

系列讲座

现代实用流行病学方法

第一讲 探索性研究方法(续)

章扬熙

三、研究的方法：通过探索性研究要了解疾病发生的频率和空间分布的特征及有关因素；所以，必须有明确统一公认的诊断标准、疾病的准确数字和准确的人口数。常用的研究方法有：①现况研究及筛检；②历史法；③疾病监测。

1. 现况研究 (prevalence study)：可分为普查和抽样调查两类。为了了解某病的患病率或健康状况，在一定的时间内对一定范围的人群中每一成员进行调查，称为普查，普查的目的主要是为了及早发现病人并给予及时的治疗。当某病患病率较高、有简易的灵敏度和特异度均高的检查或检测的手段发现病人、有有效的疗法时，可进行普查。我国曾对血吸虫病等多种寄生虫病和甲状腺肿等地方病进行过大规模普查，成绩很大。本节重点讨论抽样调查。如果不是为了及早发现和及时治疗，只是为了了解分布特征及有关因素，研究防制对策，通常可采用抽样调查，既节省人力、物力和时间，也易保证调查质量。

(1) 抽样方法和样本含量：

①单纯随机抽样 (simple random sampling)：设总体中有N个个体，使总体中的每一个体都有同等的机会被选入样本的抽样，叫做随机抽样。如果被抽的个体，在重复抽样之前，不再送回总体则叫做单纯随机抽样。通常采用抽签法或随机数目表法进行抽取样本。随机数目表法就是给每一个体一个随机数目(其位数要大于或等于K的位数)，若抽取 $\frac{1}{K}$ 个个体，首先在1~K中随机取一值为L，每个随机数目除以K，把除后的余数为L的个体选入样本之中。这种抽样有产生极端样本的机会，所以它适用于个体特征没有很大参差的场合，且只有当样本大到一定程度，才能保证所抽出的样本具有总体的代表性。

样本含量的计算公式见表4。

例3 对某地高血压患病率进行调查，希望所得的样本率与总体率的差 δ 不超过1%，基于小规模预调得

表4 单纯随机抽样样本含量的计算公式

适用条件	计算公式
对总体率作估计	$n = \left(\frac{u_{\alpha}}{\delta}\right)^2 P(1-P)$ (1)
样本率与总体率的比较(单侧)	$n = \left(\frac{u_{2\alpha} + u_{2\beta}}{\delta}\right)^2 P(1-P)$ (2)
样本率与总体率的比较(双侧)	$n = \left(\frac{u_{\alpha} + u_{2\beta}}{\delta}\right)^2 P(1-P)$ (3)
对总体均数作估计	$n = \left(\frac{u_{\alpha} S}{\delta}\right)^2$ (4)
样本均数与总体均数的比较(单侧)	$n = \left(\frac{(t_{2\alpha} + t_{2\beta}) S}{\delta}\right)^2$ (5)
样本均数与总体均数的比较(双侧)	$n = \left(\frac{(t_{\alpha} + t_{2\beta}) S}{\delta}\right)^2$ (6)

注：式中 δ 为允许误差，一般取总体率(或均值)($1-\alpha$)可信限的一半，P为样本率，S为标准差， α 及 β 分别为第一类错误及第二类错误的概率， t_{α} ， $t_{2\alpha}$ ， $t_{2\beta}$ 分别为 α ， 2α ， 2β 检验水准的t值， u_{α} ， $u_{2\alpha}$ ， $u_{2\beta}$ 分别为 α ， 2α ， 2β 检验水准的u值。

样本率P=7%，问需查多少人？(规定 $\alpha=0.05$)。

按题意， $\delta=0.01$ ， $P=0.07$ ，查u值表， $u_{0.05}=1.96$ ，代入公式(1)得：

$$n = \left(\frac{1.96}{0.01}\right)^2 \times 0.07 \times (1-0.07) = 2500$$

故需查约2500人。

例4 某地当地居民高血压的患病率为7%，在移民中进行抽样考察其患病率是否与当地居民不同，希望当移民的高血压患病率与当地居民差3%，即可查出有显著意义的差别，规定 $\alpha=0.05$ ， $\beta=0.10$ ，问需查多少人？

按题意为双侧检验， $\alpha=0.05$ ， $u_{0.05}=1.96$ ， $\beta=0.10$ ， $2\beta=0.20$ ， $u_{0.20}=1.282$ ， $\delta=0.03$ ， $P=0.07$ ，代入公式(2)得：

$$n = \left(\frac{1.96 + 1.282}{0.03} \right)^2 \times 0.07 \times (1 - 0.07) = 760$$

故需查760人。

②系统抽样 (systematic sampling): 系统抽样是按照系统顺序等距抽样的方法, 它比单纯随机抽样简便。比如, 从100个村中抽取10个村。系统抽样的方法是将村依次编为1~100号, 再从随机数目表上任取一个一位数, 假设此数为4, 则4, 14, …… , 94, 被选入样本。样本在总体中的分布比较均匀, 代表性好, 误差也较小。系统抽样是有周期性的, 如果总体的特征也存在着同步的周期性, 则抽样结果必有偏性, 所以遇到这种情况应采用几个开头数或频繁地更换开头数, 其样本含量与单纯随机抽样相同。

③分层抽样 (stratified sampling): 分层抽样适用于总体中个体的特征有较大参差的场合。分层的原则是使参差尽量表现于各层之间而不表现于各层

之内。分层抽样能提高效率, 抽样误差较小。在进行分层抽样时, 必须每个层都抽到。分层抽样的样本含量可参照单纯随机抽样计算, 至于每层的样本含量可采用最优分配分层随机抽样或采用按比例抽样。最优分配分层抽样应用下式求每层内抽查的含量。

$$\text{作为计量资料时, 用 } n_i = n \cdot \frac{N_i S_i}{\sum N_i S_i} \quad (7)$$

$$\text{作为计数资料时, 用 } n_i = n \cdot \frac{N_i \sqrt{P_i(1-P_i)}}{\sum N_i \sqrt{P_i(1-P_i)}} \quad (8)$$

式中 N_i 为第 i 层的人口数, P_i 为第 i 层的样本率, S_i 为第 i 层的标准差, n 为计划抽取的样本总含量。

例5 在14万人口的地区中调查某病患病率, 居民按职业分成甲、乙、丙三个层次, 现要抽取一个1400人的样本, 若以五年前各层患病率做参考, 按最优分配随机抽样, 问各层需查多少人?

计算应用公式(8), 具体步骤如下表。

表5 最优分配分层随机抽样各层应查人数的计算过程

职业	人口数 N_i	五年前 患病率 P_i	$\sqrt{P_i(1-P_i)}$	$N_i \sqrt{P_i(1-P_i)}$	$\frac{N_i \sqrt{P_i(1-P_i)}}{\sum N_i \sqrt{P_i(1-P_i)}}$	$n_i = n \cdot \frac{N_i \sqrt{P_i(1-P_i)}}{\sum N_i \sqrt{P_i(1-P_i)}}$
甲	60 000	0.05	0.218	13 080	0.306	428
乙	50 000	0.10	0.300	15 000	0.351	491
丙	30 000	0.40	0.490	14 700	0.344	481
合计	140 000	0.143	—	42 780	1.000	1 400

结果职业甲抽查428人, 职业乙抽查491人, 职业丙抽查481人。

按比例分层抽样的各层含量计算比较简单, 比如例5资料若用此法则为职业甲600人, 职业乙500人, 职业丙300人。分层抽样的均数与率分别用下两式计算。

$$\bar{X} = \frac{\sum N_i \bar{X}_i}{\sum N_i} \quad (9)$$

$$P = \frac{\sum N_i P_i}{\sum N_i} \quad (10)$$

式中 \bar{X}_i 为第 i 层的均数, 其余符号同前。

④整群抽样 (clustering sampling): 这是以整群或整组为单位进行随机抽样的方法。群或组可以大小相等, 也可以不等。比如, 在现况调查时, 抽查某市的几个居委会的全部居民, 这叫做单级整群抽样, 如果在某市抽查几个居委会, 每个居委会不全查, 而抽查几个居民小组, 这就是二级整群抽样, 显然, 还可以有多级整群抽样。整群抽样的优点是容易组织, 缺点是由于各群间差异较大, 抽样误差也较大, 故统计效率较

低。为此, 整群抽样的样本需大于单纯随机抽样, 才能达到同等的精确度。整群抽样的群间越一致越好, 群内的差异与总体一致。从统计效率看, 整群抽样在符合上述原则条件下, 群内容量宜小些, 群数宜多些。整群抽样的样本含量计算公式列如表6。

表6 整群随机抽样样本含量的计算公式

适用条件	计算公式
对总体率作估计	$K = \frac{u^2 S_c^2}{m^2 \delta^2} \quad S_c^2 = \frac{\sum m_i^2 (P_i - P)^2}{K - 1} \quad (11)$
对总体均数作估计	$K = \frac{u^2 S_c^2}{m^2 \delta^2} \quad S_c^2 = \frac{\sum m_i^2 (\bar{X}_i - \bar{X})^2}{K - 1} \quad (12)$

例6 某市调查肺结核患病率, 预查5个居民委, 得方差 $S_c^2 = 10112.37$, 每居民委平均人口数 $\bar{m} = 1381$, 设允许误差 $\delta = 0.02$, $\alpha = 0.05$, 问需调查多少群?

查 u 值表得 $u_{0.05} = 1.96$, 将有关数据代入公式

(11) 得
$$K = \frac{1.96^2 \times 10112.37}{0.02^2 \times 1381^2} = 50.92$$

即需查51个居民委，已查5个，再查46个。

(2) 各种抽样的标准误的计算：抽样研究要通过样本来推断总体，样本率（或均数）是对总体率（或均数）的点估计，由于抽样误差的存在，对总体值的

估计还需做区间估计，率的(1-α)可信区间用下式计算

$$P \pm u_{\alpha} S_p \quad (13)$$

均数的(1-α)可信区间用下式计算

$$\bar{X} \pm t_{\alpha} S_{\bar{x}} \quad (14)$$

不同抽样方法的标准误的计算公式是不同的，见表7。

表7 标准误的计算公式

抽样方法	率的标准误 S_p	均数的标准误 $S_{\bar{x}}$
单纯随机抽样	$\sqrt{\frac{P(1-P)}{n}} \quad (15)$	$S/\sqrt{n} \quad (16)$
分层随机抽样	$\sqrt{\sum \left(\frac{N_i}{N}\right)^2 \frac{P_i(1-P_i)}{n_i}} \quad (17)$	$\sqrt{\sum \left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i}} \quad (18)$
整群随机抽样	$\frac{1}{m} \sqrt{\frac{\sum m_i^2 (P_i - P)^2}{K(K-1)}} \quad (19)$	$\frac{1}{m} \sqrt{\frac{\sum m_i^2 (\bar{X}_i - \bar{X})^2}{K(K-1)}} \quad (20)$

当研究总体中的个体数N是有限的，样本含量n与总体含量N比值小于0.1时，可把有限总体近似地看成无限总体；若该比值大于0.1时，计算标准误后应乘以 $\sqrt{1 - \frac{n}{N}}$ ，整群抽样时计算标准误后乘以 $\sqrt{1 - \frac{k}{K}}$ 。

例7 若例5中五年前患病率值为最优分配分层抽样的调查结果，试求率的标准误。

计算过程如表8。

将表8中有关值代入公式(17)得

$$S_p = \sqrt{\left(\frac{60000}{140000}\right)^2 \frac{0.0475}{428} + \left(\frac{50000}{140000}\right)^2 \frac{0.09}{491} + \left(\frac{30000}{140000}\right)^2 \frac{0.24}{481}} = 0.0082$$

若采用单纯随机抽样，结果P=0.143，其标准误按公式(15)计算得

$$S_p = \sqrt{\frac{0.143 \times (1 - 0.143)}{1400}} = 0.0094$$

此结果较分层随机抽样的结果为大。

读者可自行对按比例分层抽样的情况进行计算，其 $S_p = 0.0086$ ，介于上述两 S_p 值之间。

例8 某市调查肺结核患病率，预查随机抽取5个居民委，结果见表9，试求总体肺结核患病率的95%可信区间。

$$P = \frac{37}{6905} \times 100\% = 0.5358\% \quad \bar{m} = 6905 \div 5 = 1381$$

将P, \bar{m} 及表中有关数据代入公式(19)得

$$S_p = \sqrt{\frac{40449.4797}{1381^2 \times 5 \times (5-1)}} = 0.0326\%$$

表8 分层随机抽样率的标准误计算

职业	人口数 N_i	抽查人数 n_i	患病率 P_i	1-Pi	$P_i(1-P_i)$
甲	60 000	428	0.05	0.95	0.0475
乙	50 000	491	0.10	0.90	0.0900
丙	30 000	481	0.40	0.60	0.2400
合计	140 000	1 400	0.143		

表9 整群抽样中率的标准误计算

居民委	人口数 m_i	患者数 C_i	患病率(%) P_i	$m_i^2(P_i - P)^2$
A	1512	8	0.5291	102.6250
B	1437	9	0.6263	16912.6124
C	1026	6	0.5848	2527.4751
D	1732	9	0.5196	787.2738
E	1198	5	0.4174	20119.4934
合计	6905	37	0.5358	40449.4797

再用公式13计算总体肺结核患病率的95%可信区间为

$$0.5358\% \pm 1.96 \times 0.0326\% = (0.4719\%, 0.5997\%)$$

(3) 研究计划：

①研究目的：研究目的必须明确，一般只设单目

标。现场的情况复杂，因素不能控制，要弄清真相实为不易，所以通常皆为单目标研究。而多目标研究只有在很必要且实现条件充分时才慎用。探索性研究的目的多为通过对疾病发生的强度及分布的了解，提出进一步分析性研究的假说。

②观察指标及测定方法：对疾病的诊断要使用明确统一公认的标准。观察指标力求选用客观的、特异的计量指标。测定方法宜选用灵敏度与特异度高的方法。观测误差要控制在一定的范围之内。

③研究对象：对研究的总体在时间、空间、人群的范围要给出明确的规定。在抽样研究中还应规定抽样方法和样本含量。

④调查项目：力求符合现场的实际情况，应有的项目具体、不漏，无需的项目不含。为此，宜先于现场做小范围的预查后，精心设计。调查项目包括分析项目及一般项目。在探索性研究中分析项目包括诊断、发病日期、性别、年龄、住址、与发病有关的因素等以便分析疾病发生的频率、三间分布及有关因素。设立一般项目是为了一旦分析项目有错漏情况时便于核查，并加以更正补充。比如，姓名、户主姓名、诊断依据、调查者等均属于此类项目。

⑤调查表：当调查内容简单、调查项目少时，调查表可采用一览表的形式（每个调查对象填写一行）。当调查项目较多时，可用个案调查表（或卡）。调查记录力求用“√、×”符号，以求快准，便于统计。有问答题应列出一切可能答案类型，也用“√、×”符号记录。对于拟用电子计算机做统计分析时，可用编码卡片，即给每一个分析项目一个编码，将编码以数据文件形式输入计算机，应用专用分析程序进行统计处理。调查员要事前培训。

⑥整理分析：整理和分析哪些统计指标以及如何整理分析都要事先有所计划。探索性研究的常用统计指标有发病率、患病率和死亡率等，按三间分布进行多维和一维分析，考察其分布特征及有关因素的关联程度。

2.筛检 (Screening)：筛检是通过快速、简便的筛检试验、检查或其他手段，在健康人群中去发现那些未被识别的病人或有缺陷的人。筛检试验不是诊断试验，仅是一个初步检查，对筛检试验阳性者必须进一步确诊，对确诊病人或有缺陷的人采取必要的治疗措施和预防措施。

(1)筛检工作的目的：有治疗和预防两个方面，通过疾病的筛选早期发现处于临床前期或早期的病人，

达到早期治疗的目的。比如筛检糖尿病病例时，用查尿糖为筛检试验，尿糖阳性者再作进一步确诊治疗。另一方面，通过对危险因子的筛检，可使某些病达到预防的目的。比如，筛检高血压和吸烟的人群，然后消除危险因子，达到预防冠心病、脑卒中的发生。又如，在孕妇中筛检乙型肝炎携带者，及早对其新生儿进行乙肝被动—自动免疫来预防乙肝的发生。

(2)筛检的必要性和可行性：评价筛检计划包括疾病、筛检试验、筛检计划的合适性。具体内容有：①被筛检的疾病或缺陷是一个重要的健康问题，有确切的诊断和治疗的方法，早治比较效果好，确实能达到早发现、早诊断、早治疗、早预防的作用；②有快速、经济、有效的筛检方法，该法无创、无痛、无害，易为群众接受；③有条件和能力对筛检后的可疑者提供进一步的诊断和治疗；④筛检、诊断、治疗全过程的经济消耗小于由此产生的收益。具备以上诸点，方能达到目的，且具有实施价值。

(3)筛检方法的评价：筛检方法除要具备快速、经济、群众易于接受的条件外，在质量上还要求真实性和可靠性。真实性用灵敏度和特异度评价。灵敏度为在有病者中查出阳性（有病）的比例，特异度为在无病者中查出阴性（无病）的比例。在考核一种筛选试验时，通常对一组已知有病和另一组无病的人进行检查，结果可归纳成表10。

表10 评价筛检试验资料的归纳表

筛检试验	有病	无病	合计
阳性	A	B	A+B
阴性	C	D	C+D
合计	A+C	B+D	N

常用的反映真实性的统计指标有五个，分别用下式计算：

$$\text{灵敏度 (真阳性率)} = \frac{A}{A+C} \times 100\% \quad (21)$$

$$\text{特异度 (真阴性率)} = \frac{D}{B+D} \times 100\% \quad (22)$$

$$\text{总一致率} = \frac{A+D}{N} \times 100\% \quad (23)$$

$$\text{调整一致率} = \frac{1}{4} \left(\frac{A}{A+B} + \frac{A}{A+C} + \frac{\bar{D}}{C+D} + \frac{D}{B+D} \right) \times 100\% \quad (24)$$

$$\text{约登指数 (Youden's index)} = \frac{A}{A+C} + \frac{D}{B+D} - 1 \quad (25)$$

理想的筛检方法，上述五个指标均为100%，但实际很难，事实是要灵敏度高就特异度低，或者反之，特异度高就灵敏度低，这是因为正常人的试验数值分布与病人的数值分布有重叠部分的缘故。筛检试验的阳性界限选定的原则是①对一些严重疾病，如早发现早治早防效果好，晚治晚防后果严重，宜选敏感度高的阳性分界线，防止筛漏；②对疗效不理想，确诊和治疗费用较贵的疾病，可选择特异度较高的阳性分界线。为了提高灵敏度或特异度，可采用几种方法并联使用，只要其中一个阳性即判为阳性，以提高灵敏度；也可采用几种方法串联使用，几种试验均为阳性才判为阳性，以提高特异度。

可靠性是某筛检试验在相同条件下，进行重复试验，获得相同结果的稳定程度。为了做到这一点，则要试验药品质量好，温度、湿度等环境条件稳定，观测者技术熟练，受试者状态稳定。

(4) 筛检效果的评价：评价筛检效果好的指征有轻型或临床前期的比重大、治愈率高、后遗症率低、生存率高、病死率低等。但就筛检试验本身效率来讲，要考察筛检试验阳性的人中有多大比例是真无病的（即阳性预测值），在筛检试验阴性的人中有多大比例是真无病的（即阴性预测值）。计算公式为

$$\text{阳性预测值} = \frac{A}{A+B} \times 100\% \quad (26)$$

$$\text{阴性预测值} = \frac{D}{C+D} \times 100\% \quad (27)$$

根据Bayes的条件概率理论，可得到下列公式

$$\text{阳性预测值} = \frac{\text{患病率} \times \text{灵敏度}}{\text{患病率} \times \text{灵敏度} + (1 - \text{患病率}) \times (1 - \text{特异度})} \times 100\% \quad (28)$$

$$\text{阴性预测值} = \frac{(1 - \text{患病率}) \times \text{特异度}}{(1 - \text{患病率}) \times \text{特异度} + \text{患病率} \times (1 - \text{灵敏度})} \times 100\% \quad (29)$$

当然阳性预测值越大效率越好，从公式(28)可显见，当灵敏度、特异度固定，患病率越高，阳性预测值越高，而患病率对阴性预测值影响不大，这就是为什么筛检试验宜用于患病率高的疾病或在患病率高

的人群中进行的原因。

例9 某医院应用某筛检试验对50万人进行肝癌筛检，结果如下表，试对筛检方法及效果评价。

表11 筛检肝癌的结果

筛检结果	肝癌病人	无肝癌病人	合计
阳性	99	24 995	25 094
阴性	1	474 905	474 906
合计	100	499 900	500 000

应用公式(21)~(27)计算得

$$\text{灵敏度} = \frac{99}{100} \times 100\% = 99\%$$

$$\text{特异度} = \frac{474905}{499900} \times 100\% = 95\%$$

$$\text{粗一致率} = \frac{99 + 474905}{500000} \times 100\% = 95\%$$

$$\text{调整一致率} = \frac{1}{4} \left(\frac{99}{100} + \frac{99}{25094} + \frac{474905}{499900} + \frac{474905}{474906} \right) = 73.60\%$$

$$\text{约登指数} = \frac{99}{100} + \frac{474905}{499900} - 1 = 0.94$$

$$\text{阳性预测值} = \frac{99}{25094} \times 100\% = 0.39\%$$

$$\text{阴性预测值} = \frac{474905}{474906} \times 100\% = 99.9998\%$$

求患病率得

$$\text{患病率} = \frac{100}{500000} \times 10\text{万}/10\text{万} = 20/10\text{万}$$

本筛检方法灵敏度为99%，特异度为95%，比较理想，但用于患病率很低的肝癌人群筛检，阳性预测值只有0.39%，查出25 094个阳性者中只有99个是真肝癌病人，说明效率是极低的。事实是在99名肝癌病人中只有20%可手术，手术后一年生存率只有40%，这又说明无论从效果还是效益上看，也都是不合算的。这项筛检工作是得不偿失的。

(5) 筛检的经济评价：包括成本效果分析、成本效益分析等，有关分析方法将于有关讲中介绍。

3. 历史法 (historical method)：又称现成资

料的再分析。现成的流行病学资料包括疫情、职业病报告卡、死亡报告卡、慢性病报告卡、个案调查卡、报表、专题报告等。这些资料以原始数据卡所含信息量多，报表由于降维，从而损失一定信息。随着时间的推移，资料不断积累，实有对现成资料再分析的必要，研究疾病模式演变规律、评价计划免疫实施的现场效果，疫势发展的趋势性和预测等，都大有可为。在利用历史资料时，首先要了解疾病的诊断标准，有无漏错，资料的可靠性和质量，哪些资料可利用。在整理资料时，把时间、空间、对象和有关因素联系起来整理成复合表，然后去粗取精，去伪存真，进行加工。对比分析时，要注意资料的可比性，遇两组资料年龄构成等差别太大时，则计算标准化率后再比较。当有诸多因素混杂时，可进行层析分析与多因素分析。电子计算机已日益普及，可应用Foxbase将慢性病发病卡片和死亡卡片数据分别形成数据库，然后进行记录联结(record Linkage)，做生存率分析。

4. 疾病监测(disease surveillance)：疾病监测是对疾病的分布及其有关因素进行连续的、细致的观察和监视，促使对疾病的控制更加有效和完善。疾病监测属于纵向的探索性研究，监测方法有面上一般人群的监测和建立监测点，二者互补，以点带面。在监测点上，系统地、规范地收集疾病资料，并通过健全基层防保机构、培训专业人员、制定工作规范和程序，开展漏报调查和定期考评，来进行质量控制。在动态地收集与分析疾病监测资料的基础上，提出假设，开展专题性研究，评价干预措施的效果，探索疾病分布特征及其有关因素，为不断改进和优化防病的对策和措施提供科学依据，以期达到更有效地控制以至消除疾病的目的。我国疾病监测工作已有长足发展。自1980年起，在卫生部和中国预防医学科学院领导下，由流行病学微生物学研究所牵头，以自愿报名参加方式逐步建立了全国综合疾病监测网，至1988年已有71个城市及农村监测点，起到了示范点的作用。1990年，在全国进行分层整群随机抽样组成了包括145个疾病监测点的全国疾病监测网，使其资料更能代表全国，为制订我国卫生防疫决策及措施更具有科学价值和指导作用。

可见，疾病监测是预防医学实践向深度和广度发展的重要步骤，它能及时、准确地提供信息，实时分析反馈信息，为疾病控制和决策服务。正像医院建立病房标志着医院水平提高一样，具有代表性的全国疾病监测网的建立是我国卫生防疫工作新的里程碑，是

卫生防疫工作走向科学化、现代化、正规化的重要标志。

四、形成假设的若干法则：通过探索性研究，可对某病的病因或有关流行因素(包括危险因子与保护因子)提出假设。首先要充分占有资料，包括该病宏观的流行病学资料，微观的理化、微生物检测和病理资料以及亚宏观的临床资料，然后进行科学的分析与综合，从客观实际出发，经过逻辑思维，形成假设。以下就形成假设时常用的若干法则，加以阐明。

1. 求同法：假如某病发生在两个或多个场合，而这些场合有一个共同点，则这仅有的共同点，就可能是所研究疾病的原因(或效应)。

例10 一个单位发生食物中毒，该单位有四个食堂，有三个食堂卖螃蟹给职工吃，在这三个食堂就餐者中都有食物中毒发生，另一个食堂没有吃螃蟹，也没有食物中毒发生，则螃蟹可能就是引起中毒的食物。

2. 求异法：某一场合发生某病，但另一场合不发生此病，除某一特点只存在于前一场合以外，这两个场合的其他各种情况相同，那么，这唯一的区别点，就可能是该病发生的原因或部分原因。

在上例中，三个发生食物中毒的食堂的共同点是吃螃蟹，依据求同法，可认为螃蟹可能就是引起中毒的食物，一个食堂未发生食物中毒，与发生食物中毒的三个食堂区别点是没有吃螃蟹，依据求异法，则进一步加强了螃蟹是引起中毒的可疑食物的可能性。

3. 共变法：当某病的发生随着某因素的变化而变化，则这个因素可能就是该病的原因(或结果)。

例11 在一次原因不明的下肢麻痹症流行的调查中，发现该病病例较多的家庭计划用粮差，食用山厘豆豆粉的比重大，曾食山厘豆量越多的患者，病情越重，依据共变法提出了山厘豆可能是引起这次下肢麻痹症的病因假设，后来该假设得到了证实。

4. 排除法：正如临床医生要注意鉴别诊断一样，流行病学医生对所获得的现场资料，运用专业知识认真加以分析，思路应广泛展开，可提出几个假设，然后再把与客观事实相矛盾的假设逐一排除，最后求得一个或几个假设。

例12 1972年7月上海市郊县广大地区，突然发生大量皮炎病例，形成爆发流行，皮炎的病原体是什么东西？提出了病原体在水中、与化工厂排出废气有关、霉菌、植物毛或花粉、臭虫、蚤、蚊、蠓、蠊叮咬、“痒辣子”、桑毛虫等许多假设，后经调查分析或试验，

一一给以否定,最后证实了一个假设,即这次流行的皮炎是寄生在桑树和杨树上的桑毛虫毒毛所引起的。

5. 类比法:一种病因不明疾病的分布与某已知病因的疾病分布相似,则这两种病可能属同类性质的疾病。

例13 某地爆发原因不明疾病,这种病的分布特征有严格的季节性(于6~7月发生),主要分布在农村,而且患者多在远离交通线、非城镇地区,其分布特征与流行性乙型脑炎相似,遂提出这种未知病可能是一种自然疫源性疾病。后来经过进一步调查,证明果然是一种自然疫源性疾病,为钩端螺旋体病。

6. 多维析因分析法:疾病的分布本来就是时间、空间、对象与有关因素多维的互相联系的动态分布,所以进行多维析因分析是形成假设的一种重要方法。而只作一维的边缘分布(时间分布、地区分布等),由于损失了许多信息,则在复杂的现场往往难找出头绪来。这正象拍胸部正位X光平片就不如拍断层片提供的信息多。

例14 某工地发生腹泻,患者在地区上分布广泛,人群年龄、性别分布也广泛,从疫情发生到调查已历时七天,疫情呈上升趋势,在食堂就餐与不在食堂就餐的发病频率也差不多。由于弄不清原因,请专家组

协助调查。专家组到现场后,绘制现场地图,并把资料做成居住地区、时间、人群三维复合表加以分析,结果立即发现最早发生的病人皆在一号井周围居住,饮用一号井水,以后逐渐扩大到一号井附近的二号井、三号井的周围居民,进而又扩大到四号、五号井周围的居民患病。这是一个由单源井水污染、居民饮用生井水造成腹泻爆发,向多个水井污染发展进而成为多源性爆发的例子,由于一周后的情况已发展成复杂的多源性,孤立地分析三间分布是看不清问题的。进行多维分析与有关因素联系的析因分析,是弄清来龙去脉的捷径。有时可疑的有关因素有多个,可按诸因素一切可能组合分层分组,考察各组与发病情况的联系,进行层析分析(析因分析)。

7. 推理法:结合专业知识,研究分布特征,进行推理,形成假设。

例15 六十年代初,大连市庄河县发生间日疟疾爆发流行,笔者应邀赴现场调查,根据当地的气温资料,结合间日疟原虫生活史的生态学知识,新发病例应最早出现在7月上旬,而该县5~3月份已有相当数量的新病例,这是用短潜伏期间日疟无法解释的,遂提出当地有长潜伏期间日疟存在的假设,以后的研究证实了这一假设。

本刊举办“高级现代流行病学培训班”的紧急通知

近年来,流行病学发展很快,现职卫生人员流行病学知识亟待更新与补充,为了满足广大卫生人员现代流行病学知识水平能与学科发展同步的迫切要求,本刊拟举办一期“高级现代流行病学培训班”。

一、时间:1994年4~7月,共4个月。

二、方法:该班以函授为主,拟聘请章扬熙主任医师等专家任教。采取做习题、发考卷、给标准答案等方法开展教学活动。

学习内容、考核及结业方式见本刊1993年第4期第256页。

三、报名及学费:报名日期为1993年10月15日~1994年3月31日。请用楷体写明姓名、性别、年龄、职称、单位、详细通讯地址及邮政编码。报名同时邮寄学费(含资料费)92元(开收据、报销)。一律寄至北京昌平流字五号《中华流行病学杂志》编辑部刘玉书收(邮政编码102206)。款到寄资料,名额有限,按报名顺序录取,未被录取者一律退款。

《中华流行病学杂志》编辑部