40	39	12	55	37	33	41	14
房间数	58	72	26	98	74	57	76	48

要求：

1. 估计平均每个居民拥有的房间数并计算估计精度；
2. 估计该条街共有多少房间及其估计的精度； 说明你上述使用的估计量是有偏的还是无偏的。

解： 1. 已知N=100, n=8, 设X代表居民数，Y代表房间数。平均每个居民拥有的房间数$R = \frac{Y}{X}$, 得到：

$$f = \frac{n}{N} = 0.08$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 33.5$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 63.625$$

$$\hat{R} = \frac{\bar{Y}}{\bar{X}} = \frac{\sum_{i=1}^{8} y_i}{\sum_{i=1}^{8} X_i} = 1.8993$$

因为：

$$S_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=189.4286$$

$$S_y^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2=463.9821$$

$$S_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=270.2143$$

$$v(\hat{R})\approx\frac{1-f}{n\bar{X}^2}(s_y^2+\hat{R}^2s_x^2-2\hat{R}s_{xy})=0.01239$$

$$\Delta=1.96\sqrt{0.01239}=0.22$$

因此平均每个居民拥有的房间数在95%置信区间下的估计为1.8993$\pm$0.22
由于采用了比例估计， 所以该估计量有偏，渐进无偏，当样本量较大时，估计量的偏倚趋向于零。

2. 该街道房间总数的点估计为：

$$\hat{Y}_{srs}=N\bar{y}=100\times63.625\approx6363$$

方差为：

$$V(\hat{Y})=N^2\frac{1-f}{n}S^2=533579.415$$

$$\Delta=1.96\sqrt{533579.415}=1432$$

所以，在95%置信度下，该街道共有房间数6363±1432
简单随机抽样的估计量是无偏的。

2. 某地有10万农户，现采用抽样调查来估计一种重要经济作物的产量。根据上一次普查结果，将全部农户按耕地面积分为7层，有关结果如下表：

层	N_h	$\bar{y}_h$	S_h^2
2	23000	0.72	2.89
1	50000	0.13	0.25
3	20000	3.34	72.25
4	5300	18.03	1225
5	1500	68.85	9025
6	120	786	40000
7	80	434	18900
合计	100000	$\bar{Y}=4.1773$	$\sum W_hS_h^2$ =286.66

1. 设样本量为3000，将第6层和第7层规模较大的200户农户全部收入样本，采用内曼分配从其余5层抽取2800个农户，求样本量在各层的分配。
2. 目标量是总产量，试求其方差估计量之值。

解：

1. 已知N'=100000 n'=3000, L=5, n=2800, N=99800
   采用奈曼分配, 各层样本量

$$n_h = n\frac{W_h S_h}{\sum_{h=1}^{3} W_h S_h} W_h S_h$$

,

$$n_1 = n\frac{W_1 S_1}{\sum_{h=1}^{5} W_h S_h} = 124$$

$$n_2 = n\frac{W_2 S_2}{\sum_{h=1}^{5} W_h S_h} = 195$$

$$n_3 = n\frac{W_3 S_3}{\sum_{h=1}^{5} W_h S_h} = 847$$

$$n_4 = n\frac{W_4 S_4}{\sum_{h=1}^{5} W_h S_h} = 924$$

$$n_5 = n\frac{W_5 S_5}{\sum_{h=1}^{5} W_h S_h} = 710$$

2. $$V(\hat{Y}) = N^2 V(\bar{y}_{min}) = N^2(\frac{1}{n}(\sum_{h=1}^{5} W_h S_h)^2 - \frac{1}{N}\sum_{h=1}^{5} W_h S_h^2) = 91287605$$

3. 检查某书稿上的错别字，每10页抽查1页上的错别字，系统抽取20页后的错别字结果如下

2	**3**	**4**	**0**	**6**	**8**	**8**	**5**	**9**	**9**
0	4	0	8	0	3	5	0	3	0

（1） 估计这本书稿平均每页的错别字数；
（2） 用合并层法估计抽样方差
解:
1)

$$\bar{Y} = \frac{1}{20}\sum_{i=1}^{20} y_i =$$

2)

$$V(\bar{y}_{sy}) = \frac{1-f}{n^2}\sum_{i=1}^{\frac{n}{2}}(y_{2i} - y_{2i-1})^2$$

4. 为调查学生购书支出，某高校在全校6000名大学生中按简单随机抽样抽出78名学生，调查了他们最近一个学期用于购书支出后，得到$\bar{y}$=102.30元， s^2=13712。
   (1)试估计该校大学生最近这一学期用于购书的总支出，并给出95%的置信区间；
   (2)若要求在置信度95%（对应的t=1.96）下，估计的相对误差不超过10%，则应该抽出多少学生进行调查?
   解:

1. 已知N=6000，n=78, f=n/N=0.013, 则购书总支出的点估计为：

$$\hat{Y} = N\bar{y} = 6000 \times 102.3 = 612003$$

$$V(\hat{Y}) = N^2 V(\bar{y}) = \frac{N^2(1-f)}{n}s^2 =$$

$$\Delta = 1.96\sqrt{} =$$

2. $r \leq 10$ 在又放回条件下的样本量为

$$n_0 = \frac{t^2 s^2}{r^2 \bar{y}^2}$$

无放回的样本量修正为

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

5. 某农村共有300块地，为了估计种植粮食的面积，采用等距抽样方法每隔10块抽取一块，取得了下列数据（粮食播种面积：亩）：

0	**0.9**	**0**	**0**	**0.3**	**0.1**	**0.5**	**3.1**	**2.8**	**2.7**
2.8	2.6	2.3	3.5	2.4	3.8	4.1	4.9	6.0	5.4
2.3	2.9	2.1	6.3	8.2	5.4	6.5	6.6	6.1	6.0

要求： 1. 估计总的粮食播种面积； 2. 使用合并层法计算$\hat{Y}$的相对标准误差；

解:

1. 已知 N=300，k=10，$n = \frac{N}{k} = 30$ 根据已知数据，计算出样本均值为：$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 3.153$ 总量的估计为：$\hat{Y} = N\bar{y} = 300 \times 3.153 = 945.9$

2. 将等距抽样视为分层抽样，根据合并层法可得:

$$V(\bar{y}_{sy}) = \frac{1-f}{n^2}\sum_{i=1}^{\frac{n}{2}}(y_{2i} - y_{2i-1})^2$$

$$V(\hat{Y}_{sy}) = N^2 V(\bar{y}_{sy})$$

相对标准误差

$$r = t\frac{\sqrt{v(\hat{Y})}}{\hat{Y}} =$$

6. 公路一侧新植树1000株，一个月后调查成活情况，若要求 n=30。

    1. 如果采用直线等距抽样，试说明抽样方法；这种方法存在什么问题，采用什么方法来处理，提出你的建议方法和实施步骤；
    2. 取样后令 $Y_i = 1$（代表成活） $Y_i = 0$（代表未活）得$\sum_{i=1}^{30} Y_i = 24$，试估计成活率及其精度；

    解:

    1. 如果采用直线等距抽样，则以$k = N/n = 1000/30 = 33$为间距，把总体分成30 段，每段33个单元，在第一段的33 个单元中随机抽出一个单元为起点，假设为r，然后每隔30 个单元编号

抽出一个单元编号，直到抽出30 个单元，即，最终抽出的样本编号为r，r+33，r+2×33，…，r+29×33。

这种方法存在的问题是，当N 不是n 的整数倍时，即抽样间距k = N/n 不是整数时，k 只能取不大于N/n 的最大整数，实际抽取的样本量与计划的样本量不一致，从而每个总体单元的入样概率也不相同，得到的样本均值是有偏的。为了得到总体均值的无偏估计，可以采用循环等距抽样方法。

具体实施步骤是：将N 个总体单元排成首尾相接的一个圆，抽样间距k 取取不大于N/n 的最大整数33。从1 到N 中随机抽取一个整数作为起始单元r，然后每隔k 抽取一个单元，直到抽满n 个单元为止。

2. 根据已知条件，可以得到抽样比f = 30/1000 = 0.03。 所以，成活率P 的估计量：

$$\hat{P} = \bar{y} = \frac{1}{n}\sum_{i=1}^{3} 0y_i = \frac{24}{30} = 0.8$$

在本案例中，系统样本来自按无关标志排列的总体，系统抽样的效果等价于简单随机抽样，从而用简单随机抽样下成活率估计量方差作为系统抽样的方差估计为：

$$v(p) = \frac{1-f}{n-1}p(1-p) = 5.352 \times 10^{-3}$$

95%置信水平下的绝对误差为：

$$\Delta = 1.96\sqrt{5.32 \times 10^{-3}} = 0.1434$$

7. 对某地区进行家庭年收入调查，以居民户为抽样单元，将居民户划分为城镇居民和农村居民两层，调查样本量为50，调查获得如下数据：

层h	N_h	S_h^2
城镇居民（层1）	600	100
农村居民（层2）	900	25

试计算：

（1） 城镇居民与农村居民分别按比例分配和按内曼分配时的样本量；

（2） 按比例分配时调查结果得到$\bar{y_1} = 10.7$万元，$\bar{y_2} = 5.3$万元，计算以95%的把握程度估计总体均值的置信区间；

（3） 在（2）中的调查结果是按如按内曼分配得到，计算以95%的把握程度估计总体均值的置信区间。

解： 已知$n = 50, N = 600 + 900 = 1500, W_1 = N_1/N = 600/1500 = 0.4, W_2 = N_2/N = 900/1500 = 0.6$

1. 按比例分配，则有：

$$n_1 = W_1 n = 20$$

$$n_2 = W_2 n = 30$$

内曼分配，则有

$$W_1 S_1 = 0.4\sqrt{100} = 4$$

$$W_2 S_2 = 0.6\sqrt{25} = 3$$

$$n_1 = n\frac{W_1 S_1}{\sum_{h=1}^{2} W_h S_h} = 28.57$$

$$n_2 = n\frac{W_2S_2}{\sum_{h=1}^{2}W_hS_h} = 21.43$$

2. 比例分配均值的估计

$$\bar{y}_{prop} = \sum_{h=1}^{L} W_h\bar{y}_h$$

比例分配均值的方差为(忽略1-f)

$$v(\bar{y}_{prop}) = \frac{1-f}{n}\sum_{h=1}^{L} W_hS_h^2 =$$

$$\Delta = 1.96\sqrt{} =$$

3. 奈曼分配均值的方差

$$\frac{1}{n}(\sum_{h=1}^{L} W_hS_h)^2 - \frac{1}{N}\sum_{h=1}^{L} W_hS_h^2 =$$

$$\Delta = 1.96\sqrt{} =$$