系列讲座

流行病学基本方法

I. 随机化方法的流行病学应用

华北煤炭医学院 吕宝成

在流行病学观察和实验方法中，随机化的应用是很重要的。实际上，这一方法在流行病学工作中往往被忽视，因此，一些研究报道即便是材料殷实，结果却出现偏差或令人置疑，前功尽弃。如一份报告指出：在中华流行病学杂志等六种刊物已发表的论文中，分析到随机化缺失率为73.1%（张揆一. 健康报1988年第2784期）。1988年1月，在泰国Pattaya召开的国际流行病学地区性科学会议上，世界卫生组织顾问B.G.Weniger指出，随机化是现代流行病学应该共同关心的方法。

随机化（randomization）是以概率统计为基础的。在观察流行病学中主要应用于随机抽样调查，在实验流行病学中主要应用于实验组和对照组随机化，要求处理组和对照组除处理条件外，其他均基本一致。随机抽样也称概率抽样，它应具备两个原则：①总体中被抽取的样本是随机的，而非随意的，抽取机会均等，分布均匀，排除主观意识抽取样本；②能测量和控制抽样误差，估计样本量，以提高调查结果的精密度。在流行病学研究中未遵循随机化原则，导致研究失败的例子不胜枚举。如英国曾有一个例子，给小学生做牛乳加餐试验，实验组和对照组均为一万人，研究结果是对照组身长体重反比试验组高。研究失败原因是样本选择未按随机化，而是根据老师的主观意识，他们很关心的把发育差的学生分配到牛乳加餐实验组。

本文主要讨论随机化方法和随机抽样方法。样本量估计方法和无应答等在以后讲座讨论。

一、随机化方法：随机化方法主要用随机数字表（按数字分组）。此外，也可用抽签（按号码分组），摸球（按颜色分组）和抓阄（按符号分组）等方法。随机数字表用于单纯随机抽样比较简便实用，应该熟练掌握。在使用随机表前，应先将总体样本按顺序进行编号。

1.随机表方法：使用随机表（表1）时，可从横、纵行或斜向顺序抽取样本，并可从任何一处开始。假如有一个总体有8059个个体，拟从中抽取300个做为样本。可从表1的横排或纵列采取，如以左上角横排开

表1 随机抽样数字表（本文有删节）

03 47 43 73 86	36 96 47 36 61	46 98 63 71 62
97 74 24 67 62	42 81 17 57 20	42 53 32 37 32
16 76 62 27 66	56 50 26 71 07	32 90 79 78 53
12 56 85 99 26	96 96 68 27 31	05 03 72 93 15
55 59 56 35 64	38 54 82 46 22	31 62 43 09 90
16 22 77 94 39	49 54 43 54 82	17 37 93 23 78
84 42 17 53 31	57 24 55 06 88	77 04 74 47 67
63 01 63 78 59	16 95 55 67 19	98 10 50 71 75
33 21 13 34 29	78 64 56 07 82	52 42 07 44 38
57 60 86 32 44	09 47 27 96 54	49 17 46 09 62
18 18 07 92 46	44 17 16 58 09	79 83 86 19 62
26 62 38 97 75	84 16 07 44 99	83 11 46 32 24
23 42 40 64 74	82 97 77 77 81	07 45 32 14 08
52 36 28 19 95	50 92 26 11 97	00 56 76 31 38
37 85 94 35 12	83 39 50 08 30	42 34 07 96 88
70 29 17 12 13	40 33 20 38 26	13 89 51 03 74
56 62 18 37 35	96 83 50 87 75	97 12 25 93 47
99 49 57 22 77	88 42 95 45 72	16 64 36 16 00
16 08 15 04 72	33 27 14 34 09	45 59 34 68 49
31 16 93 32 43	50 27 89 87 19	20 15 37 00 49
68 34 30 13 70	55 74 30 77 40	44 22 78 84 26
74 57 25 65 76	59 29 97 68 60	77 91 38 67 54
27 42 37 86 53	48 55 90 65 72	96 57 69 36 10
00 39 68 29 61	66 37 32 20 30	77 84 57 03 29
29 94 98 94 24	68 49 69 10 82	53 75 91 93 30

始，取四位数依次为0347、9774和1676等，到底后重新一行为4373、2464等；如用纵列四位数，则为0911、5186和3512等；采用斜向则为0374、6299…。这些方法还可以交替使用。为了更好地抽取样本，每次选取时可任意取一张钞票，以其末两位数做为起点，如钞票末两位是27，则这次样本可取第27个四位数为起始，按表1查出横排四位数为2467，依次为6227、8599等。凡超出总体数（本例为8599）者均不列为样本，以此类推，直至选出300个样本为止。如总体数为2000，从表1纵列顺序选出0347、9774、1676、1256、5559、1622和8442等，对首位数修正≥8减8，≥6减6，≥4减4，≥2减2，表中修正数字则依次为0347、1774、1676、1256、1559、1622和0442，使修正后要求的样本数均在2000以内。如总体数改为三位数（如200），可取表中三位数按上述的减首位法修正。如总体数为奇数（如3000）时，首位修正时应≥7减7，≥5减5，≥3减3，表中纵列数0347、9774、1676、1256、5559、1622、8442等修正为0347、2774、1676、1256、0559、1662和0442等，修正后要求样本数均在3000以内。

在实验研究中，对实验组和对照组的分组可采取随机编组方法，将所有对象顺序排列编号后，第一个取值要从任一随机数字开始依次排出。①要求实验组和对照组人数各半（$\frac{1}{2}$）时，先用随机数编排的奇偶数分为试验组和对照组，再调整人数，对较多的一组按随机化方法移入较少的组；②若实验组和对照组分为三组时，如受试对象为12个，可按随机数字表出现（如横排）1～12数值顺序编入已排好次序的12个对象中。依次列出随机数字表中0～12的顺序为03、07、12、05、09、06、04、01、10、11、08和02。将01～04、05～08和09～12分别列入1、2和3组。上列实验和对照三组的12个对象的编组随机顺序为1、2、3、2、3、2、1、1、3、3、2和1，每组为4个。

用随机表抽取样本或分组完全由机遇而定，排除了研究人员的主观决定。所以，随机化是保证研究结果确切的重要手段。

2.实验研究的随机化：为了保证在实验研究中的齐同性和分组均衡，还可以采取实验研究的随机方法。①配对随机：即试验组和对照组的人员按性别、年龄和其他影响条件配成对子，随机地分入试验组和对照组以保证均衡；②分层随机：在条件较多时，试验组和对照组可以采用分层随机方法。

例如研究风疹疫苗，先分为四个群组：

<1岁，抗体<1：10；

1岁，抗体<1：10；

<1岁，抗体≥1：10；

1岁，抗体≥1：10。

然后用两种（Ⅰ、Ⅱ型）接种风疹疫苗观察效果，列入Ⅰ、Ⅱ型疫苗接种者也要随机分组，以使潜在的因素以同样机率分配到两组中。

二、随机抽样：在流行病学现场调查中，限于人力物力等条件，不能对总体进行调查，一般采用随机抽样方法。常用的随机抽样方法是：单纯随机抽样和系统抽样、分层随机抽样、整群随机抽样以及两阶段抽样。这些抽样方法概念不同，抽样误差计算方法各异，现举例介绍如下。

1.单纯随机抽样和系统抽样：单纯随机抽样是将总体全部顺序编号，用随机方法抽取需要数量的样本。系统抽样也称等间隔抽样或机械抽样，它同样是以总体顺序排列为基础，在抽样开始时，先随机取一个观察单位（数字）作为抽样起点，以后按一定间隔数依次抽取各观察单位组成样本。这两种随机抽样用表2中公式计算抽样误差（标准误）。

表2　单纯随机抽样和系统抽样的标准误和可信区间

均数的标准误	$S_{\bar{x}}=\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}$	(2-1)
均数95%可信区间	$\bar{x}\pm t_{a,v}\cdot\frac{S}{\sqrt{n}}$	(2-2)
率的标准误	$S_p=\sqrt{\left(1-\frac{n}{N}\right)\left[\frac{P-(1-P)}{n-1}\right]}$	(2-3)
率的95%可信区间	$P\pm 1.96\sqrt{\frac{P(1-P)}{n}}$	(2-4)

S 样本标准差　P 样本率　N 总体数　n 样本数　$t_{a,v}$ 自由度n－1的t值

〔例1〕调查某校小学生2 000人的近视率，随机抽取100人，近视率10%，求抽样误差和可信区间。

代入（公式2-3），

$$Sp=\sqrt{\left(1-\frac{100}{2000}\right)\left[\frac{0.1(1-0.1)}{100-1}\right]}=0.02938$$

代入（公式2-4），95%可信区间

$$=0.1\pm 1.96\sqrt{\frac{0.1(1-0.1)}{100}}=0.1\pm 0.0588$$

〔例2〕某厂成年男子2 000人，随机抽取144人做红细胞计数，均值为537.8万/mm³，标准差43.9万/mm³。

求抽样误差和可信区间〔查表t0.05，100（近似值）=1.984〕。

代入（公式2-1），

$$S\bar{x}=\sqrt{\left(1-\frac{144}{2000}\right)\frac{43.9^2}{144}}=3.52\text{万/mm}^3$$

代入（公式2-2），95%可信区间

$$=537.8\pm 1.984\times\frac{43.9}{\sqrt{144}}=545.0\sim 530.6\text{万/mm}^3$$

〔例3〕某医院从10 000份病历中以系统抽样法随机抽取1 000份，医院内感染率8%，求抽样误差和可信区间。

$$\text{抽样间隔}=\frac{N\text{（总体量）}}{n\text{（样本量）}}=\frac{10000}{1000}=10$$，先取一个随机数（如6）确定第一个抽取的样本，以10的间隔抽取尾数为6、16、26、36…至9996共1 000个样本。

代入（公式2-3），

$$Sp=\sqrt{\left(1-\frac{1000}{10000}\right)\left[\frac{0.08(1-0.08)}{1000-1}\right]}=0.00814$$

代入（公式2-4），95%可信区间

$$=0.08\pm 1.96\sqrt{\frac{0.08(1-0.08)}{1000}}=0.08\pm 0.0168$$

2.分层抽样：根据研究目的，先按总体不同特征（如不同地区、年龄和职业等），成为若干部分(层)，然后再从各层内单纯随机抽样或系统抽样。这种抽样要求同一层次的变差小，而不同层次的变差大。分层抽样的精密度（Precision）要比同等数量单纯随机抽样为高，因此，所需样本量较小。分层抽样又分为最佳分配和按比例分配抽样，两者计算抽样误差方法不同（表3）。可信区间的计算同单纯随机抽样。

表3　　分层抽样法和抽样误差公式

均数抽样	$ni=n\frac{Ni\,oi}{\Sigma Ni\,oi}$	（3-1）
均数的标准误	$S\bar{x}=\sqrt{\left(1-\frac{n}{N}\right)\left(\frac{1}{n^2}\right)\Sigma ni\,Si^2}$	（3-2）
率按比例抽样	$ni=Ni(n/N)$	（3-3）
率按最佳分配抽样	$ni=n\frac{Ni\sqrt{\pi i(1-\pi i)}}{\Sigma Ni\sqrt{\pi i(1-\pi i)}}$	（3-4）
率按比例抽样标准误	$Sp=\sqrt{\left(1-\frac{n}{N}\right)\left(\frac{1}{n^2}\right)\Sigma\left(\frac{ni^2}{ni-1}\right)Pi(1-Pi)}$	（3-5）
率按最佳分配标准误	$Sp=\frac{1}{N}\sqrt{\Sigma Ni^2\left(1-\frac{ni}{Ni}\right)\left[\frac{Pi(1-Pi)}{ni-1}\right]}$	（3-6）

注：N 总体数 Ni 总体第i层数 n 样本数 ni 样本第i层数 o 总体第i层标准差 S 样本标准差 Si 第i层样本标准差 πi 总体第i层的率 P 样本率 Pi 第i层的样本率

式中o或πi可据文献或小范围调查估计，按比例分配需知总体各层人数，最佳分配需知oi或πi

〔例4〕某地100 000人，城、郊、乡各为60 000、10 000和30 000人，既往已知某病阳性率分别为40%、20%和60%，此次拟分层抽查2 000人，求最佳分配和按比例分配。见表4计算。

表4 最佳分配和按比例分配及抽样误差计算

层	Ni	πi	$\sqrt{\pi i(1-\pi i)}$	$Ni\sqrt{\pi i(1-\pi i)}$	$n\cdot Ni\sqrt{\pi i(1-\pi i)}$	$\frac{Ni\sqrt{\pi i(1-\pi i)}}{\Sigma Ni\sqrt{\pi i(1-\pi i)}}$	$ni=\frac{n\cdot Ni\sqrt{\pi i(1-\pi i)}}{\Sigma Ni\sqrt{\pi i(1-\pi i)}}$ *	Ni(n/N)**
1	60 000	0.4	0.4899	29 393.9	58 787 753	0.611	1 222	1 200
2	10 000	0.2	0.4000	4 000.0	8 000 000	0.083	166	200
3	30 000	0.6	0.4899	14 696.9	29 393 876	0.306	612	600
合计	100 000			48 091	168 181 629	1.000	2 000	2 000

*为最佳分配；**按比例分配（n／N）=0.02

表4计算完成最佳分配和按比例分配的各层样本数后，可以单纯随机抽样或等间隔抽样抽取样本。

3.整群抽样：将总体分为若干群组，抽取群组中的一部分整体做为样本，各群组内的样本数可相同或不同。整群抽样法比较简便经济，但随机性小，抽样误差大，故抽取样本量要比其他随机化抽样方法增加1/2，其可信区间计算同单纯随机抽样，抽样误差公式见表5。

〔例5〕某校100个班，4 200名学生，分11班（人数不等），用随机抽样法调查蛔虫感染率，求整群抽样感染率及抽样误差。演算见表6。

〔例6〕某校80个班，各班学生均为50人，随机整群抽取8个班进行锡克氏试验，求样本阳性率及抽样误差。计算见表7。

4.两阶段抽样：调查单位由初级单位组成，再由初级单位抽取样本。如疫苗接种率调查，先从初级单位（居委会、村等）提供30个单位，再从这些单位各抽取7人，共210人作为样本，这种方法称为按初级单位容量比例概率抽样调查法（probability proportional to size，简称PPS）。PPS法抽样误差为±10%，如接种率为70%，则总体接种率在60～80%之间，这样的调查结果，调查20次有19次在此范围内，即95%可信区间±10%。这种调查方法同样可应用于其他调查（如血清流行病学调查等），较节省样本（表8）。

〔例7〕某市有66个居委会，共33万人，如按PPS法抽样，如何计算？

计算程序见表8，计算步骤如下：

①抽样间隔（组距）＝累计人口数/抽样单位数＝330 000/30＝11 000

②确定第一个随机数，它应等于或小于组距，如以一张钞票确定后四位数为5306，则此数在2号居委会的秩次范围内，此次为第一个初级抽样单位。

③确定第2个以后的抽样单位：第2个抽样单位为组距＋随机数＝11000＋5306＝16306，第3个抽样单位为2×组距＋随机数＝2×11000＋5306＝27306，连续计算直至29×组距＋随机数＝29×11000＋5306＝324306，此值在最后一个第66个居委会内。

在选定的30个初级单位中，再随机抽取7个样本。如已抽取的各居委会是各个街道，则先随机抽取一条街，再从一条街内各户口随机抽取一个门牌起始，依次查找登记所需要的对象。调查完第一门后，以面朝门方向退出，不转身的右手方向进入第二门调查，

表5　　整群抽样样本率和抽样误差公式

群内观察数不等

样本均数　$x=\frac{K}{Nk}\Sigma mi\bar{x}i$　　(5-1)

均数标准误　$S^{-}=\frac{K}{N}\cdot\sqrt{\left(1-\frac{k}{K}\right)\left[\frac{1}{k(k-1)}\right]\sum_{i=1}^{k}(Ti-\bar{T})^2}$　　(5-2)

样本率　$P=\frac{K}{Nk}\Sigma ai$　　(5-3)

率的标准误　$Sp=\frac{K}{N}\cdot\sqrt{\left(1-\frac{k}{K}\right)\left[\frac{1}{k(k-1)}\right]\sum_{i=1}^{k}(ai-\bar{a})^2}$　　(5-4)

群内观察数相等

样本均数　$\bar{x}=\frac{\Sigma xi}{k}$　　(5-5)

均数标准误　$Sx=\sqrt{\left(1-\frac{k}{K}\right)\cdot\frac{\Sigma(\bar{x}i-\bar{x})^2}{k(k-1)}}$　　(5-6)

样本率　$P=\frac{1}{k}\Sigma pi$　　(5-7)

率的标准误　$Sp=\sqrt{\left(1-\frac{k}{K}\right)\cdot\frac{\Sigma(pi-p)^2}{k(k-1)}}$　　(5-8)

K 总观察群数，k 抽样群数，N 总人数，mi 群内观察数不等的第i群人数，xi 样本第i群的均数，Ti 样本第i群内观察值之和，$\bar{T}$ 各Ti的均数（$\Sigma Ti/x$），ai 样本各群阳性数，$\bar{a}$ 样本各群的平均阳性数，Pi 样本中第i群的率，P样本平均阳性率，$\bar{x}i$、$\bar{x}$与Ti、$\bar{T}$相同

表6　整群抽样（各群人数不等）的样本率和标准误计算（$\bar{a}=21.6$）

班号	受检人数mi	阳性数ai	$(ai-\bar{a})^2$	样本率 P*	标准误 Sp**
1	42	18	13.25		
2	40	20	2.69		
3	45	25	11.29		
4	44	22	0.13		
5	40	28	40.45		
6	44	24	5.57		
7	42	27	28.73		
8	46	15	44.09		
9	42	21	0.41		
10	45	20	2.69		
11	44	18	13.25		
合计	474	238	162.55	0.515	0.0273

* $P=\frac{K}{Nk}\Sigma ai=\frac{100}{4200\times11}\times238=0.515$，

即蛔虫感染率为51.5%；

** $Sp=\frac{K}{N}\sqrt{1-\frac{k}{K}\left(\frac{1}{k(k-1)}\right)\Sigma(ai-\bar{a})^2}$

$=\frac{100}{4200}\sqrt{1-\frac{11}{100}\left(\frac{1}{11(11-1)}\right)162.55}$

$=0.0273$

依次查完7例，30条街，共210人，计算率。如初级单位为农村，可选一个中心位置（如市场等），按随机方法确定东、西、南、北方向，以随机选择的这一方向划成一条直线，调查分布在此直线上的各户。或者将村户编号，随机确定第1户参照城市住户进行。

在随机抽样调查中，可以将系统、分层和整群等方法结合起来，这样往往可以收到节省人力物力，抽样误差小和精密度高的效果。

表7　整群抽样（各群人数相等）的样本率和标准误计算（P=0.34）

班号	受检人数(m)	阳性数(ai)	阳性率(pi)	$(pi-p)^2$	样本率P*	标准误Sp**
1	50	12	0.24	0.01		
2	50	17	0.34	0.00		
3	50	12	0.24	0.01		
4	50	15	0.30	0.0016		
5	50	21	0.42	0.0064		
6	50	20	0.40	0.0036		
7	50	21	0.42	0.0064		
8	50	18	0.36	0.0004		
合计	400	136		0.0384	0.34	0.0248

* $P=\Sigma ai/km=\frac{136}{8\times 50}=0.34$，

即锡克氏阳性率为34%；

** $Sp=\sqrt{\left(1-\frac{k}{K}\right)\frac{\Sigma(pi-p)^2}{k(k-1)}}$

$=\sqrt{\left(1-\frac{8}{80}\right)\frac{0.0384}{8(8-1)}}=0.0248$

习　题　和　提　示

1.随机化的主要方法。（参见本文前言第2段复习）

表8　　PPS程序

居委会编号	人口数	累计秩次范围	随机数字
1	4000	1～4000	
2	5000	4001～9000	①5306
3	6000	9001～15000	
4	4000	15001～19000	②16306
5	5000	19001～24000	
6	6000	24001～30000	③27306
7	4000	30001～34000	
⋮	⋮	⋮	⋮
64	4000	315001～319000	
65	5000	319001～324000	㉙313306
66	6000	324001～330000	㉚324306

2.预试验20人，分为试验和对照组，如何随机化？（参见本文“随机表方法”第2段①内容的方法）

3.总体数为4000，用随机数字表如何抽取？（参见本文“随机表方法”第1段总体数2000例的方法）

4.用计算器练习演算例1～7.

5.将表8中的PPS程序，完成④～㉘随机数.（参见本文表8及计算步骤）

6.计算例4最佳分配和按比例分配的分层抽样误差（参见本文公式3-4和3-5；参考答案：按比例Sp=0.0107006，最佳分配Sp=0.0106530）

韶关市幼龄儿童HBV感染的血清流行病学调查

广东韶关市卫生防疫站　李　杰　王石华　邓俊兴
广东省卫生防疫站　沙庆洪　刘家壁　刘玉霞

在韶关市开展广泛接种乙肝疫苗前，于1987年7～9月间随机抽查市区712名0～6岁未接种乙肝疫苗的健康儿童血样，用SPRIA法（RIA试剂盒由北京福瑞诊断用品联营公司提供）同时检测HBsAg、抗-HBs和抗-HBc，以全面反映该市儿童HBV感染状况。

结果表明，在这组幼儿中HBsAg、抗-HBs和抗-HBc的阳性率分别为13.06%、17.28%和30.62%。HBV总感染率为35.39%。男女性感染率分别为38.30%和34.04%，两者无显著差别（P>0.05）。而各种血清学标记的检出阳性率也不存在性别间差别。但年龄别存在显著差异。三种标记检出率是从2岁开始明显上升的，到6岁时总感染率达59.29%。而1岁以内婴幼儿的感染率在14%以下。由此启示我们要抓紧婴幼儿的疫苗接种，会获得非常明显的流行病学效果。