《抽样技术》习题解答

倪加勋　主编

中国统计出版社

《抽样技术》习题解答

倪加勋 主编

中国统计出版社

(京)新登字041号

《抽样技术》习题解答

CHOUYANG JISHU XITI JIEDA

中国统计出版社出版
(北京三里河月坛南街38号 100826)
新华书店北京发行所发行
北京通县永乐印刷厂印刷

787×1092毫米 32开本 8.75印张 18万字
1992年10月第1版 1992年10月北京第1次印刷
印数：1—3 000
ISBN 7-5037-0719-4/C·434
定价：7.80元

编写说明

W·G·科克伦的《抽样技术》一书是抽样调查理论方面比较具有权威性的著作，被国外很多大学选为研究生教材。张尧庭和吴辉二同志于1985年译成中译本并由中国统计出版社出版，已经过多次重印，在国内统计学界有较大的影响。该书每章后面均附有练习题，这些题目对于理解原书的内容有很大帮助。其中有些题目有一定难度，但该书最后只给出答案而没有解题过程，这样给自学该书的读者带来一些困难。为了满足这方面的需要，我们针对该书的练习题编写了习题的解答。

本书是在中国人民大学统计系88级研究生抽样技术课讨论班的基础上形成的，讨论班由倪加勋主持。参加本书起草的有胡忠兵、黎樟林、陈安贵、郭向军、杨玉华和范小昕，最后由胡忠兵和黎樟林进行整理，倪加勋负责审阅定稿。需说明的是：

1.《习题解答》的作用有二重性。一方面可以帮助读者更好地理解原书的内容，在读者本人解题遇到困难时从中得到启发，有助于进一步深入学习；但另一方面，一些习题要求读者自己认真思考才能得益，如果单纯依靠题解而不去认真思考，那就违背了本书的意图。

2.一个习题有时有多种解法，本书只是提供了一种解法供读者参考，这种解法不一定是最简便的，希望读者能在这

基础上作出更多更好的解题方法，那么本书就起到了抛砖引玉的作用。

3.由于编者水平有限，解题时难免有不严谨甚至有错误的地方，希望读者批评指正。

编　者

1991年8月

目　录

第1章　引言（共8个练习题）……（1）
第2章　简单随机抽样（共20个练习题）……（7）
第3章　抽样比例及百分比（共14个练习题）……（31）
第4章　样本含量的估计（共13个练习题）……（48）
第5章　分层随机抽样（共16个练习题）……（64）
第5A章　分层抽样的其他方面（共14个练习题）…（85）
第6章　比率估计量（共10个练习题）……（112）
第7章　回归估计量（共9个练习题）……（127）
第8章　等距抽样（共9个练习题）……（142）
第9章　单级整群抽样——当各群的大小相等时（共6个练习题）……（162）
第9A章　单级整群抽样——当各群的大小不相等时（共8个练习题）……（173）
第10章　抽子样本——当单位大小相等时（共10个练习题）……（192）
第11章　抽子样本——当单位大小不相等时（共11个练习题）……（205）
第12章　双重抽样（共12个练习题）……（227）
第13章　调查中的误差来源（共11个练习题）……（256）

第1章 引 言

1.1 假定你在用抽样的方法去估计一本有插图的书中的词的总数。

(a)总体的定义有什么问题吗?

(b)若(1)将一页做抽样单位;(2)将一行当作抽样单位。其赞成或反对的理由是什么?

【解答】 (a)在定义总体时,首先是插图是否算词,如果要算,要确定折算的关系,其次决定是否要把序言或索引中的词也计算在内,以及数学符号怎样当作"词"处理等。总体的范围应根据调查的目的来确定。

(b)(1)用页作为抽样单位,其优点是比较容易取得抽样框,其缺点是要数清每一个样本页上的词数,在有很多插图的情况下,由于有不印满的页数等原因,使每一页的词数可能出入很大,会使抽样误差增大,比较好的方法是将没有印满的页数单独列出来,分成二类,一类是印满的页,另一类是没有印满的页,再采用第5章所述的分层抽样方法分别抽样,可以克服这一缺点。(2)如果以行作为抽样单位,就必须将所有的行列出作为抽样框,以便直接抽行,编制这样的抽样框是很费时间的。还有每一段落最后一行的词数也可能不满,会出现差异,但总的说来每一行的词数还是比较稳定的。所以可以采用第11章所述的二级抽样方法来解决这个问题。即先抽一个由页组成的样本,然后数一下抽到页上的

行数，再在这些页上抽一个行的子样本即可。

1.2 从档案内登记在卡片上的人名（一个人名一张卡片）中抽一个样本，卡片是按连续的次序编号的。每个人名被抽入样本的概率相等，在下述经常会遇到的情况下，会出现什么问题？

(a)有些人名不属于目标总体，但在没有抽到人名之前是无法证实这一点的。

(b)有些人名在卡片上多次出现。同一人名的全部卡片具有连续的编号，因此在档案中一齐出现。

(c)有些人名在卡片上出现多次，但登录同一人名的卡片可能分散在档案的各处。

【解答】 假定对于这个问题我们以等概率抽了一个容量为 n 的样本。那么，

(a)在抽到的样本中，对于有些不属于目标总体的人名，其处理办法是，把不在目标总体内的样本姓名抛开不用。在这种情况下存在的问题是，从目标总体取得的样本姓名的含量一般少于原来规定的样本容量 n。有用卡片的数目是一个随机变量，它的值取决于抽选到的是什么样的卡片。

(b)有些卡片的姓名因重复会有较高的概率被抽中，处理的方法：一种是数一下人名重复的卡片数，采用第9A章中叙述的不等概率抽选方法；另一种是每个姓名都以等概率抽取，只有当这个姓名第一次出现时才留下，重复出现时都去掉。

(c)如果每个姓名出现的卡片数已有记录，则可以和(b)一样，采用不等概率抽样，如果没有记录则据目前所知尚无简易的方法使每个姓名有同等的机会被抽中。

1.3 找一个完整的抽样框通常是不容易的，在下述调

查中可以试用什么样的抽样框？

(a)调查一个大城市中卖皮箱的商店。

(b)调查失落在地铁或公共汽车上的东西的种类。

(c)调查去年被蛇咬过的人数。

(d)调查估计每周家庭成员用于看电视的总时数。

【解答】 (a)可以采用一本最近的百货商店和皮箱店的名册。

(b)地铁和公共汽车公司所设的失物招领处。

(c)发生过蛇咬的地区内的医院或私人医生，加上有义务向它报告蛇咬情况的公共卫生机构。所有这三种抽样框的缺点是很可能不完全。如果蛇咬事件很少发生，又没有集中报告制度，就很费钱。

(d)住户名册通常用来作为抽取家庭样本的抽样框。

1.4 一本四年前的城市居民地址录，按这条街列出地址和住在这个地址的人名。想在最近派人对本城的居民进行一次调查，这本居民录作为抽样框有什么缺点？能否由调查员在实际调查中补救？用这本居民录时，你想抽选一些地址（即居住地方）还是抽选一些人呢？

【解答】 由于有新的建筑，所以采用四年前的城市居民地址作为抽样框，其缺点是不完全。在一个由地址组成的样本中，调查人通常是可以处理新住户的。对任何一个地址样本，他可以检查在居民地址录中，这个地址和下一个地址之间是否有新的住户，若有，就把新住户列入样本。如果整个区域都是新建筑，则它们不会列在四年前的居民住址录中，因此需要另外一个抽样框。从地址录中抽样比从人名录中抽样好，因为地址更长久些。然而为了节约旅差费等原因，抽样单位可以是一个城市的街区，从街区中再抽选出住户的子

样本。（注：根据我国的情况——有户口制度，因而用户口册作抽样框是比较方便的。）

1.5 用抽样方法对一个大商行存货中的小商品的实际价值作出估计时，样本中每件商品的实际价值和帐面价值都进行了记录。对全部样本，实际价值与帐面价值之比是1.021，这个估计值近似正态分布，具有0.0082的标准误。若存货的帐面价值是8万美元，求实际价值的95%的置信限。

【解答】 设样本的实际价值为 Σy_i，帐面价值为 Σx_i，二者的比率为 R，已知

$$\hat{R} = \frac{\Sigma y_i}{\Sigma x_i} = 1.021$$

$$S(\hat{R}) = 0.0082$$

且 $\hat{R}$ 为近似正态分布，所以 R 的95%置信限为

$$\hat{R} \pm 1.96[S(\hat{R})] = 1.021 \pm 1.96(0.0082)$$

又 $\hat{Y} = \hat{R}X = 1.021(8) = 81\,630$（美元）

$$S(\hat{Y}) = XS(\hat{R}) = 656$$

所以，Y 的95%置信限为

$$\hat{Y} \pm 1.96S(\hat{Y}) = 81\,680 + 1.96(656)$$

即80 394美元～82 966美元。

1.6 有一些资料虽然初看似乎是全面调查资料，但却必须作为样本看待，这种情况是常有的。停车场的业主发现星期日上午生意是清淡的。在开业的26个星期日，平均每星期日上午收入恰为10美元。从逐周变化的资料中算得这个数字的标准误是1.2美元。每星期日服务人员的费用是7美元。如果业主期望每星期日上午将来的利润是5美元时，才愿意

在这个时候开业。请问(每个星期日上午)长期利润至少为5美元的置信概率有多大？要回答这一问题应作什么样的假定？

【解答】 要回答这一问题必须先假定将来的收入和这26笔收入的样本遵从相同的分布，并进一步假定这一分布是正态分布。

设：Y 表示收入，已知

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 10,\quad S(y_i) = 1.2$$

则欲求的概率为：

$$P(Y-7\geqslant 5) = P(Y\geqslant 12) = P\left(\frac{Y-\bar{Y}}{S(Y_i)} \geqslant \frac{12-10}{1.2}\right)$$

$$= P(t\geqslant 1.67) = 0.054$$

所以，置信概率大约是0.054。

1.7 在表1.2中，当 B/σ 趋于无穷大时，也即 MSE 全由偏差产生时，大于 $1\sqrt{MSE}$，$1.96\sqrt{MSE}$，$2.576\sqrt{MSE}$ 的概率会有什么变化？你计算的结果与表1.2中当 B/σ 由0向0.6变动时的变化方向是否一致？

【解答】 当 B/σ 趋于无穷大时，说明 $\sigma = 0$，其误差全部来自偏差。因而，误差 $= \sqrt{MSE}$ 的概率为1，这时其误差也不可能 $\geqslant 1.96\sqrt{MSE}$，更不可能 $\geqslant 2.567\sqrt{MSE}$，因此其概率为0。这与表1.2中当 B/σ 由0向0.6变动时的变化方向是一致的。

1.8 当有必要比较误差 $(\hat{\mu}-\mu)$ 的频率分布不同的两个估计量时，在一些特殊问题中，有时可能计算由于任何给定大小的误差 $(\hat{\mu}-\mu)$ 所产生的费用与损失。在其他方面都相同时，给出较小的期望损失的估计量是较好的，请证明

若损失是误差的二次函数 $\lambda(\hat{\mu}-\mu)^2$，我们应选用有较小均方误差的估计量。

【解答】 已知损失函数 $L=\lambda(\hat{\mu}-\mu)^2$，要使期望损失最小，即 $E(L)=E[\lambda(\hat{\mu}-\mu)^2]=\lambda E(\hat{\mu}-\mu)^2=\lambda MSE(\hat{\mu})$ 达到最小。因为 λ 是常数，所以也即 $MSE(\hat{\mu})$ 达到最小，故应选择有较小均方误差的估计量。

第 2 章　简单随机抽样

2.1　$N=6$ 的总体中，y_i的值是 8，3，1，11，4，7。请对所有可能的含量为 2 的样本，计算样本均值 $\bar{y}$。证明$\bar{y}$是$\bar{Y}$的一个无偏估计，它的方差是定理2.2给出的那个值。

【解答】　所有可能的含量为 2 的样本及均值如下：

样本	均值	样本	均值	样本	均值	样本	均值
8，3	5.5	8，1	4.5	8，11	9.5	8，4	6
8，7	7.5	3，1	2	3，11	7	3，4	3.5
3，7	5	11，4	7.5	11，7	9	1，11	6
1，4	2.5	1，7	4	4，7	5.5		

所有可能的含量为 2 的样本共有：$C_6^2=15$(个)

它们均值的均值 $=\Sigma\bar{y}/15\doteq 5.67$，而 $\bar{Y}=\Sigma y_i/N=\frac{34}{6}\doteq 5.67$。所以，$\bar{y}$ 是 $\bar{Y}$ 的一个无偏估计。由所有样本均值 $\bar{y}$ 计算的方差：$\sigma_{\bar{y}}^2=4.49$，由定理 2.2 给出的是：$V(\bar{y})=\frac{1-f}{n}S^2$，$n=2$，$f=\frac{2}{6}$，$S^2=13.47$，故 $V(\bar{y})=\frac{1}{2}\times\frac{2}{3}\times 13.47=4.49$，所以 $\sigma^2(\bar{y})=V(\bar{y})$。

2.2　对同一个总体，请计算含量为 3 的所有简单随机样本的 s^2，并证明：$E(s^2)=S^2$。

【解答】　所有可能的含量为 3 的样本及方差如下

样本	方差	样本	方差	样本	方差
8，3，1	13	8，3，11	16.33	8，3，4	7
8，3，7	7	8，1，11	26.33	8，1，4	12.33
8，1，7	14.33	8，11，4	12.33	8，11，7	4.33
8，4，7	4.33	3，1，11	28	3，1，4	2.33
3，1，7	9.33	3，11，4	19	3，11，7	16
3，4，7	4.33	1，11，4	26.33	1，11，7	25.33
1，4，7	9	11，4，7	12.33		

所有 s^2 的均值 $E(s^2)=13.47$，故可证明 $E(s^2)=S^2$。

2.3 设在此总体中有放回地抽取含量为 2 的样本，请用找出所有可能的样本的方法来证明 $V(\bar{y})$ 满足等式

$$V(\bar{y})=\frac{\sigma^2}{n}=\frac{S^2}{n}\cdot\frac{N-1}{N}$$

【解答】 含量为 2 的样本（有放回）及均值如下：

样本	均值	样本	均值	样本	均值	样本	均值
8，8	8	8，3	5.5	8，1	4.5	8，11	9.5
8，4	6	8，7	7.5	3，8	5.5	3，3	3
3，1	2	3，11	7	3，4	3.5	3，7	5
1，8	4.5	1，3	2	1，1	1	1，11	6
1，4	2.5	1，7	4	11，8	9.5	11，3	7
11，1	6	11，11	11	11，4	7.5	11，7	9
4，8	6	4，3	3.5	4，1	2.5	4，11	7.5
4，4	4	4，7	5.5	7，8	7.5	7，3	5
7，1	4	7，11	9	7，4	5.5	7，7	7

由所有可能的样本的均值 $\bar{y}$ 计算的方差 $V(\bar{y})=5.61$ 而

$$V(\bar{y}) = \frac{S^2}{n} \cdot \frac{N-1}{N} = \frac{13.47}{2} \times \frac{5}{6} = 5.61$$

即 $V(\bar{y})$满足 $V(\bar{y}) = \frac{S^2}{n} \cdot \frac{N-1}{N}$。

2.4 从一个有 14 848 个住户的市区中抽取一个30个 住户的简单随机样本。样本中每一住户的人数是 5，6，3，3，2，3，3，3，4，4，3，2，7，4，3，5，4，4，3，3，4，3，3，1，2，4，3，4，2，4。估计这一市区居民的总数，并算出这个估计算在真值的±10%范围内的概率。

【解答】 已知 $n = 30$，$N = 14\,848$，$\bar{y} = 3.46667$

$$s^2 = 1.499$$

$$\hat{Y} = N\bar{y} = 14\,848 \times \frac{104}{30} = 51473(\text{人})$$

$$v(\bar{y}) = \frac{1-f}{n} s^2 = 0.0499$$

$$P\left\{\frac{|N\bar{y} - N\bar{Y}|}{N\bar{Y}} < 10\%\right\} = P\left\{\frac{|\bar{y} - \bar{Y}|}{\bar{Y}} < 10\%\right\}$$

$$= P\left\{\frac{|\bar{y} - \bar{Y}|}{\sqrt{V(\bar{y})}} < \frac{10\%\bar{Y}}{\sqrt{V(\bar{y})}}\right\}$$

用 $\bar{y}$ 代替 $\bar{Y}$，用 $v(\bar{y})$代替 $V(\bar{y})$，

$$t = \frac{10\%\hat{y}}{\sqrt{v(\bar{y})}} = 1.553$$

查表得 $P = 0.88$。

2.5 探索用抽样的方法来节省盘查库存的工作 量时，将库内36个货架上的货物的价值列一清单。将价值都折算成美元（用四舍五入的方法）后如下：29，28，42，44，45，

47, 51, 53, 53, 54, 56, 56, 56, 58, 58, 59, 60, 60, 60, 60, 61, 61, 61, 62, 64, 65, 65, 67, 67, 68, 69, 71, 74, 77, 82,85，从一个样本所得的总的价值的估计值的误差不超过200美元，想控制超过的机会小于1/20，有人建议：用一个含量为12个货架的简单随机样本就可达到要求。你是否同意呢？ $\Sigma y = 2.138$，$\Sigma y^2 = 131\,682$

【解答】 已知 $S^2 = \dfrac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1} = \dfrac{\sum_{i=1}^{N} y_i^2 - N\bar{Y}^2}{N-1} = 134.53$

$n = 12$，$N = 36$

$$\sigma_{\hat{Y}}^{\wedge} = \sqrt{Var(N\bar{y})} = \sqrt{N^2 \frac{1-f}{n} S^2} = 98.42$$

所以 $P(|\hat{Y} - Y| > 200) = P\left(\dfrac{|\hat{Y} - Y|}{\sigma_{\hat{Y}}^{\wedge}} > \dfrac{200}{98.42}\right)$

$$\doteq 0.04 < \frac{1}{20}$$

同意用这种方法进行盘查。

2.6 从表2.2（第41页）的样本取得以后，数得签满了名的张数（每张有42个签名）为326，利用这些信息作一个对总签名数的改进的估计，并求出你这个估计值的标准误。

【解答】 已知 $N = 676$，$N_j = N - 326 = 350$

$n = 50$，$n_j = 27$

除去了签满名的单位后，$y = \Sigma f_i y_i = 505$，$\Sigma f_i y_i^2 = 13\,925$

$$\hat{Y} = \hat{Y}_{\text{签满名}} + \hat{Y}_{\text{未签满名}} = 42 \times 326 + \frac{350}{27} \times 505 = 20\,238$$

$$S^2 = \frac{1}{n_j - 1}\left(\Sigma f_i y_i^2 - \frac{(\Sigma f_i y_i)^2}{\Sigma f_i}\right) = 172.293$$

所以 $V(\hat{Y}_j)=\frac{1-f_j}{n_j}\cdot N_j^2\cdot S^2=849^2$

（$\hat{Y}_j$：未签满名的总签名数的估计。）

$\hat{Y}$ 的标准误为 $\sigma(\hat{Y})=\sqrt{V(\hat{Y})}=\sqrt{V(\hat{Y}_j)}=849$。

2.7 从468个二年制的学院中抽取一个100个学院的简单随机样本。样本中有54所公立学院，46所私立学院。学生数(y)和教师数(x)的数据如下：

	n	$\Sigma(y)$	$\Sigma(x)$
公立学院	54	31 281	2 024
私立学院	46	13 707	1 075

	$\Sigma(y^2)$	$\Sigma(xy)$	$\Sigma(x^2)$
公立学院	29 881 229	1 729 349	111 090
私立学院	6 366 785	431 041	33 119

(a)对这个总体中的每一类学院，估计比率（学生数/教师数）。

(b)算出你的估计值的标准误。

(c)对公立学院，求整个总体的学生数/教师数的90%的置信限。

【解答】 设对应于公立学院的变量下标为“1”，私立学院对应的为“2”，则

(a) $\hat{R}_1=\frac{\bar{y}_1}{\bar{x}_1}=\frac{\Sigma(y_1)}{\Sigma(x_1)}=\frac{31\,281}{2\,024}=15.46$

$\hat{R}_2=\frac{\bar{y}_2}{\bar{x}_2}=\frac{\Sigma(y_2)}{\Sigma(x_2)}=\frac{13\,705}{1\,075}=12.75$

(b) $S(\hat{R})=\frac{\sqrt{1-f}}{\sqrt{n}\,\bar{X}}$