扩展：　　抽样调查

常见的抽样方法

（1）简单随机抽样

对北航学生的研究能力进行抽样测试。在北航全校学生中随机抽取 n 名学生。

（2）分层抽样

分层次抽样：专科、本科、研究生、博士、博士后。

（3）整群抽样

在本科生中，随机抽取若干个班，观察每个班的全部学生。

（4）分段抽样

全国调查，随机抽取若干省，再随机抽取若干市，再随机抽取若干区，...

（5）非随机抽样

1 简单随机抽样方法

简单随机抽样：

每一个容量为 n 的可能样本被抽到的概率都是一样的。

原则：调查者不能根据主观意图挑选调查单位。而是在总体中，按照随机原则和纯粹偶然性的方法抽取样本。

-方法：（1）抽签法

（2）随机数字表，随机数发生器

抽签法：先将调查总体的每个单位编上号码，然后将号码写在卡片上搅拌均匀，任意从中选取。抽到一个号码，就对上一个单位，直到抽足预先规定的样本数目为止。

优点：　可以获得一个无偏倚的样本

使用限制：实施操作并不简单

（1）保证样本点分布均匀；

（2）有时,调查人员要了解所有样本中的个体有时是很困难的。

（3）样本容量较小时，一些比例少但是很重要的个体不能入样，使样本的代表性受到影响。

例如：在人民银行随机抽取100名职员，可能会抽不到高层管理人员。

TBT调查在全国抽1000家企业，可能会有许多大型企业不能入样。

(1) 总体均值的估计

- 放回抽样

总体均值的点估计

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

$$D(\bar{x}) = \frac{\sigma^2}{n}$$

总体均值的区间估计（抽样误差）

$$(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}})$$

- # 不放回抽样

总体均值的点估计

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

N—总体中的个体数量

n —样本容量

$$D(\bar{x})=\boxed{\frac{N-n}{N-1}}\cdot\frac{\sigma^2}{n}$$

$$\frac{N-n}{N-1}$$ 数

当$N\to\infty,\ \dfrac{N-n}{N-1}\to 1$。

意

$$\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}<\frac{\sigma^2}{n}$$

同样样本容量下，不放回抽样的误差更小！

总体均值的区间估计

$$\bar{x}\pm t_{\alpha/2}\sqrt{\frac{N-n}{N-1}}\frac{s}{\sqrt{n}}$$

[自由度 $df=(n-1)$]

例：

某居民区共有 $N = 200$户居民，随机抽取 $n = 20$ 位居民，他们每日收看电视的时间如下：

60 90 100 30 90 60 180 80 70 90
180 120 30 60 90 120 80 80 100 90

求该居民区居民平均每日收看电视时间的点估计和区间估计；

求该居民区居民平均每日收看电视时间的点估计和区间估计；

$$\bar{x} = \frac{1}{20}[60 + 90 + 100 + \boxed{?} + 90] = 90 \text{ 分钟}$$

$$s^2 = \frac{1}{20-1}\left[(60-90)^2 + (90-90)^2 + \boxed{?} + (90-90)^2\right] = 1515.7895$$

$$s = \sqrt{1515.7895} = 38.93$$

$$\text{取}\alpha = 0.05 \quad \Rightarrow \quad t_{\alpha/2}(19) = 2.093$$

$$90 \pm 2.093\sqrt{\frac{200-20}{200-1}} \cdot \frac{38.93}{\sqrt{20}} \approx 90 \pm 17$$

相对误差为：　17 / 90 = 19%　（显然，样本容量不够大）

(2) 总体比例的估计 (大样本)

- **放回抽样**

总 $$p=\frac{A}{N}$$

样 $$\hat{p}=\frac{a}{n} \quad ($$

$$E(\hat{p})=p, \qquad D(p)=\frac{1}{n}p(1-p)$$

区 $$\boxed{\hat{p}\pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-p)}{n}}}$$

●不放回抽样

(1)[illegible]

$$\hat{p} = \frac{a}{n}$$

$$E(\hat{p}) = p, \qquad D(\hat{p}) = \boxed{\frac{N-n}{N-1}} \cdot \frac{p(1-p)}{n}$$

[illegible] $\frac{N-n}{N-1}$ [illegible]

(2)[illegible]

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{N-n}{N-1}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

例题：某城市想要估计下岗职工中女性所占的比例，随机抽取了100名下岗职工，其中65人为女性。试估计该城市下岗职工中女性比例，并指出估计误差。置信水平要求为95%。

已知 $n=100$，$\alpha=0.05$，$z_{\alpha/2}=1.96$　$\hat{p}=65\%$

放回抽样的置信区间为：

$$0.65 \pm 1.96\sqrt{\frac{0.65\times 0.35}{100}}$$

$$=65\% \pm 9.35\%$$

不放回抽样的置信区间半长： $N \to \infty$

$$z_{\alpha/2}\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\hat{p}(1-p)}{n}} \approx 1.96\sqrt{\frac{0.65\times 0.35}{100}} = 9.35\%$$

7.2 样本容量的确定

问题：估计某地区的平均收入

$$D = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

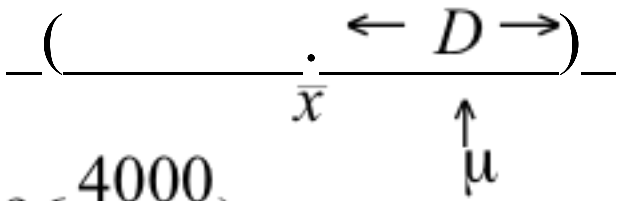
假若已知：□ = 4000 ¥

希望抽样误差 D= $|\bar{x}-\mu| \le 500$

并且要求置信度为（1- □）= 0.95

问：<u>样本容量应该多大？</u>

95% C. I.is



$$(\bar{x}-1.96\frac{4000}{\sqrt{n}},\bar{x}+1.96\frac{4000}{\sqrt{n}})$$

要求 $D = \left|\bar{x}-\mu\right| \leq 500$

则:

$$D = 1.96\frac{4000}{\sqrt{n}} \leq 500$$

$$n = \frac{1.96^2(4000)^2}{500^2} = 245.86$$

样本容量应不少于 246 人。

1、估计总体均值时需要的样本容量

放回抽样

在构造总体均值 μ 的置信度为 $100(1-\alpha)$ 的置信区间时 **（总体方差已知）**

$$(\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

置信区间的半长 D 等于

$$D=\frac{z_{\alpha/2}\sigma}{\sqrt{n}}\Rightarrow\sqrt{n}=\frac{z_{\alpha/2}\sigma}{D}\qquad n=\frac{z_{\alpha/2}^{2}\sigma^{2}}{D^{2}}$$

例题： $n = \dfrac{z_{\alpha/2}^{2}\sigma^{2}}{D^{2}}$

某厨具代理商欲了解其长期用户每月平均购买支出额。**问至少要抽取多大容量的样本**，才能使样本均值与总体均值的绝对误差在置信度不低于95%的条件下小于1?

问题1.总体标准差 σ 在抽样之前未知！

问题2. 在未确定样本容量 n 之前，无法计算样本标准差！

预抽样：

先在该公司固定用户中随机抽取 n=30 的样本， 经计算得到：$\bar{x}=110$, $s=13.12$

95% C.I.

$$(110-1.96\frac{13.12}{\sqrt{30}},\ 110+1.96\frac{13.2}{\sqrt{30}})$$

$$=(110-4.7,\ 110+4.7)$$

精度不够（要求误差为110□1）： $D=1$

$$n=\left(\frac{1.96\times 13.12}{1}\right)^2=661$$

如何确定调查所需要的精度 D

$$n = \frac{4s^2}{D^2} \qquad \bar{x} = 100, \quad D = 10$$
$$\bar{x} = 1000, \quad D = 10$$

应用时，由于存在量纲问题，可以采用相对误差：

$$\frac{D}{\bar{x}} = r \quad \Rightarrow \quad \boxed{D = r \cdot \bar{x}}$$

$$\bar{x} = 100, \quad r = 5\%, \quad D = r \times \bar{x} = 5$$
$$\bar{x} = 1000, \quad r = 5\%, \quad D = r \times \bar{x} = 50$$

所以常用的方法是： $$n = \frac{4s^2}{(r \cdot \bar{x})^2}$$

不放回抽样

置信区间：

$$\left(\bar{x}-t_{\alpha/2}\sqrt{\frac{N-n}{N-1}}\cdot\frac{s}{\sqrt{n}},\ \bar{x}+t_{\alpha/2}\sqrt{\frac{N-n}{N-1}}\cdot\frac{s}{\sqrt{n}}\right)$$

抽样误差范围：

$$D=\sqrt{\frac{N-n}{N-1}}\cdot\frac{t_{\alpha/2}s}{\sqrt{n}}$$

要求样本容量为：

$$n\approx\frac{n_0}{1+n_0/N}<n_0$$

例：　　假如固定用户：$N = 2000$

$$n_0 = \left(\frac{1.96\times 13.12}{1}\right)^2 = 661$$

$$n \approx \frac{n_0}{1+{n_0}/{N}} = \frac{661}{1+661/2000} = 496.8 \approx 497$$

注： 有时为计算方便起见，常取简单随机抽样所需要的样本容量代替n。这是一种保守的做法，但计算简单，在实际调查中经常使用。

2、估计总体比率时需要的样本容量

放回抽样

置信度为 (1-□)，总体比率 p 的置信区间为

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n},\ \hat{p} + z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}\right)$$

置信区间的宽度为

$$D = z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} \Rightarrow \sqrt{n} = \frac{z_{\alpha/2}\sqrt{\hat{p}\hat{q}}}{D}$$

样本容量为
$$n = \frac{z_{\alpha/2}^{2}\hat{p}\hat{q}}{D^{2}}$$

问题：在调查之前 $\hat{p}$ 是未知的

解决的办法：

$\hat{p}$	1	0.1	0.2	0.3	0.4	0.5	0.6	……
$1-\hat{p}$	0	0.9	0.8	0.7	0.6	0.5	0.4	……

取 $\hat{p}=0.5$ $1-\hat{p}=0.5$

所以样本容量 n 的最大值是：

$$n=\frac{0.25z_{\alpha/2}^{2}}{D^{2}}$$

注意教材P179：记置信区间长度 d=2D

$$n=\frac{\frac{1}{4}z_{\alpha/2}^{2}}{\left(\frac{1}{2}d\right)^{2}}=\left(\frac{z_{\alpha/2}}{d}\right)^{2}$$

教材P180： $(1-\alpha)=0.95 \qquad z_{\alpha/2}=1.96$

$$n=\left(\frac{z_{\alpha/2}}{d}\right)^2$$

d	0.14	0.12	0.1	0.08	0.06	0.04	0.02	0.01
n	196	266.78	384.16	600.25	1067.11	2401	9604	38416

向上取整：

d	0.14	0.12	0.1	0.08	0.06	0.04	0.02	0.01
n	196	267	385	601	1068	2401	9604	38416

教材3.4 （P180） 2009年3月，有政协委员提出恢复繁体字的提案。为了广泛了解民意，需要对该提案的支持率进行估计。

（1）要求置信度位0.95，置信区间长度$\boxed{\text{不超过0.01}}$，应抽取多少人?

抽样误差为0.5%

（2）如果随机抽取了4万人，其中有5600人支持该提案，计算支持率的置信区间，置信度为0.95。

解：

$$n=\left(\frac{1.96}{0.01}\right)^2=38416$$

放回抽样，至少抽取38416人

（1）

$$\hat{p}=\frac{5600}{40000}=0.14$$

14%的人表示赞同

(2)

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-p)}{n}}=1.96\times\sqrt{\frac{0.14\times0.86}{40000}}=0.0034$$

置信区间: $[0.1366,\ 0.1434]$

抽样误差为0.34%

例题：

北京地区观众调查网的置信度要求90%,误差要求不超过3%。求所需要的样本容量。

解： $(1-\alpha)=0.90$, $z_{\alpha/2}=1.65$, $D=0.03$

$$n=\frac{0.25\times1.65^2}{0.03^2}=756(人)$$

不放回抽样：

$$n\approx\frac{n_0}{1+{n_0}/{N}}<n_0$$

7.3 简单随机抽样总体总值的估计

1. 例题：

某工厂欲了解工人由于停工待料及机器故障所造成的每周工时损失。全厂共有750人。从中抽取50个工人进行调查，得到每个工人平均每周的工时损失数为 $\bar{y}=10.31$ 小时，且 $s^2=2.25$。估计**全厂由于停工待料及机器故障造成的工时损失数**。(α=0.05)

知 $N=750$, $n=50$, $\bar{y}=10.31$, $s^2=2.25$

求 估 Y

2. 点估计方法

计算公式：

$$\hat{Y} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

问题：为什么要先求样本均值 $\bar{y}$，再求 $\hat{Y} = N\bar{y}$？

为什么不直接用公式：$\hat{Y} = \sum_{i=1}^{n} y_i$

答案：（1）$\sum_{i=1}^{n} y_i \neq \sum_{i=1}^{N} y_i$

（2） 样本均值的波动小于个别观测值yi的波动。$D(\bar{y}) = \sigma^2/n$

例如，我们很可能从总体中抽取一个身高1.80的个体，但却不可能抽取一个身高平均值为 $\bar{y} = 1.80$ 的10个人的样本。在样本中，高、中、矮个子互相平均后，对总体的概括性更强。

点估计：

$$\hat{Y} = N\bar{y}$$

区间估计：

$$\hat{Y} \pm t_{\alpha/2}\sqrt{D(\hat{Y})}$$

$$D(\hat{Y}) = D(N\bar{y}) = N^2 D(\bar{y}) \begin{cases} N^2 \cdot \dfrac{\sigma^2}{n} & (\text{[illegible]} \\ N^2 \cdot \dfrac{N-n}{N-1} \cdot \dfrac{\sigma^2}{n} & (\text{[illegible]} \end{cases}$$

由此可见，总值估计的抽样误差要比均值估计的抽样误差**扩大 N 倍**。**但是相对误差不变**。

例题:

$N = 750,\quad n = 50,\quad \bar{y} = 10.31,\quad s^2 = 2.25$

求 体 Y

$$\hat{Y} = N\bar{y} = 750 \times 10.31 = 7732.5$$ (小

$$D(\hat{Y}) = N^2 \frac{N-n}{N-1} \cdot \frac{s^2}{n} = 750^2 \times \frac{750-50}{750-1} \times \frac{2.25}{50} = 24470.52$$

$(1-\alpha$

$$7732.5 \pm 1.96 \times \sqrt{24470.52} = 7732.5 \pm 1.96 \times 156.43$$

$$= (7426.20 \quad 8039.40$$

$$\frac{1.96 \times 156.43}{7732.5} = 0.04$$

7.4 系统抽样

又称“**等距抽样**”或“**机械抽样**”

特点：组织形式简单：不需要在抽样前对每一个单位进行编号。只要确定抽样起点和间隔，就可以确定整个样本单位。

（1）按照无关标志排队，按间隔抽取

例如：调查某企业职工收入时，按照姓氏比画排列职工名单，进行抽样。显然，职工工资与姓氏比画之间没有必然联系；

（2）按照有关标志排队，按间隔抽取

例如：进行农产量调查时，将总体单位按照上一年度的产量高低排序。这样，可以使标志值高低不同的单位均进入样本，样本单位在总体中分布均匀，抽样误差较小。

（3）按照自然位置顺序排列，按间隔抽取

例如：工业产品检验时，按照生产时间顺序，每间隔一定时间抽取一定数量的样本；检验一打发票时，可以按照顺序，每隔10张抽取1张；在估计果园的产量时，每隔7株抽取1株。

方法：随机起点，等距抽取。

（1）按照某种顺序给总体中的N个单元排列编号；

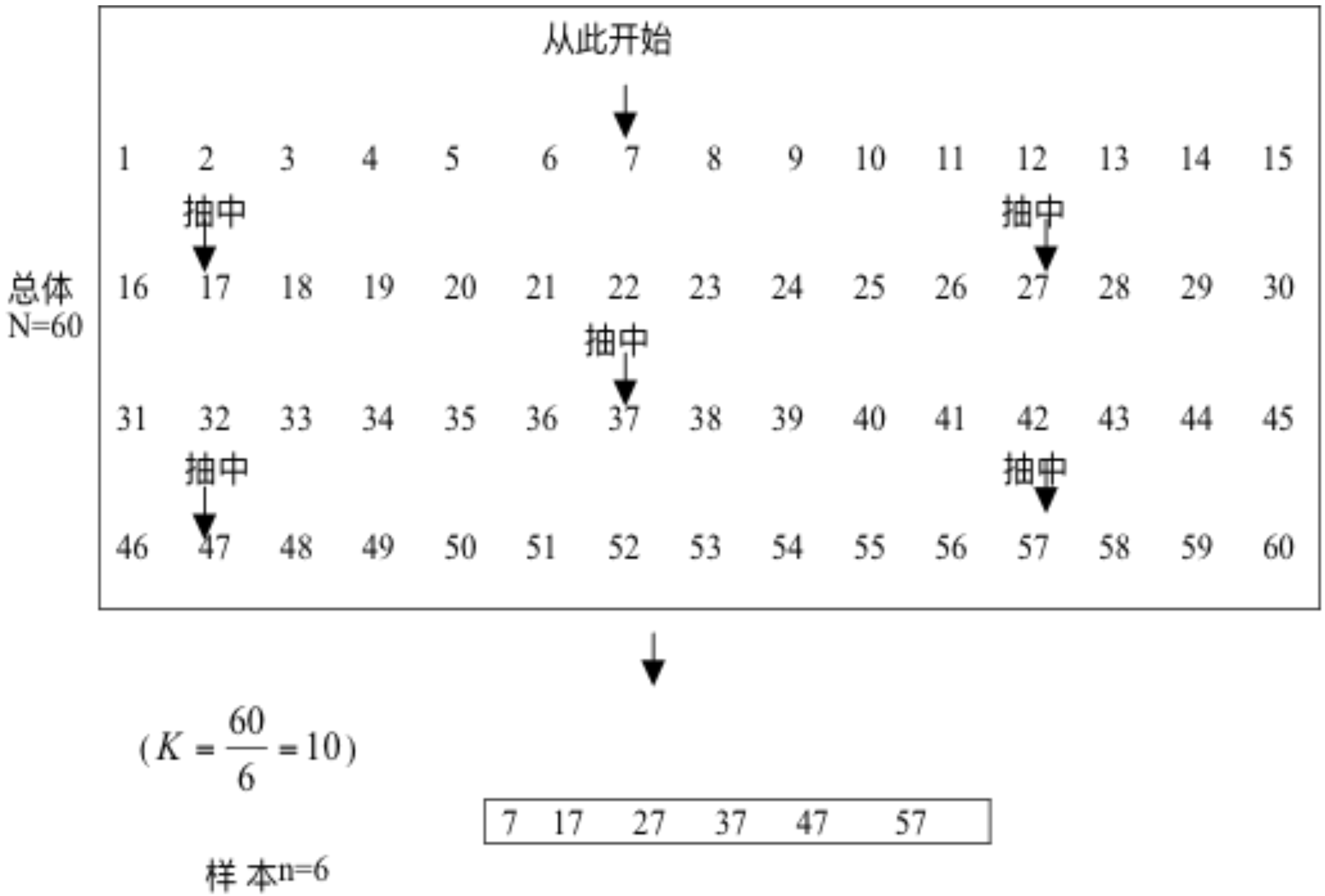
（2）按照随机数表，随机抽取一个编号 i 作为样本的第一个单元；

（3）计算间距：

$$k=\left[\frac{N}{n}\right]$$

(4)起始的样本点编号选取$1\sim k$之间的随机数。然后依次抽取编号如下的 n 个单元作为样本点。

$$i,i+k,i+2k,\boxed{?},i+(n-1)k$$


从此开始
总体
N=60
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
抽中
抽中
抽中
抽中
抽中
$(K=\frac{60}{6}=10)$
7 17 27 37 47 57
样 本n=6

例如：中央电视台在建立收视率调查网时，要在某居委会拥有电视的512户中抽取5个样本户。

$$N=512,\quad n=5,\quad k=\frac{512}{5}\approx 102$$

在[0,512]中任意确定一个三位数，例如是071。则被抽中的5户为：

71，173，275，377，479

抽样误差的大小与总体单位的排列顺序有关：

（1）如果总体中所有单元的排列编号是随机的，并且 n 比N小得多的话，那么等距抽样的精度和简单随机抽样的精度是十分相近的。

（例如，按照姓氏比画或按照行政单位编号排序。）

（2）如果总体单元是按照某个与调查项目有关的变量的大小排序，由于等距抽样的样本点分布更加均匀，则等距抽样的精度将高于简单随机抽样。

（例如，调查机械加工企业的工业增加值时，以用电量排序。）

（3）如果总体各单位的标志值存在周期变化趋势，而循环周期恰好等于抽样间隔，则等距抽样的精度低于简单随机抽样。

1,2,3,4,5,6; 1,2,3,4,5,6; 1,2,3,4,5,6; 1,2,3,4,5,6; 1,2,3,4,5,6

7.5　分层随机抽样

一、分层抽样方法

例如：

（1）对北航学生的研究能力进行抽样测试。学生层次有：专科、本科、研究生、博士、博士后。

（2）对央行的某项政策意见进行调查。可以根据调查内容分层：不同的职务层次， 或者不同的部门、不同地区。

在所调查的指标上，各层的相似程度高，而且层间差异大

分层的原则：

例如 TBT影响调查：按照36个地区进行分层？ （行政管理力度大）

按照[illegible]类出口地区分层？ （[illegible]情况类似）

分层抽样的特点：

采用分层抽样，使每一层内的差异大大缩小，而每一个样本单位对各层均有较高的代表性。

- 利用已知信息，提高抽样调查的精度；
- 便于组织实施；
- 在调查中，除了得到总体的有关信息外，还可以得到一些子总体的信息.

同样的样本容量下，分层抽样的抽样误差更小。

应用. TBT影响调查的分层方法： — 按照产品分层

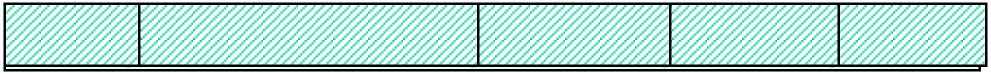
— 按照地区管理

二. 总体均值的估计

例：

对某市600个个体商户的月零售额进行抽样调查，现申报资金分为大、中、小三类，根据调查结果的数据整理如下表。试估计该市个体户的平均月零售额，并以95%的可靠性作出区间估计。

层次	N_i	n_i	$\bar{y}_i$	s_i^2
大	60	30	20	16
中	240	40	8	4
小	300	40	1	0.5

$\hat{\bar{Y}} = \frac{1}{N}$ 

计算方法:

(1) 第 i [illegible] $\qquad \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}\ , \quad i = 1,2,\boxed{?},r$

(2)第 i [illegible] $\qquad \hat{Y}_i = N_i \bar{y}_i$

(3)[illegible] $\qquad \hat{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{r} N_i \bar{y}_i = \sum_{i=1}^{r} W_i \bar{y}_i$

[illegible] $\qquad W_i = N_i / N$

总体均值 = 各层均值的加权和

方差估计：

$$\hat{\bar{Y}} = \sum_{i=1}^{r} W_i \bar{y}_i$$

（1）放回抽样

$$D(\hat{\bar{Y}}) = \sum_{i=1}^{r} W_i^2 D(\bar{y}_i) = \sum_{i=1}^{r} W_i^2 \left(\sigma_i^2 / n_i \right)$$

（2）不放回抽样

$$D(\hat{\bar{Y}}) = \sum_{i=1}^{r} W_i^2 \cdot \frac{\sigma_i^2}{n_i} \cdot \frac{N_i - n_i}{N_i - 1} \approx \sum_{i=1}^{r} W_i^2 \cdot \frac{s_i^2}{n_i} (1 - f_i)$$

抽样比

其 $f_i = \dfrac{n_i}{N_i}$， $\sigma_i^2 \approx s_i^2$

例题：某市个体商户的月零售额的抽样调查

$N=600$

$N1=60$

$n1=30$

$N2=240$

$n2=40$

$N3=300$

$n3=40$

$$f_1=\frac{n_1}{N_1}=\frac{30}{60}=0.5\ ,\quad \bar{y}_1=20\qquad s_1^2=16$$

$$f_2=\frac{n_2}{N_2}=\frac{40}{240}=0.17\ ,\quad \bar{y}_1=8\qquad s_1^2=4$$

$$f_3=\frac{n_3}{N_3}=\frac{40}{300}=0.13\ ,\quad \bar{y}_3=1\qquad s_3^2=0.5$$

$$\hat{\bar{Y}}=\frac{1}{N}\sum_{i=1}^{3}N_i\bar{y}_i=\frac{1}{600}\left[60\times20+240\times8+300\times1\right]=5.7$$

$$D(\hat{\bar{Y}})=\sum_{i=1}^{r}W_i^2\cdot\frac{\sigma_i^2}{n_i}(1-f_i)$$

$$\left[\left(\frac{60}{600}\right)^2\frac{16}{30}(1-0.5)+\left(\frac{240}{600}\right)^2\frac{4}{40}(1-0.17)+\left(\frac{300}{600}\right)^2\frac{0.5}{40}(1-0.13)\right]$$

$$=0.0187$$

区 $\quad 5.7\pm1.96\times\sqrt{0.0187}\Rightarrow(5.43,5.97)$

$0.268/5.7=0.047$

三. 样本数目在层间的分配

问题： 总的样本容量为 n，总体分为 r 层。

每一层的样本容量应为多大?

（一）等比例分层抽样

1. 分配方案计算方法 I

在任意一层中，样本容量所占的比例都相同。

[illegible] $N = N_1 + N_2 + \boxed{?} N_r$

[illegible] n

记 $f = \dfrac{n}{N}$

[illegible] i[illegible] $n_i = f \cdot N_i$

例：N=1000，　N1=600，　N2=200，　N3=200

要抽取容量为 n=200 的样本，问每一层应抽取多少个体?

解：

$$f=\frac{n}{N}=\frac{200}{1000}=0.2 \qquad f_i=f\ ,\ \ i=1,2,3$$

因

$$n_1=0.2\times 600=120$$

$$n_2=0.2\times 200=40$$

$$n_1=0.2\times 200=40$$

2.分配方案计算方法 II

记 $W_i = \dfrac{N_i}{N}$ ， $i = 1,2,\boxed{?},r$

则 $\boxed{n_i = W_i \cdot n}$

$$\frac{N_i}{N} = W_i = \frac{n_i}{n}$$

$$n_i = \frac{n}{N} \cdot N_i = f \cdot N_i$$

例：N=1000, $N1$=600, $N2$=200, $N3$=200

$$W_1 = 0.6,\quad W_2 = 0.2,\quad W_3 = 0.2$$

$n = 200$:

$$n_1 = 200 \times 0.6 = 120$$

$$n_2 = 200 \times 0.2 = 40$$

$$n_3 = 200 \times 0.2 = 40$$

3. 等比例分层抽样，总体均值的估计量

点估计：

$$\hat{\bar{Y}} = \sum_{i=1}^{r} W_i \cdot \bar{y}_i$$

区间估计：

（1）放回抽样的方差

$$D(\hat{\bar{Y}}) = \sum_{i=1}^{r} W_i^2 \frac{\sigma_i^2}{n_i} = \sum_{i=1}^{r} W_i \cdot \frac{\cancel{n_i}}{n} \cdot \frac{\sigma_i^2}{\cancel{n_i}}$$

$$= \frac{1}{n} \boxed{\sum_{i=1}^{r} W_i \sigma_i^2} = \bar{\sigma}^2 / n$$

其中，$\bar{\sigma}^2$ 表示平均层内方差。

（2）不放回抽样的方差

$$D(\hat{Y}) = \sum_{i=1}^{r} W_i^2 \frac{\sigma_i^2}{n_i}(1-f_i) = (1-f)\sum_{i=1}^{r} W_i \cdot \frac{\cancel{n_i}}{n} \cdot \frac{\sigma_i^2}{\cancel{n_i}}$$

$$= \frac{1}{n}(1-f)\sum_{i=1}^{r} W_i \sigma_i^2 = (1-f)\overline{\sigma^2}/n$$

由于各层内的单元变化程度比较小，分层后有

$$\sigma_i^2 < \sigma^2$$

$$\overline{\sigma^2} = \sum_{i=1}^{r} W_i \sigma_i^2 < \sum_{i=1}^{r} W_i \sigma^2 = \sigma^2 \sum_{i=1}^{r} W_i = \sigma^2$$

因此，同样的样本容量下，分层抽样的抽样误差更

四. 总体比例的估计

$$\hat{p} = \sum_{i=1}^{r} W_i \hat{p}_i$$

不放回抽样

$$D(\hat{p}) = \sum_{i=1}^{r} W_i^2 D(\hat{p}_i) = \sum_{i=1}^{r} W_i^2 \frac{N_i - n_i}{N_i - 1} \cdot \frac{p_i(1-p_i)}{n_i}$$

$$= \sum_{i=1}^{r} W_i^2 (1 - f_i) \cdot \frac{p_i(1-p_i)}{n_i}$$

例题： $\hat{p}=\sum_{i=1}^{r}W_i\hat{p}_i;$ $D(\hat{p})=\sum_{i=1}^{r}W_i^2(1-f_i)\cdot\frac{p_i(1-p_i)}{n_i}$

某广告公司要了解电视广告的作用，拟在有关对象中调查看电视广告的比例。设对象分为三层：

$$N1=155,\ N2=62,\ N3=93,$$

样本容量为40。**采用等比例分层抽样**，调查结果为：第一层看电视广告的比例为0.8，第二层的比例为0.25，第三层的比例为0.5。试以95%的可靠性，估计调查对象中收看电视广告比例的置信区间。

N=155 + 62 + 93 =310 $\qquad f = 40/310 = 0.129$

n =40

$$W_1 = \frac{155}{310} = 0.5 \quad , \quad n_1 = 0.5 \times 40 = 20$$

$$W_2 = \frac{62}{310} = 0.2 \quad , \quad n_2 = 0.2 \times 40 = 8$$

$$W_3 = \frac{93}{310} = 0.3 \quad , \quad n_3 = 0.3 \times 40 = 12$$

由调查结果：

$\hat{p}_i = 0.8 \qquad \hat{p}_i = 0.25 \qquad \hat{p}_i = 0.5$

则 $\qquad \hat{p} = 0.5 \times 0.8 + 0.2 \times 0.25 + 0.3 \times 0.5 = 0.6$

方差估计：

$$D(\hat{p})=\sum_{i=1}^{3}W_i^2(1-f)\frac{\hat{p}_i(1-p_i)}{n_i}$$

$$=0.5^2(1-0.129)\frac{0.8\times0.2}{20}+0.2^2(1-0.129)\frac{0.25\times0.75}{8}$$

$$+0.2^2\times(1-0.129)\frac{0.5\times0.5}{12}=0.0042$$

$$s=\sqrt{0.0042}=0.065$$

所 95%

$$0.6\pm1.96\times0.065=0.6\pm0.1274$$

从总体看，观看广告的比例约为 60%，估计误差约为 $\pm13\%$，估计的可靠性为95%。

7.6　抽样调查的误差来源

调查误差 = 抽样误差 + 非抽样误差

抽样误差：由于抽选样本的随机性而产生的误差

（由于概率抽样方式不同所造成，是可以估计的）

非抽样误差：除抽样误差外，由其他各种原因而引起的误差。

产生非抽样误差的主要原因：

（1）抽样框误差：目标总体不等于抽样总体，如遗漏了有关单位，或包含了非目标单位；观测之间的复合连接；分层方案设计不当等。

（2）无应答误差：受调查人有意识不合作；无意识（由于客观原因无法接受调查，填写问卷时粗心）；

（3）计量误差：问卷设计不合理、调查指标含义不清、计量单位不标准，选择的统计量和推算方法不适当等。

案例1：《文学摘要》民意测验

抽 ≠

1936年美国总统选举

F.D. Roosevelt (罗斯福）任美国总统的第一任期届满(民主党)

A. Landon (兰登）Kansas州州长(共和党)

经济背景：国家正努力从大萧条中恢复，失业人数高达九百万人。

The literary Digest《文学摘要》进行民意测验，将问卷邮寄给一千万人，他们的名字和地址摘自电话簿或俱乐部会员名册。其中240万人寄回答案（回收率24%）。

预测结果：Roosevelt 43%，Landon 57%

竞选结果：Roosevelt 62%，Landon 38%

主要原因：

选择偏倚——将一类人排除在样本框之外（当时四个家庭中，只有一家安装电话）

不回答偏倚——低收入和高收入的人倾向不回答

1936年美国总统竞选（Gallup的预测）

- 样本容量3000人，在《摘要》公布其预测结果之前，仅以一个百分位数的误差预言了《摘要》的预测结果。

- 利用一个约5万人的样本，正确地预测了Roosevelt的胜利。

	Roosevelt的百分数
盖洛普预言《摘要》的预测结果	44
《摘要》预测的选举结果	43
盖洛普预测的选举结果	56
选举结果	62

方法：从《摘要》要用的名单中随机选取3000人，并给他们每人寄去一张明信片，询问他们打算怎样投票。

大样本并不能防止偏倚： 当抽样框不正确时，抽取一个大的样本并无帮助，它只不过是在较大的规模下，去重复基本错误。

Gallup1936~1948年采用定额抽样

定额抽样：样本被精心挑选，以使在某些关键特征上与总体相似。 **在规定定额内，访问人员可以自由选取任何人。**

例如：在 St. Louis 的访问人员访问13个对象，并规定其中

- 6人住在近郊，7人住在市中心；
- 男人7名，女人6名；
- 在男人中，3人40岁以下，4人40岁以上；1名黑人，6名白人。
- 6名白人支付的月租：1人支付的金额不少于44.01$

  3人支付的金额为18.01~ 44.00 $

  2人支付的金额不超过18.00 $

年份	预测共和党得票	共和党实际得票	有利于共和党的偏差
1936	44	38	6
1940	48	45	3
1944	48	46	2

Gallup民意测验在1948年后总统选举中的记录

（随机抽样：访问员无任何自主处理的权利）

年份	样本容量	获胜候选人	预测值	选举结果	误差
1952	5385	艾森豪威尔	51.0%	55.4%	+4.4%
1956	8144	艾森豪威尔	59.5%	57.8%	-1.7%
1960	8015	肯尼迪	51.0%	50.1%	-0.9%
1964	6625	约翰逊	64.0%	61.3%	-2.7%
1968	4414	尼克松	43.0%	43.5%	-0.5%
1972	3689	尼克松	62.0%	61.8%	-0.2%
1976	3439	卡特	49.5%	51.1%	+1.6%
1980	3500	里根	55.3%	51.6%	-3.7%
1984	3456	里根	59.0%	59.2%	-0.2%
1988	4089	布什	56.0%	53.9%	-2.1%

案例2　可口可乐问卷设计失败

问题与思考：

20世纪80年代，美国可口可乐公司耗资500万美元，进行了历时2年的市场调查，调查了近20万名消费者。决定放弃传统配方，推出一代新的可口可乐。却几乎产生灾难性的后果。

可口可乐发展将近百年。但在20世纪80年代，它的市场销售增长率从平均每年13%猛降到2%。市场占有率从曾是百事可乐的2倍，变成只领先2.9个百分点。

市场调查与决策：

（1）出动2000名调查员，在10个主要城市调查消费者的口味。**问卷的主要问题是：“如果在可口可乐配方中增加一种新的成分，使它喝起来更柔和，您愿意吗？**结果有一多半的人表示接受，只有11%的人表示不安。

（2）公司投资400万美元进行大规模的口味尝试活动。13个大城市的19.1万消费者参与口味尝试活动。在众多口味饮料中，消费者对新口味可乐青睐有加。55%的品尝者认为新口味超过传统配方。**结论：立即生产新可乐。**

结果：

新饮料上市4个小时，可口可乐公司接到650个抗议电话。10天后，每天接到5000多个抗议电话。更有雪片似的抗议信件。有人甚至说要改喝茶水来代替可乐。公司不得不开辟83个热线，雇佣大量的公关人员来处理这些抱怨和抗议。

3个月以后，市场调研表明，只有不到30%的消费者说新可乐的好话了。愤怒的情绪在美国蔓延。社会学家认为，可口可乐公司把一个神圣的象征毁掉了。

罗伯特.戈伊朱埃塔不得不率领公司全体高层管理者站在可口可乐的标志下，向公众道歉，并宣布立即恢复传统配方生产。全国一片沸腾。有议员在参议会回上发表演说："这是美国历史上一个非常有意义的时刻，它表明有些民族精神使不可更改的。"

问题的根源是什么？

耗资巨大、范围广泛、被调查者反映良好

（决策是：放弃老饮料）

其他案例：调查中的非抽样误差

1、分层抽样方案设计不当，造成选择偏倚：按产品分层（样本分配原则是出口额高的产品多抽；对于一个产品，根据其出口额在全国各地分布分配样本。）

问题：一些出口总额小的地区会不能入样。

2、样本点之间的复合连接，造成重复统计

例如：企业类型（生产型企业、流通型企业）

3、抽样框中包含非目标单位：若以上年企业出口额作为抽样依据；但该企业的受调查产品当年没有出口。减少有效样本数量

4、避免调查表中内容的歧异："所调查的产品" → "本问卷所调查的产品"

"进口国" → "贸易对象国"；

5、加强调查人员的责任意识：采取登记制度和汇总结果的报告制度。

抽样调查作业

采用抽样调查方法，估计全班同学的平均身高

1、首先：计算总体均值和方差 （留做参考）

2、预抽样（n=30)：估计样本均值与方差

3、选取抽样的相对误差 5%或10%

4、计算样本容量 $n_0 = \dfrac{4s^2}{(r \cdot \bar{x})^2}$ $n \approx \dfrac{n_0}{1 + n_0 / N} < n_0$

5、等距抽样：随机起点，等距抽取

6、给出点估计和区间估计 $\bar{x} \pm 2\sqrt{\dfrac{N-n}{N-1}}\dfrac{s}{\sqrt{n}}$

7、对比总体参数，对于你的分析过程和结论进行评价与思考

阅读与练习

分层抽样总体总值的估计

Excel 软件应用

一、分层抽样总体总值的估计

1. 点估计

$$\hat{Y} = N\hat{\bar{Y}} = N\sum_{i=1}^{r}\frac{N_i}{N}\bar{y}_i = \sum_{i=1}^{r}N_i\bar{y}_i$$

2. 区间估计

$$D(\hat{Y}) = N^2 D(\hat{\bar{Y}})$$

(1)[illegible]

$$D(\hat{Y}) = N^2 D(\hat{\bar{Y}}) = N^2\sum_{i=1}^{r}W_i^2\frac{\sigma_i^2}{n_i}$$

(2)[illegible]

$$D(\hat{Y}) = N^2\sum_{i=1}^{r}W_i^2\frac{\sigma_i^2}{n_i}\cdot\boxed{\frac{N_i - n_i}{N_i - 1}}$$

[illegible]

$$s_i^2 \approx \sigma_i^2$$

例题：某市个体户的月零售额的抽样调查，估计全市个体户**总的月销售额**。

根据前面计算：

$$N = 600 \qquad \hat{\bar{Y}} = 5.7 \qquad D(\hat{\bar{Y}}) = 0.0187$$

所以有：

$$\hat{Y} = N\hat{\bar{Y}} = 600 \times 5.7 = 3420 \quad (千$$

$$D(\hat{Y}) = N^2 D(\hat{\bar{Y}}) \quad = 600^2 \times 0.0187 = 6732$$

置

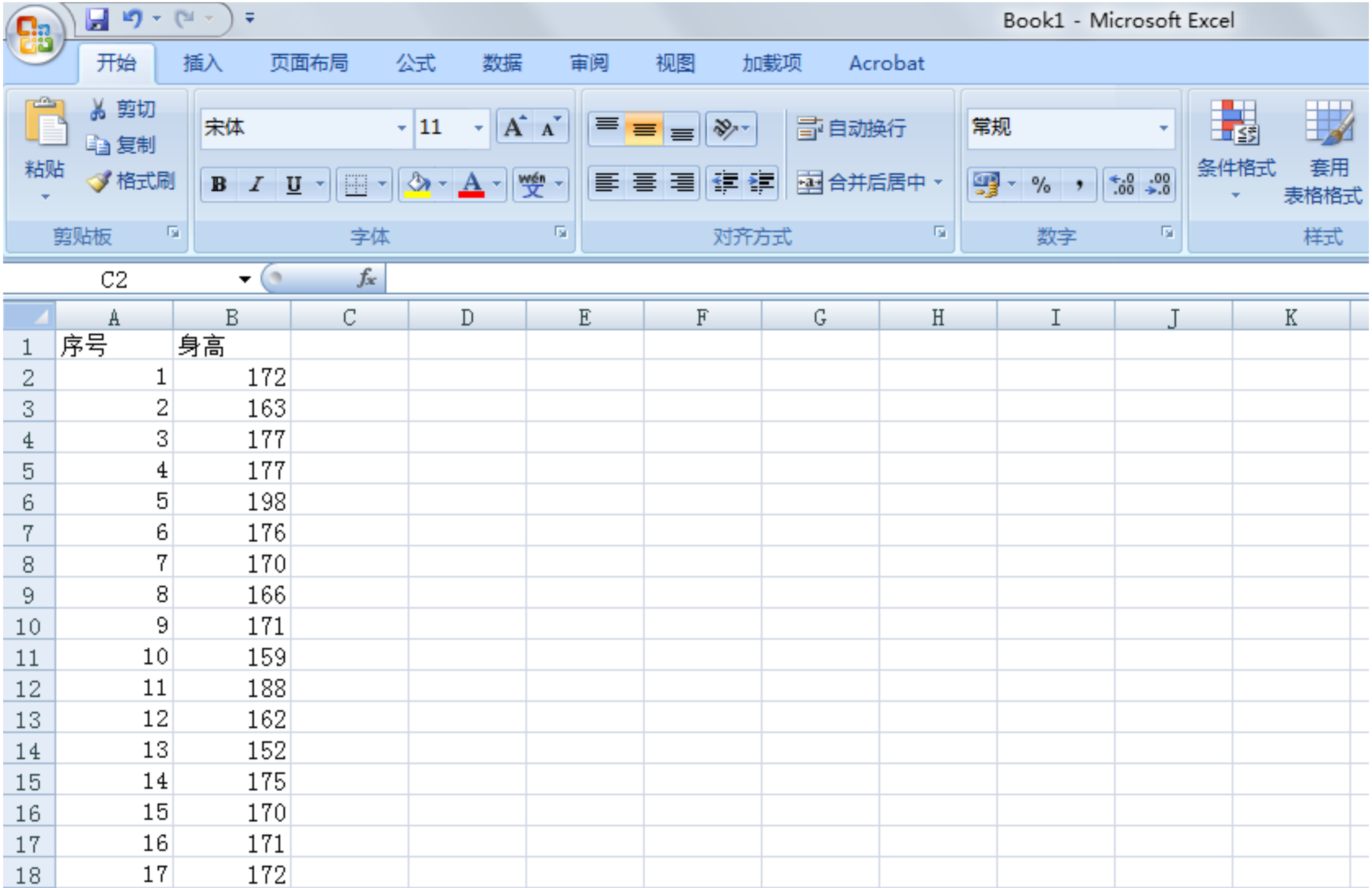
$$3420 \pm 1.96 \times \sqrt{6732} = 3420 \pm 160.8156$$

$$(3259.184, 3580.816)$$

160.8156/3420=0.047

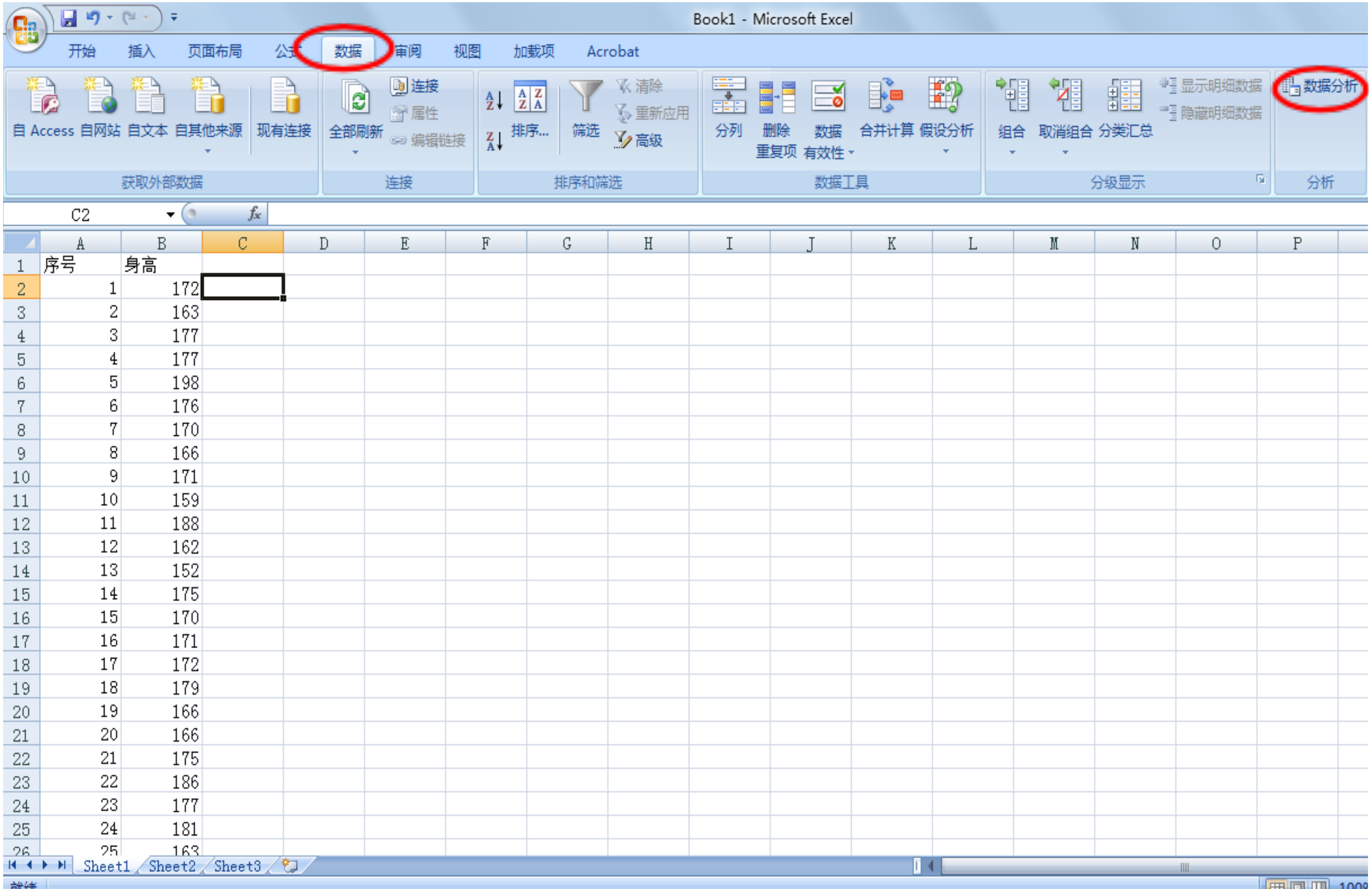
总值估计的与总体均值的相对误差不变。

二、随机抽取30个样本点的方法

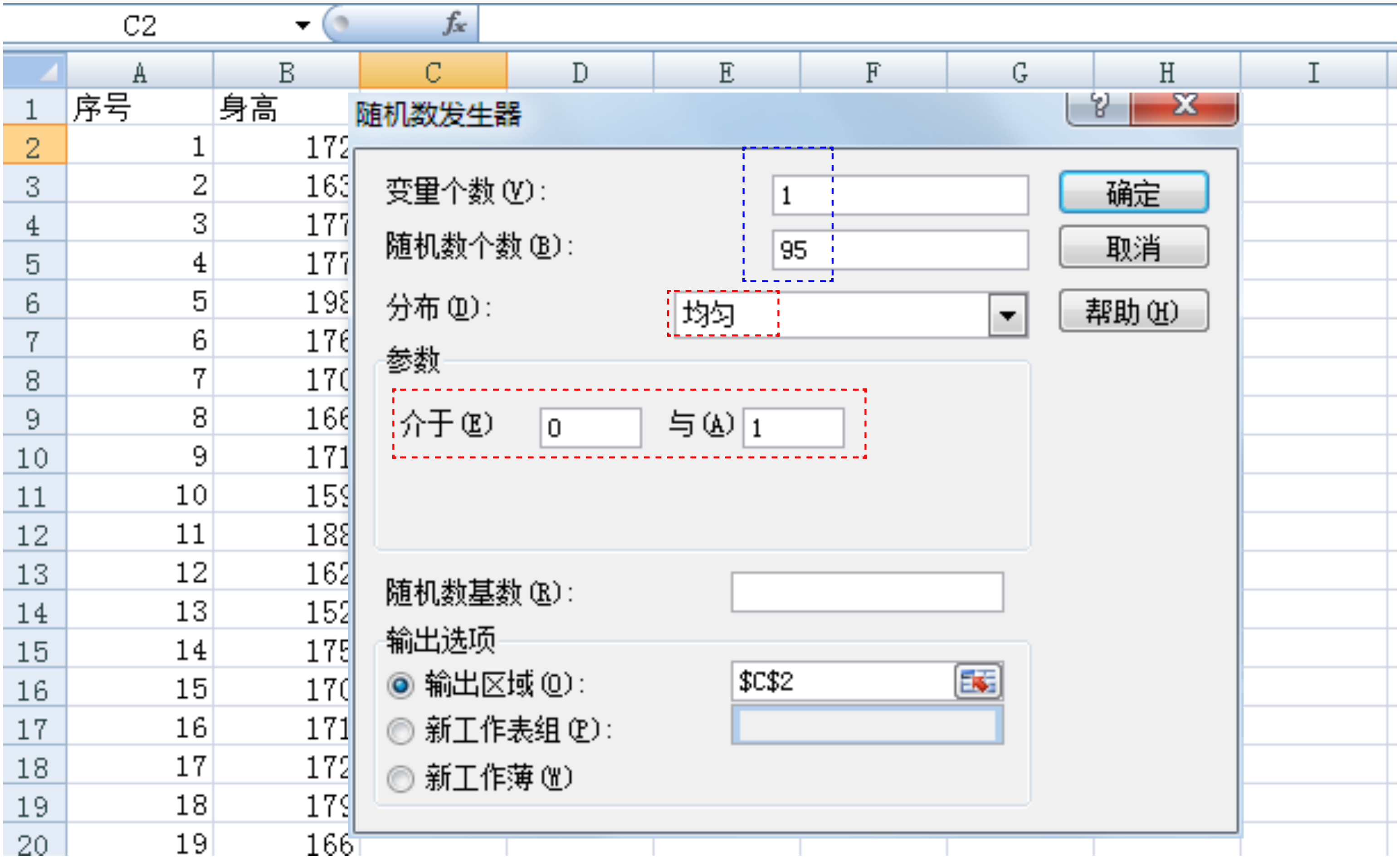


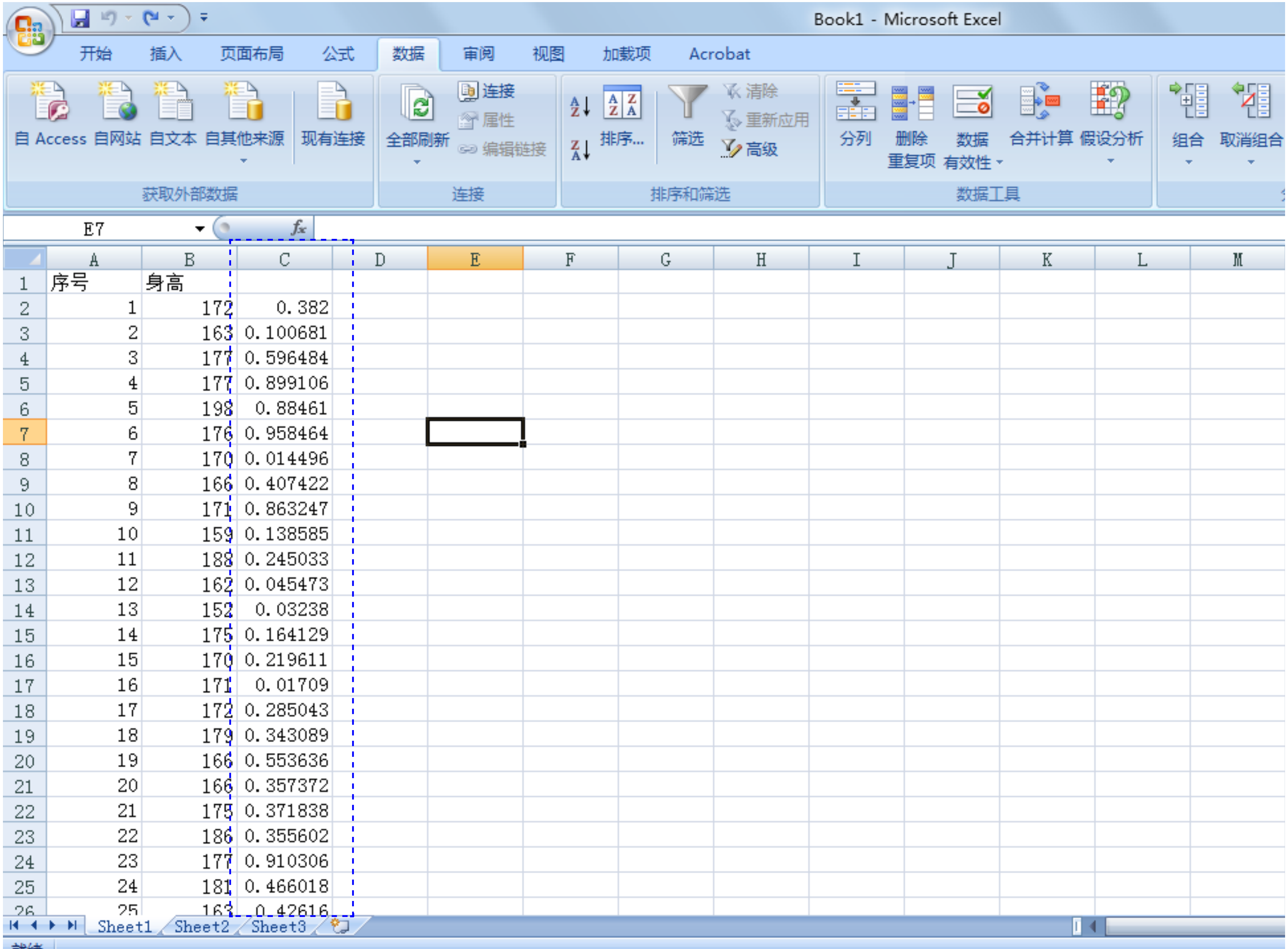
序号	身高
1	172
2	163
3	177
4	177
5	198
6	176
7	170
8	166
9	171
10	159
11	188
12	162
13	152
14	175
15	170
16	171
17	172

数据□数据分析

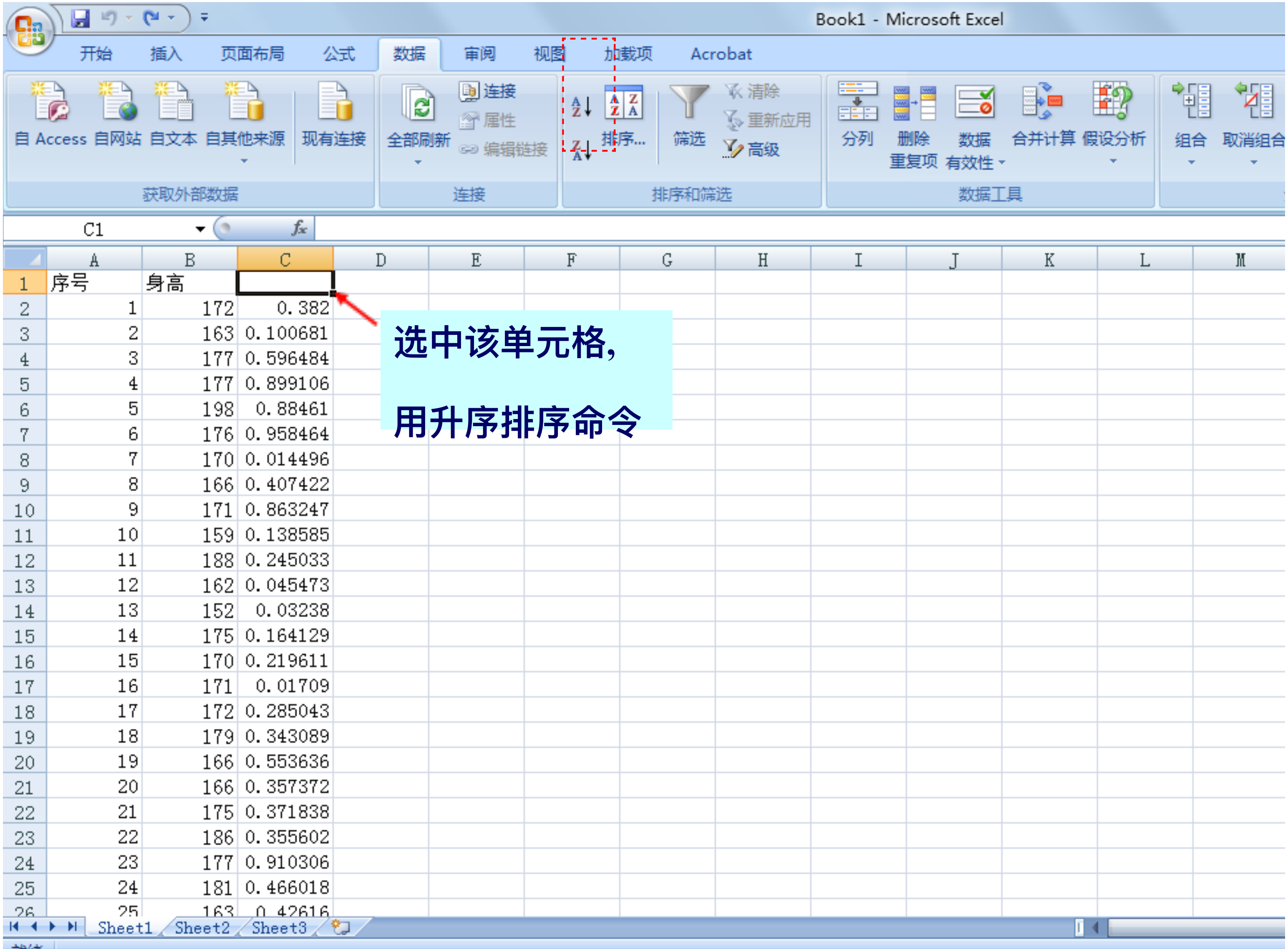

Book1 - Microsoft Excel
开始
插入
页面布局
公式
数据
审阅
视图
加载项
Acrobat
自 Access
自网站
自文本
自其他来源
现有连接
获取外部数据
全部刷新
连接
属性
编辑链接
连接
排序...
筛选
清除
重新应用
高级
排序和筛选
分列
删除重复项
数据有效性
合并计算
假设分析
数据工具
组合
取消组合
分类汇总
显示明细数据
隐藏明细数据
分级显示
数据分析
分析
C2
序号
身高
1 172
2 163
3 177
4 177
5 198
6 176
7 170
8 166
9 171
10 159
11 188
12 162
13 152
14 175
15 170
16 171
17 172
18 179
19 166
20 166
21 175
22 186
23 177
24 181
25 163
Sheet1
Sheet2
Sheet3
就绪

数据□数据分析□随机数发生器

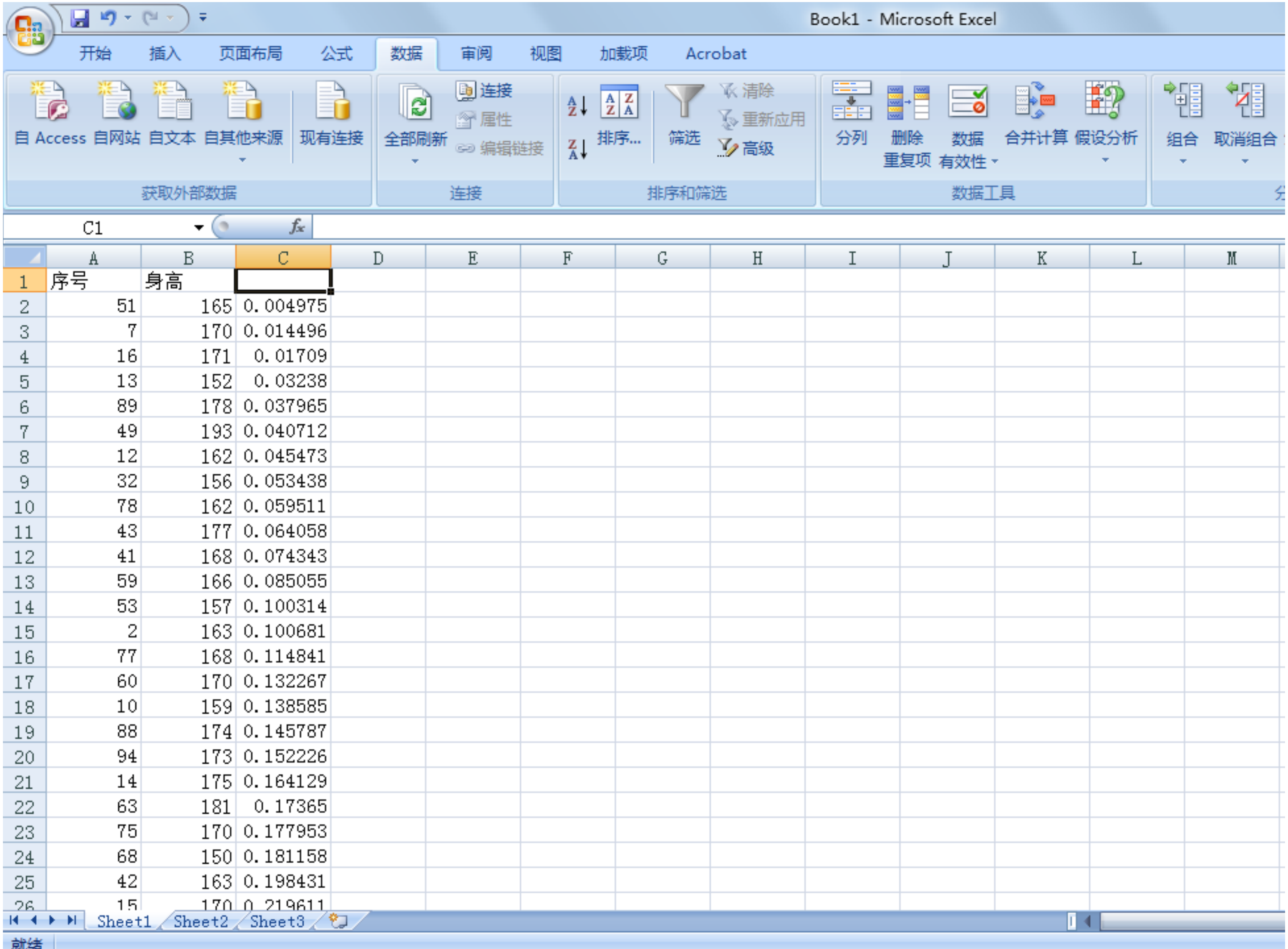




	A	B	C
1	序号	身高	
2	1	172	0.382
3	2	163	0.100681
4	3	177	0.596484
5	4	177	0.899106
6	5	198	0.88461
7	6	176	0.958464
8	7	170	0.014496
9	8	166	0.407422
10	9	171	0.863247
11	10	159	0.138585
12	11	188	0.245033
13	12	162	0.045473
14	13	152	0.03238
15	14	175	0.164129
16	15	170	0.219611
17	16	171	0.01709
18	17	172	0.285043
19	18	179	0.343089
20	19	166	0.553636
21	20	166	0.357372
22	21	175	0.371838
23	22	186	0.355602
24	23	177	0.910306
25	24	181	0.466018
26	25	163	0.42616



	A	B	C
1	序号	身高	
2	1	172	0.382
3	2	163	0.100681
4	3	177	0.596484
5	4	177	0.899106
6	5	198	0.88461
7	6	176	0.958464
8	7	170	0.014496
9	8	166	0.407422
10	9	171	0.863247
11	10	159	0.138585
12	11	188	0.245033
13	12	162	0.045473
14	13	152	0.03238
15	14	175	0.164129
16	15	170	0.219611
17	16	171	0.01709
18	17	172	0.285043
19	18	179	0.343089
20	19	166	0.553636
21	20	166	0.357372
22	21	175	0.371838
23	22	186	0.355602
24	23	177	0.910306
25	24	181	0.466018
26	25	163	0.42616



	A	B	C
1	序号	身高	
2	51	165	0.004975
3	7	170	0.014496
4	16	171	0.01709
5	13	152	0.03238
6	89	178	0.037965
7	49	193	0.040712
8	12	162	0.045473
9	32	156	0.053438
10	78	162	0.059511
11	43	177	0.064058
12	41	168	0.074343
13	59	166	0.085055
14	53	157	0.100314
15	2	163	0.100681
16	77	168	0.114841
17	60	170	0.132267
18	10	159	0.138585
19	88	174	0.145787
20	94	173	0.152226
21	14	175	0.164129
22	63	181	0.17365
23	75	170	0.177953
24	68	150	0.181158
25	42	163	0.198431
26	15	170	0.219611

作业

一、 教材《习题九》： 9.16题 （请注意对计算结果的解释）

二、《统计学》各章练习题： 7.13 题

三、假若要在1000个人的总体中抽取100人，调查对某种商品的接受程度（用 x=5 表示非常喜欢、x=1 表示非常不喜欢的5分评分制）。已知该在总体中，青年人、中年人、老年人所占比例分别为55%、30%和15%。

问：

（1）如果采用按比例无放回的分层抽样，应分别青年人、中年人、老年人中个抽取多少人?

（2）抽样调 以及抽样误差。

层次	青年人	中年人	老年人
平均分	4.80	2.70	3.60
层内方差	1.50	1.20	1.80