章

解：1 这种抽样方法是等概率的。在每次抽取样本单元时，尚未被抽中的编号为 1～64 的这些单元中每一个单元被抽到的概率都是 $\frac{1}{100}$

2 这种抽样方法不是等概率的。利用这种方法，在每次抽取样本单元时，尚未被抽中的编号为 1～35 以及编号为 64 的这 36 个单元中每个单元的入样概率都是 $\frac{2}{100}$，而尚未被抽中的编号为 36～63 的每个单元的入样概率都是 $\frac{1}{100}$。

3 这种抽样方法是等概率的。在每次抽取样本单元时，尚未被抽中的编号为 20 000 21 000 中的每个单元的入样概率都是 $\frac{1}{1000}$，所以这种抽样是等概率的。

2.2 解：

项目	相同之处	不同之处
定义	都是根据从一个总体中抽样得到的样本，然后定义样本均值为 $\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i$。	抽样理论中样本是从有限总体中按放回的抽样方法得到的，样本中的样本点不会重复；而数理统计中的样本是从无限总体中利用有放回的抽样方法得到的，样本点有可能是重复的。
性质	(1) 样本均值的期望都等于总体均值，也就是抽样理论和数理统计中的样本均值都是无偏估计。 (2) 不论总体原来是何种分布，在样本量足够大的条件下，样本均值近似服从正态分布。	(1) 抽样理论中，各个样本之间是不独立的；而数理统计中的各个样本之间是相互独立的。 (2) 抽样理论中的样本均值的方差为 $V(\bar{y})=(1-f)\frac{S^2}{n}$，其中 $S^2=\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$。 在数理统计中，$V(\bar{y})=\frac{\sigma^2}{n}$，其中 σ^2 为总体的方差。

2.3 解：首先估计该市居民日用电量的 95%的置信区间。根据中心极限定理可知，在大样本的条件下，近似服从标准正态分布，$\bar{Y}$ 的 1 95% 的置信区间为 $[\bar{y}-z_{\frac{\alpha}{2}}\sqrt{V(\bar{y})},\ \bar{y}+z_{\frac{\alpha}{2}}\sqrt{V(\bar{y})}]$，$[\bar{y}-1.96\sqrt{V(\bar{y})},\ \bar{y}+1.96\sqrt{V(\bar{y})}]$。

$V(\bar{y}) = \frac{S^2}{n}$ 中总体的方差 S^2 是未知的，用样本方差 s^2 来代替，置信区间为 $\bar{y} - 1.96\sqrt{(1-f)\frac{s^2}{n}}, \bar{y} + 1.96\sqrt{(1-f)\frac{s^2}{n}}$

由题意知道，$\bar{y} = 9.5, s^2 = 206$，而且样本量为 $n = 300, N = 50\,000$，代入可以求得 $v(\bar{y}) = (1-f)\frac{s^2}{n} = (1 - \frac{300}{50000})\frac{206}{300} = 0.682\,5$。将它们代入上面的式子可得该市居民日用电量的 95% 置信区间为 $(7.880\,8, 11.119\,2)$。

下一步计算样本量。绝对误差限 d 和相对误差限 r 的关系为 $d = r\bar{Y}$。

根据置信区间的求解方法可知

$$P\left\{\frac{|\bar{y} - \bar{Y}|}{\sqrt{V(\bar{y})}} \le \frac{r\bar{Y}}{\sqrt{V(\bar{y})}}\right\} \ge 1 - \alpha$$

根据正态分布的分位数可以知道 $z_{\frac{\alpha}{2}} = 1$，所以 $V(\bar{y}) = V_y = \frac{r\bar{Y}^2}{z_{\frac{\alpha}{2}}}$

也就是 $n^{-1} = \left(\frac{1}{N}\right)S^2 = \left(\frac{r\bar{Y}}{z_{\frac{\alpha}{2}}}\right)^2 \Rightarrow n = 1\Big/\left[\frac{1}{N} + \frac{\left(r\bar{Y}\right)^2}{z_{\frac{\alpha}{2}}^2 S^2}\right]$

把 $\bar{y} = 9.5, s^2 = 206, r = 10\%, N = 50\,000$ 代入上式可得，$n = 861.75 \approx 862$。所以样本量至少为 862。

2.4 解：总体中参加培训班的比例为 P，那么这次简单随机抽样得到的 P 的估计值 $p = \frac{1}{n}$ 的方差 $V(p) = P(1-P)\frac{1-f}{n-1}$，利用中心极限定理可得 p 在大样本的条件下近似服从标准正态分布。在本题中，样本量足够大，从而可得 P 的 $1-95\%$ 的置信区间为 $(p - z_{\frac{\alpha}{2}}\sqrt{V(p)}, p + z_{\frac{\alpha}{2}}\sqrt{V(p)})$。

而这里的 $V(p)$ 是未知的，我们使用它的估计值

$1 f_5$

$V p v p p 1 p 9.652 10^{5}$ P 的 1 95% 的置信区间可以写为 $p z_2 v p, p z_2 v p$，将 $p 0.35, n 200, N 10\,000$ 代入可得置信区间为 $0.284\,4, 0.415\,6$。

2.5 解：利用得到的样本，计算得到样本均值为 $y 2\,890 / 20 144.5$，从而估计小区的平均文化支出为 144.5 元。总体均值 Y 的 1 95% 的置信区间为 $y z_2 V y, y z_2 V y$，用 $v y {}_1 n f s^2$ 来估计样本均值的方差 $V y$。

$_2 1 f_2 1 0.1$

计算得到 $s^2 826.025\,6$，则 $v y \quad s^2 \quad 826.025\,6$，$37.172$

$n 20$

$z_2 V y 1.96\ 37.172\ 11.95$，代入数值后计算可得总体均值的 95% 的置信区间为 $132.55, 156.45$。

2.6 解：根据样本信息估计可得每个乡的平均产量为 1 120 吨，该地区今年的粮食总产量 Y 的估计值为 $\hat{Y} 350\, \bar{y} 350\ 1120\ 3.92\ 10^{5}$（吨）。

总体总值估计值的方差为 $V Y \quad {}_Y N 1 f s^2$，总体总值的 1 95% 的置信区间 n

$\hat{Y} z_2 \sqrt{V \hat{Y}}, \hat{Y} z_2 \sqrt{V \hat{Y}}$，把 $\hat{Y} 3.92\ 10^{5}, S^2 25\,600, n 50, N 350,$

$f_N n, z_2 1.96$ 代入，可得粮食总产量的 1 95% 的置信区间为 $377\,629, 406\,371$。

2.7 解：首先计算简单随机抽样条件下所需要的样，把 $N 1\,000, d 2, 1 \quad 95\%, S^2 68$ 带入公式 $n_0 1 \quad \frac{1 \quad d^2}{N z^2_{/2} S^2}$ 最后可得 $n_0 61.3\ 62$。

如果考虑到有效回答率的问题，在有效回答率为 70% 时，样本量应该最终确定为 $n n_0 70\% 88.57\ 89$。

2.8 解：去年的化肥总产量和今年的总产量之间存在较强的相关性，而且这种相关关系较为稳定，所以引入去年的化肥产量作为辅助变量。于是我们采用比率估计量的形式来估计今年的化肥总产量。去年化肥总产量为 $X 213\,5$。利用去年的化肥总产量，今年的化肥

$_R RX\ y_{\_}\ X\ 2\ 426.14$ 吨。
x

2.9 解：本题中，简单估计量的方差的估计值为 $v\ y$ $1f_2$ s^2 =37.17 。
n

利用比率估计量进行估计时，我们引入了家庭的总支出作为辅助变量，记为 X 。文化 支出属于总支出的一部分， 这个主要变量与辅助变量之间存在较强的相关关系， 而且它们之 间的关系是比较稳定的，且全部家庭的总支出是已知的量。

文化支出的比率估计量为 $y_R RX\ \hat{y}\ \bar{X}$ ，通过计算得到 $y\ 2\ 890/\ 20\ 144.5$ ，而
x

$\bar{\ }\hat{y}\ 144.5\ \bar{\ }$
$x\ 1580$ ，则 $R\ \bar{\ }\ 0.0915$ ，文化支出的比率估计量的值为 $y_R\ 146.3$ （元）。

，则 $x\_\ 1580$ ，文化支出的比率估计量的值为 $_R$ （元）。

现在考虑比率估计量的方差，在样本量较大的条件下， $1f_{222}$
$V\ y_R\ MSE\ y_R$ $S^2\ 2RS\ S_x R^2 S_x^2$ ，通过计算可以得到两个变量的样
n

本方差为 $s^2\ 826,\ s_x^2\ 9.958\ 1\ 40$， Y 和 X 之间的相关系数的估计值为 0.974 ，

代入上面的公式，可以得到比率估计量的方差的估计值为 $v\ y_R$ 1.94 。这个数值

比简单估计量的方差估计值要小很多。全部家庭的平均文化支出的 1 95% 的置

信区间为 $y_R z_{\ 2} v\ y_R, y_R z_{\ 2} v\ y_R$ $y_R\ 1.96$ $v\ y_R, y_R\ 1.96\ v\ y_R$

把具体的数值代入可得置信区间为 $143.57,149.03$ 。

$V\ y_R$ $v\ y_R$
$RR\ 1.94\ V\ y\ v\ y\ 37.17$

接下来比较比估计和简单估计的效率， 0.052 ，这是比估计的设计效应值，从这里可以看出比估计量比简单估计量的效率更高。

2.10 解：利用简单估计量可得 y $y_i\ n\ 1\ 630/10\ 163$ ，样本方差为 $s^2\ 212.222$ ，

$1\ f_{\ 2}\ 1\ 10/120$
$N\ 120$ ，样本均值的方差估计值为 $v\ y$ s^2 $212.222\ 19.453\ 7$ 。
$n\ 10$

利用回归估计的方法， 在这里选取肉牛的原重量为辅助变量。 选择原重量为辅助变量是 合理的， 因为肉牛的原重量在很大程度上影响着肉牛的现在的重量， 二者之间存在较强的相

关性，相关系数的估计值为 0.971 ，而且这种相关关系是稳定的，这里肉牛的原重量的数值已经得到，所以选择肉牛的原重量为辅助变量。

回归估计量的精度最高的回归系数 的估计值为 $_s\ 0.971$ $_{14.568}\ 1.368$ 。 s_x 10.341

现在可以得到肉牛现重量的回归估计量为 y_{lr} $y X x$，代入数值可以得到

$_{lr}$ 159.44

回归估计量 y_{lr} 的方差为 $V\left(\bar{y}_{lr}\right) \approx MSE\, y_{lr} \approx \frac{1-f}{n} S^2 1^2$，方差的估计值为

$v\left(\bar{y}_{lr}\right) \frac{1f}{n} s2\ 1\ \hat{}2$，代入相应的数值，$\left(\bar{y}_{lr}\right) \frac{1f_2}{n} s^2 \quad 1\ \hat{}2$ 1.112，显然有 $v\, y_{lr}\, v\, y$。在本题中，因为存在肉牛原重量这个较好的辅助变量，所以回归估计量 的精度要好于简单估计量。

第 章

3.1 解：在分层随机抽样中，层标志的选择很重要。划分层的指标应该与抽样调查中 最关心的调查变量存在较强的相关性，而且把总体划分为几个层之后，层应该满足：层内之 间的差异尽可能小，层间差异尽可能大。这样才能使得最后获得的样本有很好的代表性。对 几种分层方法的判断如下：

(1) 选择性别作为分层变量，是不合适的。首先，性别这个变量与研究最关心的变量 (不同职务，职称的人对分配制度改革的态度)没有很大的相关性；其次，用性别作为分层变量后，层内之间的差异仍然很大，相反，层之间的差异不是很大，因为男性和女性各自内部的 职务，职称也存在很大的差别；最后，选择性别作为分层变量后，需要首先得到男性和女性 的抽样框，这样会更加麻烦，也会使抽样会变得更加复杂。

(2) 按照教师、行政管理人员和职工进行分层，是合适的。这种分层的指标与抽样调查 研究中最关心的变量高度相关，而且按照这种方法分层后，可以看出层内对于分配制度改革 的态度差异比较小，因为他们属于相同的阶层，而层之间的态度的差异是比较大的。这样选 取出来的样本具有很好的代表性。

(3) 按照职称(正高、副高、中级、初级和其他)分层，也是合理的。理由与(2)相同，这样进行分层的变量选择与调查最关心的变量是高度相关的，分层后的层满足分层的要求。所以，按照职称进行分层是合理的。

(4) 按照部门进行分层，是合理的。因为学校有很多院、系或者所，直接进行简单随机 抽样，有可能样本不能很好地代表各个院系，最关心的变量与部门也存在一定的相关性。这 样分层后，每个层的总体数目和抽取的样本量都较小，最终的样本的分布比较均匀，比简单 随机抽样更加方便实施。

3.2 解：设计的方案如下：

第一种方案：可以按照不同的专业进行分层，但是考虑到如果在每层都抽取，不能保证 每个新生的入样概率相等，因为每个专业的人数比例未知， 8 个人的样本量无法在每个层之

所以采取如下方法：对所有的新生按照专业的先后顺序进行编号，使得每个专 业的人的编号在一起，然后随机选取出一个号码，然后选取出这个号码所在的专业，选取出 这个专业，再在这个专业的所有新生中按照简单随机抽样的方法选取出　个人。这样就可以 保证每个人入选的概率是相等的。

第二种方案：也可以按照性别进行分类，对他们进行编号，为　　1～800，使得男生的编号都在一起，女生的编号也都在一起，然后随机选取出一个号码，然后看这个号码所对应的 性别，然后从这个性别的所有人中按照简单随机抽样的方法选取出　　8 个新生。这样就可以保证所有的新生的入样概率是相同的。

第三种方案：　随机地把所有的人分成 8 组，而且使得每组的人都是 100 个人，这样分组完成后，每个组的新生进行编号为　　1～ 100，然后随机抽取出一个号码，再从所有的小组中抽取出号码所对应的新生，从而抽取出 8 个人。

解：(1) 首先计算出每层的简单估计量，分别为　　$y_1 11.2, y_2 25.5, y_3 20$，其中，$N_1 256, N_2 420, N_3 168, N 844$，则每个层的层权分别为；

$$W_1 \frac{N1}{NNN} 0.303\ 3, W_2 \frac{N2}{NNN} 0.497\ 6, W_3 \frac{N3}{NNN} 0.1991$$

则利用分层随机抽样得到该小区居民购买彩票的平均支出的估计量　　y_{st}　　$W_h y_h$，代入数值可以得到 y_{st}　　$W_h y_h 20.07$。

$_{-3}$ 1 f 购买彩票的平均支出的的估计值的方差为 $V y_{st}$　　$W_h^2 {}_{h}{}^{1f} S_h^2$，此方差的估计值为 $v y_{st}$　　$W_{h1}^{2} {}_{h}{}^{1f} s_h^2$，根据数据计算可以得到每层的样本方差分别为：

$$s_1^{2} 94.4, s_2^2 302.5, s_1^2 355.556$$

其中 $n_1 n_2 n_3 10$，代入数值可以求得方差的估计值为　　$v y_{st}$　　9.473 1，则估计的标准差为 $s y_{st}$　　$\sqrt{v y_{st}}$　　9.473 1 3.08。

(2) 由区间估计可知相对误差限满足

$$P y_{st} Y rY 1 \quad y_{st} Y P V y_{st} \quad V y_{st} rY$$

$$rY_1$$

$$\text{所以 } V_{rY} y_{st} z_2, \quad V y_{st} \quad \frac{rY^2}{z_2}$$

样本均值的方差为 $V(y_{st}) = \sum_{h=1}^{} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2 \left(\frac{1}{n}\right) - \sum_{h=1}^{} \frac{W_h S_h^2}{N}$，从而可以得到在置信度为，相对误差限为 r 条件下的样本量为 $n = \frac{\sum W_h^2 S_h^2 \sum}{V(y_{st}) + \frac{1}{N}\sum W_h S_h^2} = \frac{\sum W_h^2 S_h^2 \sum}{r^2 \left(\bar{Y}/z_{\alpha/2}\right)^2 + \frac{1}{N}\sum W_h S_h^2}$。

对于比例分配而言，有 $W_h = $ $_h$ 成立，那么 $n = \frac{z_{\alpha/2}^2 \sum W_h S_h^2}{r^2 \bar{Y}^2 + \frac{1}{N}\sum W_h S_h^2}$，把相应的估计值和数值 $1-95\%, r = 10\%$ 代入后可以计算得到样本量为 $n = 186$，相应的在各层的样本量分别为 $n_1 = 56.4 = 57, n_2 = 92.6 = 93, n_3 = 186 - n_1 - n_2 = 36$。

②按照内曼分配时，样本量在各层的分配满足 $_h W_h S_h = W_h S_h$，这时样本量的计算公式变为 $n = \frac{z_{\alpha/2}^2 \sum W_h S_h^2}{r^2 \bar{Y}^2 + \frac{1}{N}\sum W_h S_h^2}$，把相应的数值代入后可得 $n = 175$，在各层中 $n_1 = 33, n_2 = 87, n_3 = 186 - n_1 - n_2 = 66$。

解：(1) 首先计算得到每层中在家吃年夜饭的样本比例为 $p_1 = 0.9, p_2 = 0.933\,3, p_3 = 0.9, p_4 = 0.866\,7, p_5 = 0.933\,3, p_6 = 0.966\,7$，那么根据每层的层权，计算得到该市居民在家吃年夜饭的样本比例为

$$p_{st} = \sum_{h=1}^{6} W_h p_h = 92.4\%$$

层中在家吃年夜饭的样本比例的方差为 $V(p_h) = \frac{1 - f_h}{n_h - 1} P_h(1 - P_h)$，则该市居民在家吃年夜饭的比例的方差，在 $N_h - 1 \approx N_h$ 的条件下，$V(p_{st}) = \sum_{h=1}^{6} W_h^2 V(p_h) = \sum_{h=1}^{N} \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h - 1} \frac{P_h(1-P_h)}{n_h}$

而其中每层的吃年夜饭的样本比例的方差的估计值为 $v(p_h) = \frac{1 - f_h}{n_h - 1} p_h(1 - p_h)$，则样本比例的方差的估计值 $v(p_{st}) = \sum_{h=1}^{6} \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h} \frac{p_h(1-p_h)}{n_h - 1}$

st $W_h^2 v\, p_h$ $W_h^2\, 1\, f_h\, ph\, 1\, ph$ ，把相应的数值代入计算可得方差的
$h\,1\,h\,1$ $n_h\, 1$

估计值为 $v\, p_{st}\, 3.9601\ 10^{-4}$ ，从而可以得到该估计值的标准差为 $s\, p_{st}\, 0.019\,9$ 。

(2)利用上题的结果， n hhhhhh ，这里的方 $W_h^2 S_h^2\ h$ $W_h^2 S_h^2\ h$
$1\ _2{}^2\ 1$
$V\, p_{st}$ $_N{}^2\ W_h^2 S_h^2$ $rP\, Z_{\ 2\,N}$ $W_h S_h^2$

差是 S_h^2 $_N^h$ $P_h\ \ 1\ \ P_h$ ，在 $N_h\, 1\, N_h$ 的条件下，近似有 S_h^2 $P_h\, 1\ \ P_h$ 。
$N_h\, 1$

比例分配的条件下，有 W_h $_h$ 成立，那么 n hh ，把相应的 $^{W\,S^2}$
$_2\, 1\ _2 rP\, z_{\ 2\,1}\, W_h S_h^2$
$_2 N$

估计值和数值代入可以求得最终的样本量应该是 $n\ 266\ 3$ ，样本量在各层的

分配是 $n_1\ 479.\ 34\ 4\ 7n9_{\ 2}$ ， $559.\ 23\ n5_{\ 3}\ 59,$ $3\ 7\ 2.$ ， $8\ 3n_4\ 327339.67\ 240$ ，

$n_5\ 426.08\ 426,\ n_{\ 6}\ 585.86\ 586$ 。

②内曼分配条件下， $_h W_h S_h$ $W_h S_h$ ，则 n $\dfrac{W_h S_h{}^2}{rP\, Z^2\ 1\ W_h S_h^2}$ ，代入相
$_2 N$

应的估计值和数值可以计算得到样本量为 $n\ 256\ 5$ ，在各层中样本量的分配

为 $n_1\ 536,\ n_{\ 2}\ 520,\ n_{\ 3}\ 417,\ n_{\ 4}\ 304,\ n_{\ 5}\ 396,\ n_{\ 6}\ 392$ 。

解：总体总共分为 10 个层，每个层中的样本均值已经知道，层权也得到，从而可

以计算得到该开发区居民购买冷冻食品的平均支出的估计值为 y_{st} ^{10} $W_h y_h\ 75.79$ 。
$h1$

下一步计算平均支出的 95% 的置信区间，首先计算购买冷冻食品的平均支出的估计值的

方差，其中 V $_-{}^{10}\ _2\, y_{st}\ W_h\ \ 1\, f_{\ 2\,h}\, S_h^2$ ，但是每层的方差是未知，则样本平均支出的方差的
n_h

估计值为 v $_-{}^{10}\ 1\, f$ $W_h^2\ _{1\,f\,h}\, s_h^2$ ，每个层的样本标准差已知，题目中已经注明各层的抽
y_{st}
$h\,1\ n_h$

样比可以忽略，计算可以得到 v y_{st} $_-{}^{10}\ {}^1 W_h^2\ _1$ $_h\, s_h^2\ 59.825\ 4$ 。则这个开发区的居民 $_{nh}$
$h1$

购买冷冻食品的平均支出 1 95% 置信区间为 $\left[\bar{y}-z_{\alpha/2}\sqrt{v\left(\bar{y}_{st}\right)},\bar{y}+z_{\alpha/2}\sqrt{v\left(\bar{y}_{st}\right)}\right]=$

$y\ 1.96\ v\ y_{st}, y\ 1.96$ $\sqrt{v\left(\bar{y}_{st}\right)}$

。

解：首先计算简单随机抽样的方差，根据各层的层权和各层的总体比例可以得到

3 总体的比例为 $P \quad W_h P_h\ 0.28$ ，则样本量为 100 的简单随机样本的样本比例的方差为

$V\ p \quad S^2$ ，不考虑有限总体校正系数，$V\ p \quad \frac{1-f}{n} S^2$ ，其中 $S^2 \quad P\ 1\ P$ ，在 $N\ 1\ N$ 的条件下，通过简单随机抽样得到的样本比例的方差为

$$V\ p_{nn} \quad S^2 \quad P\ 1\ P\ 2.016\ 10^{-3}$$

方差为 $V\ p_{st} \quad W_h^2 \frac{1-f_h}{n_h} S_h^2$ ，但是因为不考虑有限总体校正系数，而且抽样方式是比例抽样，所以有 $N_h\ W_{h\,h}\ n_h$ 成立，样本比例的方差近似为 $V\ p_{st} \quad nn \quad {}_{W\,S}{}^{h\,h}{}_{1}\ WhSh^2$ 。对于每一层，分别有 $Sh \quad {}_{N}{}^{h}\ Ph\ 1\ Ph$ ，在 $N_h\ 1\ N_h$ 的条件下，近似的有 $S_h^2\ P_h\ 1\ P_h$ 成立，有

$$S_1^2\ 0.09, S_2^2\ 0.16, S_3^2\ 0.24$$

$W_h S_h^2$

样本量应该满足 $n\ h\,h$ ，同时这里要求分层随机抽样得到的估计的方差和简单抽样的方差是相同的，$V\ p_{st}\ V\ p$ ，层权分别为 $W_1\ 0.2, W_2\ 0.3, W_3\ 0.5$ ，代入数值，可以计算得到最终的样本量为 $n \quad \frac{WhSh \quad 0.186}{V\ p_{st} \quad 2.016\ 10^{-3}} \quad 3\ 92.26\ 93$ 。

3.7 解：事后分层得到的总体均值的估计量和估计量的方差分别为

$E\ y_{pst}\ Y, E\ Var\ y_{pst}$

$\frac{1-f}{n} S_h^2, \frac{1-f_2}{n_2} W_h S_h^2$ ，估计量的方差的估计值 $v\ y_{pst}$ $\frac{1\ W_h}{n\ n\ n}$ $\frac{1}{n^2}\ 1\ W_h\ s_h$ 。

$^2 W_h s_h^2$

（ ）事后分层比简单随机抽样产生更加精确的结果，这个说法是错误的。从事后分层 得到估计量的方差的估计值来看， 它的方差不一定比简单随机抽样的要小， 而且从事后分层 得到的样本是利用简单随机抽样的方法得到的， 只是在计算估计量和估计量的方差时是按照 分层随机抽样来处理， 而且事后分层要求层权是已知的， 但是当层权未知从而利用样本来估 计层权时，就会产生偏差，事后分层不见得比简单随机抽样产生更精确的结果。

（2）事后分层比按比例分配产生更精确的结果，这个说法是错误的。从事后分层得到 的估计量的方差的估计值可以看出， 它的第一项就是按照比例分层抽样得到的估计量方差的 估计值，公式中的第二项表示的是按事后分层时各层样本量与按照比例分层时各层样本量发 生偏差所引起的方差的增量。

（3）事后分层的最优分配产生更精确的结果，这种说法是错误的。事后分层在样本量 足够大的条件下是与比例分层相当的， 但是在一般条件下， 事后分层的精度仍然低于比例分 层的，那么事后分层的精度也会高于最优分配的精度。

（4）在抽样时不能得到分层变量，这个说法是正确的。事后分层在抽样时，是利用简 单随机抽样的方法， 在抽样时不涉及按照变量进行分层， 至于按变量进行分层， 是在抽样完 成后，然后根据具体的变量来对样本进行分层。

（5）它的估计量的方差与真正按照比例分层随机抽样的方差差不多，只有在样本量足 够大的条件下才成立。在样本量足够大的条件下，从事后分层的方差的计算公式可以看出， 它的第二项会趋于 0，这时事后分层的估计量的方差和分层随机抽样的方差差不多。

解：（1） 根据简单随机抽样的公式， 登记原始凭证的差错率的估计值为 $p \frac{3}{100}$ 3%，在考虑到 $f 0,N N 1$ 的条件下，登记的原始凭证的差错率的估计量的方差近似

$v p \frac{1 f}{n} s2 \frac{1 f}{n} \frac{N1}{nN 1 n} P 1 P P1 P$

$p 1 p 2.91 10 4$，则估计量的方差的估计值为 $v p p 1 p$ ，计算得 $v p 1 1 4$

则原始凭证的差错率的估计的标准差为 $s p v p 1.71 10 2$ 。

2）这里，每个层的层权是事先知道的，那么利用事后分层来计算登记原始凭证的差

错率的估计值为 $p_{pst} \sum_{h1}^{2} W_h p_h 2.68\%$ ，在这里 $p_1 4_1 3 2.33\%,p_2 3.51\%$ 。

$1f$ 利用事后分层得到的原始凭证的差错率的估计量的方差的估计值为 $v p_{pst} \frac{1 f_{pst}}{n} 11 W_h s_h^2 \frac{1 W_h}{n} 2 v p_{pst}$ s_h^2 ，在不考虑有限校正系数的条件下，又可以写为 n^2

$W_h \frac{n_h}{n_h 1 n} p_h 1 p_h 12 1 W_h \frac{n_h}{n_h 1} p_h 1 p_h$ ，其中 $W_1 0.7,W_2 0.3,$

$n_1 43,n_2 57$ ，可以得到 $v p_{pst} 2.689 5 10 4$ ，则相应的标准差为 $s p_{pst}$

$1.64 10 2$ 。

1）所有可能的样本的数量为 $C_3^2\ C_3^2\ 9$ ，所有的样本如下：

3,0 , 5,3 , 8,6 , 15,9 ,　　3,0 , 5,3 , 8,6 , 25,15　　, 3,0 , 5,3 , 25,15 , 15,9　　,

3,0 , 10,6 , 8,6 , 15,9 ,　　3,0 , 10,6 , 8,6 , 25,15　　, 3,0 , 10,6 , 25,15 , 15,9　　,

5,3 , 10,6 , 8,6 , 15,9 ,　　5,3 , 10,6 , 8,6 , 25,15　　, 5,3 , 10,6 , 25,15 , 15,9

（2）我们用 9 个样本中的一个来计算，假定抽中的样本为 $5,3 , 10,6 , 8,6 , 25,15$ 。

首先按照分别比估计来估计 Y ，首先可以得到分层后的辅助变量的总体均值分别为 $X_1\ 6, X_2\ 16$ 。在这个样本中，经计算得到 $x_1\ 7.5, x_2\ 16.5, y_1\ 4.5, y_2\ 10.5$ ，$R_1\ 0.6, R_2\ 0.64$ ，而且 $W_1\ W_2\ 0.5$ ，则根据分别比估计可得 Y 的估计值为 y_{RS} _ $W_h\ y_{Rh}$　　$W_h\ R_h\ X_h\ 6.891$ 。

利用联合比估计时，首先计算得到辅助变量的总体均值 $X\ 11$ ，然后利用样本得到的 _ 主要变量和辅助变量的样本均值为 $y_{st}\ 7.5, x_{st}\ 12, R_c\ 7.5\ 12\ 0.625$ ，则利用联合比 _ _ _ 估计得到的 Y 的估计值为 $y_{RC}\ R_c\ X\ 6.875$ 。

在计算分别比估计和联合比估计的偏差，这里的方法是利用所有可能的样本，然后计算 出比估计和联合估计的估计值，按照与上面相同的计算方法，计算得到其他样本时比估计和 联合估计值　按照上面的样本的排列顺序 ）为：

$y_{RS1}\ 6.342, y_{RC1}\ 6.387, y_{RS2}\ 6.216, y_{RC2}\ 6.439, y_{RS3}\ 5.925, y_{RC3}\ 6.188,$

$y_{RS4}\ 6.602, y_{RC4}\ 6.243, y_{RS5}\ 6.476\ y_{RC5}\ 6.457, y_{RS6}\ 6.185, y_{RC6}\ 6.227,$

$y_{RS7}\ 7.017, y_{RC7}\ 6.947, y_{RS8}\ 6.6, y_{RC8}\ 6.6, y_{RS9}\ 6.891, y_{RC9}\ 6.875$

分别计算可得 $E\left[\bar{y}_{RS} \right.$ $\left. \sum_{9\ h\ 1}^{19} \bar{y}_{RSh} \right]\ 6.473, E\ \bar{y}_{RC}$ $\sum_{9\ h\ 1}^{19} y_{RCh}\ 6.485$ ，而且可以计算得到 $\mathrm{var}\ \bar{y}_{RC}\ 0.076$ ，$\mathrm{var}\left(\bar{y}_{RS} \right)\ 0.121$ 。总体的实际均值为 $Y\ 39/6\ 6.5$ 。则分别比估计和联合比估计的偏差分别为 $E\left(\bar{y}_{RS} \right) - \bar{Y}\ 6.473\ 6.5\ 0.027, E\left(\bar{y}_{RC} \right) -$ $Y\ 6.485\ 6.5\ 0.015$ 。

$\left| E\left(y_{RC}\right) - Y \right| = 0.015$ $\left| E\left(y_{RS}\right) - Y \right| = 0.027$，所以联合比估计的偏差比分别比估计的偏差要小。

接下来计算分别比估计和联合比估计的均方误差。在这里样本量很小，不可以利用教材中的近似公式。

$$MSE\left(y_{RS}\right) = \left| \mathrm{var}\left(\bar{y}_{RS}\right) \right| + \left| E\left(\bar{y}_{RS}\right) - Y \right| = 0.121\ 0.000\ 729\ 0.122$$

$$MSE\left(y_{RC}\right) \mathrm{var}\left(y_{RC}\right) + \left| E\left(y_{RC}\right) - \bar{Y} \right| = 0.076\ 0.000\ 25\ 0.076\ 3$$

$$MSE\left(y_{RC}\right) 0.076\ 3 < MSE\left(y_{RS}\right) = 0.122$$

(3) 从分别比估计和联合比估计的偏差和均方误差可以看出，联合比估计的偏差和均方误差都要小于分别比估计，也就是说在本题中，联合比估计要比分别估计好。在本题中，各层的比率和总体的比率相差基本差不多，从整个样本出发进行的联合比估计比基于每层的分别比估计更好一些，偏差更小，均方误差也更小。

第　章

4.1 解:由题意知,平均每户家庭的订报份数为：

$$\bar{y} = \sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}/nM = (19+20+16+20)/10/4 = 1.875\ 2 \quad (份)$$

总的订报份数为：

$$Y_{?} = N\bar{y} = 4\ 000\ 1.875\ 7\ 500 \quad (份)$$

$$s_b^2 = \frac{M}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 2^2 = 0.358\ 333$$

所以估计方差为

$$v(\bar{y}) = \frac{1-f}{nM} s_b^2 = \frac{1-0.01}{4\ 10} 0.358333 = 0.008\ 869$$

$$v(Y_{?}) = N^2 v(\bar{y}) = N^2 \frac{1-f}{nM} s_b^2 = 141\ 900$$

4.2 解：

总人数	赞成人数	赞成比例 p_i

	51	42	0.823 529
2	62	53	0.854 839
3	49	40	0.816 327
4	73	45	0.616 438
5	101	63	0.623 762
6	48	31	0.645 833
7	65	38	0.584 615
8	49	30	0.612 245
9	73	54	0.739 726
10	61	45	0.737 705
11	58	51	0.879 31
12	52	29	0.557 692
13	65	46	0.707 692
14	49	37	0.755 102
15	55	42	0.763 636

n

$$M_i$$

(1) $m_{i\,1}$ $=60.733\ 33$

n

y_{y} $=70.91\%$ m

其估计的方差为：

$$v(y)\ N^2\ 1\ f\ \frac{\sum_{i\,1}^{n}(yi\ \ y)^2}{nM_0^2\ \ n\ 1}$$

$$=0.001\ 37$$

所以其估计的标准误为

$$s(y)\qquad v(y)\ =3.7\%$$

(2) $s(y)$ $v(y)\ =8\%$

$$v(y)\ \ \frac{1\ f\ \sum_{i\,1}^{n}(y_i\ y)^2}{N^2\ \ nM_0^2 n1}\ \ \frac{\sum_{i\,1}^{n}(y_i\ y)^2}{n(mN)^2\ \ n\ 1}\ \ \frac{(y_i\ y)^2}{nm^2\ \ n1}$$

$$\frac{1f}{n^2m}\ \ \frac{\sum_{i1}^{n}(y_i\ y)^2}{n1}\ =0.006\ 4$$

得 n=6.2 ，所以应抽取 7 个单位作样本。

4.3 解 :该集团办公费用总支出额为：

$$Y_{?N}^{Nn}\quad y_i=48/10\times(83+62+\quad+67+80)=3\ 532.8(\quad 百元)$$

$$v(Y^{?})\ N^{21\ f}\ {}_{i\ 1}^{n}\ \frac{(y_i\ y)^2}{n\quad n\ 1}\ 1\ f\ =72\ 765.44$$

$$s(Y?)\quad v(Y?)\ =269.750\ 7(\quad 百元)$$

所以其置信度为　　95% 的置信区间为 :[3 004.089 , 4 061.511]

4.4 解:$m\ {}_{i\ 1}^{n}\ \frac{M_i}{n}\ =52.3$

所以整个林区树的平均高度为

$$y\ y\ =5.9(\ 米)\ m$$

其估计的方差为：

$$N^{2\ 1}\ {}_{nM0}^{2}\ {}_{2}^{1\ v(y)}\ {}_{i\ 1}^{f}\ \frac{(y_i\ y)^2}{n\ 1}\ \frac{N\ n(mN)^2}{n}\ {}_{1\ f}^{1}\ {}_{i\ 1}^{n}\ \frac{(y_i\ y)^2}{n1}$$

$$1\ f\ {}_{i\ 1}^{n}\ \frac{(y_i\quad y)^2}{nm^{\ 2}\quad n1}\ {}_{(y}^{i}\ =0.06$$

$$s(y)\ v(y)\ =0.246(\ 米)$$

其 95% 的置信区间为 :[5.42 ,6.38]

4.5 解:拍摄过艺术照的女生比例为

$$y\ {}_{nm\ i\ 1\ j\ 1}^{1nm}\ y_{ij}=9/30=30\%$$

其估计的方差为：

$$v(y)\ {}_{n}^{1}\ {}_{n^1}^{s^{12}}\ {}_{f\ (1\ f\ )\ s2}^{1\quad 2\quad 2}\ {}_{nm}\ =0.005\ 891$$

其估计的标准差为：

$$s(y)\quad =\sqrt{v(\bar{y})}\ =7.68\%$$

4.6 解:$m_{opt}=\frac{S_2}{S_1}\sqrt{\frac{C_1}{C_2}}=\frac{188}{316.8}\sqrt{\frac{10}{1}}\approx$

$$s_u^2 \; s_1^2 \; 326^2 \qquad {}^2 188 \qquad 100\ 385.33$$

M6

所以最优的样本学生数为 2。

代入 $c_c 0\ c_1 n\ c_2 nm$ 得到

$$n_{opt} \; 20$$

所以最优的样本宿舍数为 20。

4.7 解:(1)简单估计:居民总的锻炼时间为:

$$Y_u^{?} \quad N \sum_{i} M_i \frac{m_i}{n_{i1} m_i} \sum_{j1} y_{ij} \quad =1\ 650$$

居民平均每天用于锻炼的时间为

$$\frac{Y_u^{?}}{M_0} \quad =3.3 \text{ (即 33 分钟)}$$

$$v(y) \quad \frac{1}{2} N^2 (1 f_1) \frac{n}{n-1} (Y_i^{?} Y_u^{?})^2 \frac{N}{n} M_i^2 \frac{f_1}{n} (1 f_2) s_{22i} M_0^2 \frac{n}{i1}$$

$$=0.163\ 421$$ 其估计的标准差为:

$$s(y) \; v(y) \; =0.404\ 254$$ (2) 比率估计:居民总的锻炼时间为:

$$Y_R^{?} \quad \frac{\sum_{i1}^{n} M_i \frac{m_i}{m_i} \sum_{j1} y_{ij}}{M_0}$$

居民平均每天用于锻炼的时间为

$$\bar{y} = \frac{Y_R^{?}}{M_0} \quad \frac{\sum_{i1}^{n} M \frac{m_i}{m_i} \sum_{j1} y_{ij}}{M} \quad =3.95 \text{ (即 39.5 分钟)}$$

$$v(y) \quad \frac{M^2}{2} \frac{1}{n} \frac{N^2 (1 f_1)}{n-1} (Y_i^{?} Y_u^{?})^2 \frac{N}{n} \frac{n}{M_0^2 n} \sum_{i1} M_i^2 \frac{f_{1m}}{m_i} (1 f_2) s_{22i}$$

$$=0.071\ 509$$ 其估计的标准差为:

$$s(y) \; v(y) \; =0.267\ 411$$

:

r=0.404 254/3.3 =12.25%　比估计下的相对误差为：

r=0.267 411/3.95=6.77%　所以比估计的估计效果好。

第 5 章

5.1 解：(1)代码法列出下表：

PUS	Z_i	$Z_i \times 1\ 000\ 000$	累计 $Z_i \times 1\ 000\ 000$	代码
1	0.000 110	110	110	1～110
2	0.018 556	18 556	18 666	111～18 666
3	0.062 999	62 999	81 665	18 667～81 665
4	0.078 216	78 216	159 881	81 666～159 881
5	0.075 245	75 245	235 126	159 882～235 126
6	0.073 983	73 983	309 109	235 127～309 109
7	0.076 580	76 580	385 689	309 110～385 689
8	0.038 981	38 981	424670	385 690～424 670
9	0.040 772	40 772	465 442	424 671～465 442
10	0.022 876	22 876	488 318	465 443～488 318
11	0.003 721	3 721	492 039	488 319～492 039
12	0.024 971	24 971	517 010	492 040～517 010
13	0.040 654	40 654	557 664	517 011～557 664
14	0.014 804	14 804	572 468	557 665～572 468
15	0.005 577	5 577	578 045	572 469～578 045
16	0.070 784	70 784	648 829	578 046～648 829
17	0.069 635	69 635	718 464	648 830～718 464
18	0.034 650	34 650	753 114	718 465～753 114
19	0.069 492	69 492	822 606	753 115～822 606
20	0.036 590	36 590	859 196	822 607～859 196
21	0.033 853	33 853	893 049	859 197～893 049
22	0.016 959	16 959	910 008	893 050～910 008
23	0.009 066	9 066	919 074	910 009～919 074
24	0.021 795	21 795	940 869	919 075～940 869
25	0.059 185	59 185	1 000 054	940 870～1 000 054

表中，Z_i 不是整数，乘以 1 000 000 使其变为整数，这样就可以赋予每个单元与其相等 的代码数。

先在[1,1 000 054] 中产生第一个随机数为 825 011 ，其对应的单元为 20 号，则得到第 一个入样单元 20 ；

去掉，剩余的 24 个单元，累计代码数为 1 000 054-36 590=963 464 ，在 [1,963464] 中产生第二个随机数为 456 731 ，得到第二个入样单元 9；

再把单元 9 去掉，剩余的 23 个单元，累计代码数为 963 464-40 772=922 692 ，在 [1, 922 692] 中产生第三个随机数为 857 190 ，得到第三个入样单元 24 ；

依此类推，直至抽出所需的样本。

最后抽得的 10 个入样单元为 20,9,24,3,4,25,21,16,7,5 。

（2） 拉希里法”。

令 $Z^* = \max Z_i = 0.078\ 216$ ， $N = 25$ ，在 [1,25] 和[1, 0.078 216] 中分别产生随机数 $6, 0.021\ 313$ ， $Z_6 = 0.073\ 983 > 0.021\ 313$ ，第 6 号单元入样；

把单元 6 去掉，剩余的 24 个单元， $\max Z_i$ 仍旧等于 0.078 216 ，在 [1,24] 和 [1, 0.078 216] 中分别产生随机数 $10, 0.031\ 543$ ， $Z_{10} = 0.022\ 876 < 0.031\ 543$ ，第 10 号单元不入样，重新抽取随机数；

依此类推，直至抽出所需的样本。

最后抽得的 10 个入样单元为 6,9,18,4 ， 1,5,19,21 ， 16,13 。

PSU	M_i	Z_i	y_{ij}	t_i
1	5	0.2	3,5,4,6,2	20
2	4	0.16	7,4,7,7	25
3	8	0.32	7,2,9,4,5,3,2,6	38
4	5	0.2	2,5,3,6,8	24
5	3	0.12	9,7,5	21

5.2. 解：首先计算 $\dfrac{4Z_iZ_j(1-Z_i-Z_j)}{(1-2Z_i)(1-2Z_j)(1-\sum_{i=1}^{1}2Z_i)}$ 出各 PSU 单元的入样概率，$M_0 = 25$ 。

由 π_{ij} 可得所有可能样本的 π_{ij}

样本	π_{ij}	$Y?$
1,2	0.068 091	128.125
1,3	0.192 926	109.375
1,4	0.090 434	110
1,5	0.048 549	137.5

	0.147 531	137.5
2,4	0.068 091	138.125
2,5	0.036 286	165.625
3,4	0.192 926	119.375
3,5	0.106 617	146.875

：代码法列出下表：

i	Z_i	$Z_i \times 1\,000$	累计 $Z_i \times 1\,000$	代码
1	0.104	104	104	1～104
2	0.192	192	296	105～296
3	0.138	138	434	297～434
4	0.062	62	496	435～496
5	0.052	52	548	497～548
6	0.147	147	695	549～695
7	0.089	89	784	696～784
8	0.038	38	822	785～822
9	0.057	57	879	823～879
10	0.121	121	1 000	880～1 000

4,5	0.048 549	147.5

霍维茨－汤普森估计量的方差为 $\hat{V}(Y)$ $\sum_{i=1}^{n}\sum_{j=1}^{n}(\pi_{ij}-\pi_i\pi_j)(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j})^2$ 3 787.572 。

表中，Z_i 不是整数，乘以 1 000 使其变为整数，这样就可以赋予每个单元与其相等的代 码数。

在[1,1 000] 之间产生三个随机数 659，722，498，则它们所对应的第 6， 7，5 号单元 被抽中，即得到的 n=3 的 PPS 样本包括单元 6、单元 7 和单元 5。

5.4 解：由题意知 n=3, 总体总量的估计为：

$\hat{Y}_{HH}$ $\frac{1}{3}\sum_{i=1}^{3}\frac{y_i}{Z_i}=\frac{1}{3}(\frac{320}{0.138}$ $+\frac{120}{0.062}+\frac{290}{0.121})=2\,217.006\,2$

总量估计的标准差为：

$s(Y_{?HH})$ $v(Y_{?HH})$ $\sqrt{\frac{1}{3(3-1)}\sum_{i=1}^{3}(\frac{y_i}{Z_i}-\hat{Y}_{HH})^2}$

$=\sqrt{((320-2\,217.006\,2)^2+(120-2\,217.006\,2)^2+(290-2\,217.006\,2)^2)/6}$

142.544 1

5.5 解：由题意知 $n=2$，$M_0=23$，$Z_i=\frac{M_i}{M_0}$，每个单元的入样概率 $_i=nZ_i$