进入率与样本的偏性

上海第一医学院流行病学教研组 苏德隆

流行病学研究的对象往往是总体的一个部份，即所谓样本。样本是否适当，主要看它对总体有无充分的代表性。如果样本对总体的代表性不高，则不可能通过调查分析得出一个一般正确的结论，也就不能达到调查分析的原定目标。

影响样本代表性的因素很多。本文仅讨论中其的一个——进入率（admissionrate）。例如，总体中具有某种特性的有100个，如果其中有10个进入研究样本，则进入率为$\frac{10}{100}=10\%$，经过医院诊断的某种病患仅是某病患者总体的一部份。由于病情不同的患者入院的机会——入院率也是进入率的一种——不是固定的，不能想象入院病人的病情及其他条件与未入院的同病患者相同，故也不能代表同病患者的总体。

本文主要用概率论的方法证明进入率能影响疾病与其他事件的联系，举例如下(四格表)：

城乡脊髓灰质炎住院病人后遗症分析

	乡区病人 A	城区病人 $\bar{A}$	合计	乡区病人比例
有后遗症者B	55	45	100	$P_1=.55$
无后遭症者$\bar{B}$	258	642	900	$P_2=.29$
合 计	313	687	1,000	

如果凭这个四格表的x^2检验结果，将会做出这样一个结论："根据现有资料，入院的脊髓灰质炎患者有后遗症的比例与病人之来自城乡的比例有联系"。但我们要问：这项经验能否一般化地反映医院外广大城乡居民中脊髓灰质炎后遗症的比例与患者的住址有联系呢？可能脊髓灰质炎后遗症的有无与城乡的住址无关，但在住院病人中则有联系的表现。

我们不可只凭手中数据的计算结果而下结论，还要看其他因素。在一个交通不便，缺医少药的农村，一个患脊髓灰质的病人被送入城市医院治疗的机会，和城市一个同病患者入院的机会有很大的差别。在同一医院中，城乡患者的病情也可能不同。在此情况下，一个城区医院治疗脊髓灰质炎的经验总结就不能反映城乡广大居民间的一般情况。

现在用概率论的方法来验证。

设：B表示一个患脊髓灰质炎并有后遗症的事件。

$\bar{B}$表示一个患脊髓灰质炎但无后遗症的事件。

P(B)表示城乡所有居民中患脊髓灰质炎并有后遗症的比例。

$P(\bar{B})=1-P(B)$表示城乡所有居民中患脊髓灰质炎但无后遗症者的比例。

A表示一个人住在乡区的事件。

$\bar{A}$表示一个人住在城区（即非乡区）的事件。

P(A)表示乡区居民中占的比例，

$P(\bar{A})=1-P(A)$表示城区居民占的比例。

P(AB)表示全部居民中乡区居民患脊髓灰质炎并有后遗症者的比例。

$P(AB)=P(A)P(B)$如果A和B是独立的。

H表示一个人住医院的事件。

$P(H\mid BA)$表示乡区患脊髓灰质炎并有后遗症者住院的比例。

$P(HB\bar{A})$表示城区患脊髓灰质炎并有后遗症者住院的比例。

$P(H\mid\bar{B}A)$及$P(H\mid B\bar{A})$的定义可仿上定出。我们要求的P_1和P_2的值：

$P_1=P(A\mid BH)$,

$P_2=P(A\mid\bar{B}H)$,

P_1：住院患脊髓灰质炎并有后遗症者来自乡区的比例。

P_2：住院患脊髓灰质炎但无后遗症者来自乡区的比例。

$$P_1 = P(A \mid BH) = \frac{P(BA \mid H)}{P(B \mid H)}$$

分母是住院患脊髓灰质炎并有后遗症的比例；分子是住院患脊髓灰质炎并有后遗症者来自乡区的比例。这分子按贝依斯Bayes原理作如下的推论。

$$P(BA \mid H) = \frac{P(H \mid BA)P(BA)}{P(H)}$$

$$= \frac{P(H \mid BA)P(B)P(A)}{P(H)}$$

因假定A和B是独立的

$$P(BA) = P(B)P(A)$$

P_1的分母，$P(B \mid H)\dfrac{P(H \mid B)P(B)}{P(H)}$

求$P(H \mid B)$时，我们利用一个事实即总率是各概率的加权均值。

那么$P(H \mid B) = P(H \mid BA)P(A \mid B) + P(H \mid \overline{BA})P(A \mid B) = P(H \mid BA)P(A) + P(H \mid \overline{BA})P(A)$

因假定A和B是独立，可利用$P(A \mid B) = P(A)$以及$P(\overline{A}) \mid B) = P(\overline{A})$的关系。因而：

$$P = \frac{P(H \mid B)P(B)}{P(H)}$$

$$= \frac{P(B)\left[P(H \mid BA)P(A) + P(H \mid B\overline{A})P(\overline{A})\right]}{P(H)}$$

$$P_1 = \frac{P(BA \mid H)}{P(B \mid H)}$$

$$= \frac{\dfrac{P(H \mid BA)P(B)P(A)}{P(H)}}{\dfrac{P(B)P(H \mid BA)P(A) + P(H \mid B\overline{A})P(\overline{A})}{P(H)}}$$

$$= \frac{P(H \mid BA)P(A)}{P(H \mid BA)P(A) + P(H \mid \overline{BA})P(\overline{A})}$$

同样 $P_2 = \dfrac{P(H \mid \overline{B}A)P(A)}{P(H \mid \overline{B}A)P(A) + P(H \mid \overline{B}\overline{A})P(\overline{A})}$

两式中的$P(A)$是乡区人口的比例，设为80；$P(\overline{A})$，城区人口的比例为.20。

P_1和P_2两式除$P(A)$和$P(\overline{A})$外，余为4个入院率。病人入院率不是固定的，而是随各种条件而变的。4个入院率变动，则P_1和P_2亦随之变动。我们可以做一个实际计算，如果要得到如原题中的P_1和P_2的值，只要使4个入院率做如下安排就行了。

设：

乡区居民患脊髓灰质炎并有后遗症的入院率，$P(H \mid BA) = .15$。

城区居民患脊髓灰质炎并有后遗症的入院率，$P(H \mid B\overline{A}) = .50$。

乡区居民患脊髓灰质炎但无后遗症的入院率，$P(H \mid \overline{B}A =) .02$。

城区居民患脊髓灰质炎但无后遗症的入院率，$P(H \mid \overline{B}\overline{A}) = .02$。

代入公式：

$$P_1 = \frac{.15 \times .80}{.15 \times .80 + .50 \times .20} = \frac{.12}{.22} = .55$$

$$P_2 = \frac{.20 \times .80}{.20 \times .80 + .20 \times .20} = \frac{.016}{.056} = .29$$

如果4个入院率做别的安排，则P_1和P_2就随之不同。

脊髓灰质炎后遗症的比例与地区的关系，可以通过精心设计的现场调查得到。而不能用住院病人——经过选择的样本——推算。

上面讨论了从医院的经验得出两个事件有某种联系时，不一定能正确的反映医院外广大人群间有同样的联系存在。有时相反的差错也可能发生。即在医院的小天地里，两个事件之间不见有联系。则在社会人群中则有联系存在。举例如下：

某血吸虫流行区的血吸虫病患者居住在乡间的比例为.60，住在城市的比例为.40，非血吸虫病患者住在乡间的比例为.20，住在城市的为.80来自乡间患血吸虫病而入某寄生虫病医院治疗的概率为.0033，住在城市患血吸虫病而入某医院治疗的概率为.009。乡间居民非患血吸虫病而入院的概率为.0025；城市居民非患血吸虫病而入院的概率为.00114。

设某流行区血吸虫病患者共有10万人，非患者共有100万人。

先分析患血吸虫病的10万人：

1、住在乡间的有：100,000×.60＝60,000人；

2、这批人入某医院治疗的有：60,000×.0033＝198人；

3、住在城市患血吸虫病的有：100,000

×.40=40,000人；

4、住在城市患血吸虫病者入某医院的有：40,000×.09=360人。

再分析非患血吸虫病的100万人：

1、住在乡间的有：1,000,000×.20=200,000人。

2、这批人入某医院的有：200,000×.0025=500人。

3、住在城市非血吸虫病患者有：1,000,000×.80=800,000。

4、这批人入院的有：800,000×.00114=912人。

以上数据可构成四格表如下：

某医院血吸虫病患者与非患者住址分析

	来自乡间	来自城市	合计	来自乡间的比例
血吸虫病患者	198	360	558	$P_1=.3548$
非血吸虫病患者	500	912	1412	$P_2=.3541$

从四格表的检验结果看来，住院治疗的血吸虫病患者的比例与他们的住址设有联系。但在医院外的广大人群中，患血吸虫病与住址之间有联系存在。因为：患血吸虫病者住在乡间的比例为.60，非血吸虫病者住在乡间的比例为.20。在本例，医院的经验同样不能反映广大人群中间的真实情况。因经过这个医院筛选出来的样本对总体不具有代表性。下面再举几个有关进入率的例子。

某院医生说：“今年乙型脑炎较去年的凶险。因我院今年的乙脑病人的病死率为14%，而去年只有12%。”这句话并没有说明什么。因为即使这两年的乙脑在凶险的程度上没有显著不同，亦可能由于其他种种原因——包括入院率——使两年的住院乙脑病人的病死率的表面值不同。

尸体解剖的材料也必须十分谨慎地使用。因尸解的进入率的变异特别显著。半个世纪前，一位赫赫有名的生物统计学家R.Pearl引用大量的尸解材料之后说，在患癌死亡的尸体中，结核病变的发现较在非因癌而死的尸体中为少，因而推想结核与癌可能有互相制约的作用。不久之后他自已发现这项推论不妥而收回了他的意见。因为除非所有的死者都以同等机会进入尸群范围。则将尸解病例中所发现的联系推广（外推）到活人中去，是不适当的。很可能在活人中没有联系。而在经过选择病例进行尸解之后，在尸解例中有显著的联系存在。

举例说，某妇产科医院总结一年来住院产妇因难产而丧失生命的占30%。而某地段接生站在同期中没有一个产妇因难产而死亡。如果按产院和接生站的难产妇死亡的表面数值分析的话，可得四格表如下：

	产妇死亡	产妇存活	合计	假设死亡的比例
产院难产妇	a	b	a+b	$P_1=\frac{a}{a+b}=.30$
接生站难产妇	c	d	c+d	$P_2=\frac{c}{c+d}=0$

如果下结论说，某妇产科医院处理难产的技术水平比一个地段接产站还不好，那将是不真实的。凭常识也能想到难产妇入妇产专科医院和地段接生站的机会——进入率——很不相同；相对的说，接生站愿收留难产的事是极少的，可谓条件过严；产科医院接收难产是责无旁贷的，可谓条件过宽。这两单位所收的难产妇的比例均不代表广大社会中难产占产妇人数的比例。现在这两单位所收的难产妇的难产程度很不相同，故难产死亡率不能相比。

启东的鸭也患肝癌。1973年某节日杀鸭取鸭肝检验肝癌，雌雄鸭的肝癌检出比例如下。

	检验鸭数	患肝癌鸭数	检出比例（%）
雄　鸭	1,111	6	0.54
雌　鸭	135	37	27.20

从这项资料的表面看来，雌鸭患肝癌的比例高出雄鸭50倍。但仔细推敲，就发现这项结论是不妥当的。因为被宰杀的雄鸭往往是年龄小的；而雌鸭往往是年老而不生蛋的或因病（癌？）而不生蛋的。另有证据显示，鸭的年龄越高患肝癌的机率越大，送去宰杀的“进入率”雌雄不同，故不可能自节日宰杀之鸭推算

雌雄的肝癌患病率有何不同。

回顾性调查和前瞻性调查也有进入率的问题。对照组和“试验组”不仅在年龄、性别等条件上要求均衡。两组成员进入两组的进入率亦须相等。任何组的成员如果是经过某种方法挑选的，有时是受调查者热心主动要求调查，有时是经过反复劝说才肯接受调查的，这些情况都将影响进入率。

例如有人用经过医院诊断治疗的或报告的“肝病”、“肝硬化”“迁慢肝”和“黄疸型肝炎”的病例为研究对象，进行回顾性配对调查或倒推起点日期的前瞻性调查，结果发现调查的“肝病”病例“转归”为肝癌的相对危险性远高于对照组，因而做出肝炎与肝癌有关的种种推论。但不能说自广大地区随机抽查生产大队的肝炎（不包括“肝病”）病例转归为肝癌的比例也是如此。不然的话，就不会出现这么一个现象：肝炎在我国普遍流行，而肝癌仅限于在某些局部地区高发。研究肝炎与肝癌有无因果关系时，关键在于不使在取样本的方法上出现偏差。经过医院选择的肝炎病例（即使不包括肝炎以外的“肝病”的话）不能代表广大人群中肝炎病例的总体。用一些经过高度选择的病例为研究对象，就不能希望其结果和在广大地区随机抽样的调查分析的结果相一致。

各地医院收留肝炎病人的标准和条件不同，病例报告的完整性不一致。各地肝炎病人要求入院治疗的比例亦不相同。由于各地医院的肝炎病例不能算为肝炎病例的“完善”样本，故各地区的研究结论，往往既不一致，也不能反映广大人群间的真实情况。但必须了解，不是说经过医院诊断为“迁慢肝”“肝硬化”、“肝肿大”等与肝癌无关，只是说这些经过特殊选择的“肝病”不能代表广大人群中的“肝炎”总体。因此也不能根据住院的肝炎和“肝病”的转归推出“肝炎与肝癌有因果关系”的结论。

调查的结论不可仅凭数据的统计分析。要记得，从调查的第一步起——即从选择调查对象之时起，陷井是到处有的，偶一不慎就可能得出荒谬的结论。

参 考 文 献

1. Berkson J：Biometrics Bull（Now Biometrics），2：47，1946.
2. Mianland D：Elementary Medical Statistics，2nd ed P 117，114，322，W B Saunders，Philadelphia 1963.
3. Fliss JL：Statistical Methods for Rates and Propo，rtions，p 3～13，J Wiley，New York，1973.

“百白破”对预防百日咳的流行病学效果观察

邬加生* 李劲云* 黄敬享** 郭予宋* 王云清* 史习舜**

在没有可能用空白对照组来考核百日咳的保护效价的情况下，我们对卢湾区（共九个街道104个里委）每一个街道抽查二个里委共18个里委，于1980年1月把1974～77年出生的全部常住户口学龄前儿童共11079人，作为观察对象，采用回顾性的前瞻性调查方法，按实际接种“百白破”混合菌苗前后，对“百白破”未接种及不同接种情况分成六个组，分别计算其暴露人月。凡观察对象中出现的百日咳病人也以发病时实际完成的“百白破”接种情况列入上述相应的组，以分别计算“百日咳”的保护效价。

凡能完成全程，虽未到加强年龄，其保护效价可达56.6%，未全程者（只注射一针或二针）其保护效价仅25.8%；凡能按照正规要求加强者，其保护效价可达76.1%，不按照正规要求，但已加强者，其保护效价可在53.3%，全程后未加强者，其保护效价仅36.6%，说明上海市目前使用的“百白破”混合制剂，对百日咳尚有一定的保护作用，但效果不太理想，远逊于白喉的预防效果。

本资料显示：严格按照计划程序开展“百白破”的免疫接种，是应该坚持的，并应引起广大医防人员的重视。

* 上海市卢湾区卫生防疫站

** 上海第一医学院流行病学教研组