讲 座

疾病聚集性分析

上海医科大学* 詹绍康

在流行病学研究中，人们常对某种疾病是否有聚集性感兴趣，希望通过调查，分析某疾病在时间、空间、人群或家庭的聚集性。例如：

1.某种先天畸形发生率在某些年份是否异乎寻常的高；

2.生活或工作在某地区的人是否更易患某种病；

3.某些家庭的某病发病率是否高于一般。

在分析这些问题时，要用到一些专门的统计方法，以鉴别所观察到的现象是否可能纯属随机或巧合的原因。

时间聚集性

疾病的时间聚集性，可以表现为：某一年或几年的发病率高于一般；某一个月、几个月或若干天的发病率高于一般或呈周期性季节变化；一天中某些时间的事件发生率高于一般或呈24小时为一个周期的周期性变化。

一、列联表的χ^2检验：如果把时间按某种方式分成若干个(往往都取相等间隔)区间，那么在发病率相等的假设下，理论上该人群在各时间区间的发病率与区间长度成比例。假定人口数不变或基本不变，就可以认为在理论上发病数与区间长度呈比例。

例如，有人抽样调查了某省1963年到1969年脊髓灰质炎发病情况，其发病率如表1，问该病是否有时间上的聚集性（各年的发病率是否都相等）？

如果不存在时间聚集性，就意味着各年份的发病率相等。在统计学上可以假设各总体率相等，以一般的2×k表的χ^2检验来对此假设作检验。首先，可用合计的发病率$\left(\frac{221}{5936846}\right)$与各年的人口数相乘算得发病的期望数，并用减法算得不发病的期望数（表2），然后用公式（1）计算χ^2值。

表1 某地1963～1969年脊髓灰质炎发病率

年 份	1963	1964	1965	1966	1967	1968	1969
发病数	49	31	24	73	19	17	8
人口数	839316	841834	845201	847737	851128	854107	857523
发病率(/10万)	5.84	3.68	2.84	8.61	2.23	1.99	0.93

$$\chi^2=\Sigma\frac{(A-E)^2}{E} \qquad (1)$$

表2 对表1资料作χ^2检验用表

年 份		1963	1964	1965	1966	1967	1968	1969	合计
发病	观察数A	49	31	24	73	19	17	8	221
	期望数E	31.2	31.3	31.5	31.6	31.7	31.8	31.9	221
不发病	观察数A	839267	841803	845177	847664	851109	854090	857515	5936625
	期望数E	839284.8	841802.7	845169.5	847705.4	851096.3	854075.2	857491.1	5936625

此χ^2值服从于自由度为k−1（本例为7−1）的χ^2分布。对于表2资料，可得：

* 邮政编码 200032

$$\chi^2=\frac{(49-31.2)^2}{31.2}+\frac{(839267-839284.8)^2}{839284.8}$$

$$+\cdots\cdots+\frac{(857515-857491.1)^2}{857491.1}$$

$$=96.07$$

查表得$\chi^2_{0.01}(6)=16.81$，故$\chi^2>\chi^2_{0.01}$，$P<0.01$，拒绝原假设，认为这7年的总体发病率并不全相等，或者说时间上有聚集性。

如果相比较的各年人口数变化不大，特别是对于发病率极低的疾病，不发病项对公式（1）中χ^2值的贡献很小，为了简化计算，可把不发病项省去，即为：

$$\chi^2=\frac{(49-31.2)^2}{31.2}+\frac{(31-31.3)^2}{31.3}$$

$$+\frac{(24-31.5)^2}{31.5}+\frac{(73-31.6)^2}{31.6}$$

$$+\frac{(19-31.7)^2}{31.7}+\frac{(17-31.8)^2}{31.8}$$

$$+\frac{(8-31.9)^2}{31.9}=96.07$$

在确实人口数变化小而发病率极低的情况下，可近似地把发病数看作Poisson分布变量，而应用下面的χ^2检验：

$$\chi^2=\frac{\Sigma(x-\bar{x})^2}{\bar{x}} \qquad (2)$$

式中x表示观察（发病）数；$\bar{x}$表示观察（发病）数的平均数。χ^2值服从自由度为k－1的χ^2分布（k为变量值x的个数）。对表1资料，可得：

$$\bar{x}=\frac{221}{7}=31.6$$

$$\chi^2=\frac{1}{31.6}[(49-31.6)^2+(31-31.6)^2$$

$$+(24-31.6)^2+(73-31.6)^2$$

$$+(19-31.6)^2+(17-31.6)^2$$

$$+(8-31.6)^2=95.05$$

χ^2值与前相差甚微，结论不变。

二、周期聚集性检验：可以用周期性回归分析法来判断疾病的发生有无周期性现象。当所用的资料是完整的周期（如12个月或24个月）时，其回归方程式为：

$$\hat{y}=b_0+b_1\sin\theta+b_2\cos\theta \qquad (3)$$

式中$\hat{y}$表示估计的发病数；θ表示观察时点（或时间区间的中点）所对应的角度；b_0，b_1和b_2表示回归系数。

如果观察资料是12个月的发病资料，第i个月对应的近似角度是$\theta i=(i-\frac{1}{2})\left(\frac{360}{12}\right)$；如果观察资料是24个小时的畸胎儿出生资料，第i个小时对应的角度是$\theta i=(i-\frac{1}{2})\left(\frac{360}{24}\right)$。计算θ的一般形式为：

$$\theta i=(i-\tfrac{1}{2})\left(\frac{360}{M}\right) \qquad (4)$$

式中M表示一个完整的周期中所划分的时间区间个数。用公式（4）时要求各区间的时间长度相等。由于各月份包含的天数不相等，所以这里各月对应的角度θ只是一种近似，即：1月：$\theta_1=(1-\frac{1}{2})\left(\frac{360}{12}\right)$

$=15$（度）；2月：$\theta_2=(2-\frac{1}{2})\left(\frac{360}{12}\right)=45$（度）

……；12月：$\theta_{12}=(12-\frac{1}{2})\left(\frac{360}{12}\right)=345$（度）。

例如，某医师为研究脊髓灰质炎在某年内有无时间聚集现象（即各月的发病频率是否一致），收集了如下资料（表3）。

对于这种一年内包含12个月的完整周期的资料，配合回归方程〔式（3）〕的方法如下。

表3　某年某地脊髓灰质炎发病资料

月　份	1	2	3	4	5	6	7	8	9	10	11	12	合计
发病数(y)	70	89	92	85	64	42	19	4	0	8	31	48	552
角　度(θ)	15	45	75	105	135	165	195	225	255	285	315	345	

$$b_0=\bar{y} \qquad (5)$$

$$b_1=\frac{1}{6}\Sigma y\sin\theta \qquad (6)$$

$$b_2=\frac{1}{6}\Sigma y\cos\theta \qquad (7)$$

对本例：$b_0=\frac{552}{12}=46$

$$\Sigma y\sin\theta=70\times\sin15^\circ+89\times\sin45^\circ$$

$$+\cdots\cdots+48\times\sin345^\circ$$

$$=258.3269$$

$$\Sigma y\cos\theta=70\times\cos15^\circ+89\times\cos45^\circ$$
$$+\cdots\cdots+49\times\cos345^\circ$$
$$=95.7096$$

$$b_1=\frac{1}{6}\times258.3269=43.0545$$

$$b_2=\frac{1}{6}\times95.7096=15.9516$$

因而得周期性回归方程：

$$\hat{y}=46+43.0545\sin\theta+15.9516\cos\theta$$

对总体回归系数$\beta_1=\beta_2=0$的假设作检验，可用公式：

$$\chi^2=\frac{2}{\Sigma y}\left[(\Sigma y\sin\theta)^2+(\Sigma y\cos\theta)^2\right] \quad (7)$$

此χ^2值服从于自由度为2的χ^2分布。

对本例：

$$\chi^2=\frac{2}{552}\left[258.3269^2+95.7096^2\right]$$
$$=274.98.$$

因$\chi^2_{0.01}(2)=9.21$，$\chi^2>\chi^2_{0.01}$，$P<0.01$。拒绝β_1和β_2都为0的假设。认为脊髓灰质炎在各月份的发病数呈周期性变化。而对于下面的表4资料：

表4 某年某地先天缺陷新生儿出生数

月份	1	2	3	4	5	6	7	8	9	10	11	12	合计
出生数	41	22	35	24	20	9	37	19	33	16	16	28	300

可算得：

$$\hat{y}=25+1.7878\sin\theta+3.3261\cos\theta$$

$$\chi^2=\frac{2}{300}(10.7267^2+19.9563^2)=3.42$$

因$\chi^2<9.21$，$P>0.05$。故认为先天缺陷新生儿的出生数并无季节差异。

如果打算把各月份所对应的角度算得确切一些，可按表5所列步骤进行。第（7）列就是各月份中点所对应的确切角度，与表3所列的近似角度略有差别。

表5 各月份所对应的确切角度的计算表

月份	天数	累计天数	(3)下移一格	(2)×1/2	(4)+(5)	(6)/365×360
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	31	31	0	15.5	15.5	15.28
2	28	59	31	14	45	44.38
3	31	90	59	15.5	74.5	73.48
4	30	120	90	15	105	103.56
5	31	151	120	15.5	135.5	133.64
6	30	181	151	15	166	163.73
7	31	212	181	15.5	196.5	193.81
8	31	243	212	15.5	227.5	224.38
9	30	273	243	15	258	254.47
10	31	304	273	15.5	288.5	284.55
11	30	334	304	15	319	314.63
12	31	365	334	15.5	349.5	344.71

空间聚集性

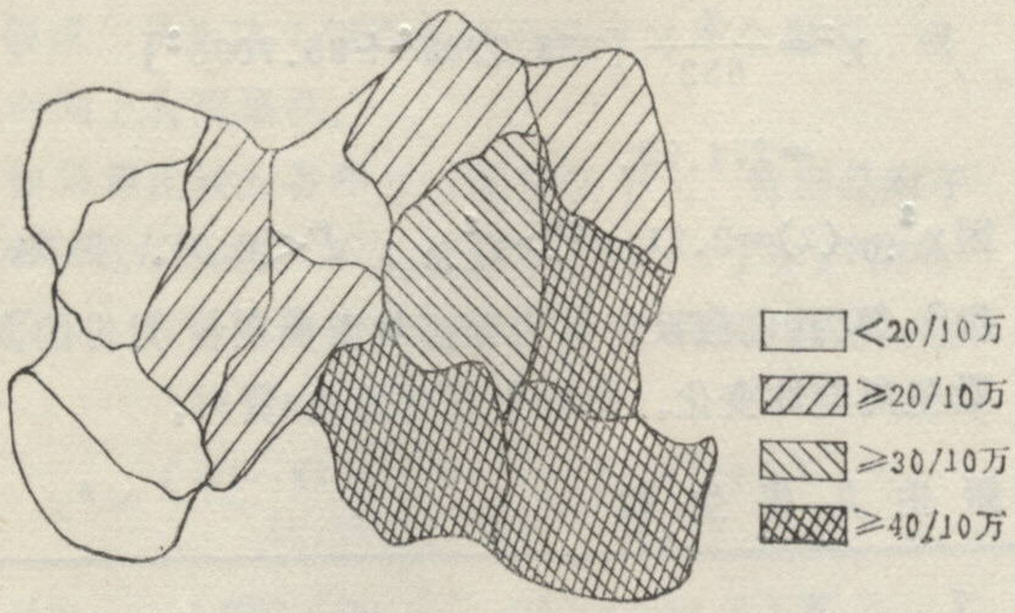
疾病的空间聚集性，就是指各地区疾病发病频率的不一致性。常用于分析各乡、各县、各区或各市某病发病率的差别。

一、发病率的地区差别：如果要分析某疾病在各行政区域（乡、区、市、县等）有无聚集现象，由于一个行政区域内包含的人口数往往都有几万、几十万或几百万，在一般情况下，发病率或患病率的抽样误差都很小，所以常省去统计学检验，直接用所得资料分析地区差别。由于许多疾病的发病率受人口的性别、年龄构成的影响，用按性别与年龄作调整的率（标准化率）作比较，效果会更好一些。表6是13个地区居民动脉硬化性心脏病死亡率资料。从表中数据可见，在CV、ZN和ZC地区居民的动脉硬化性心脏病死亡率比较高。

表6 某年13个地区居民动脉硬化性心脏病死亡率

地区	JN	JT	ZP	ZC	YD	CV	LY	WM	DZ	LC	TA	ZN	HZ
死亡率(/10万)	25.2	34.3	19.2	36.2	21.4	53.8	19.4	16.6	29.5	23.3	12.5	35.3	15.3
标化死亡率(/10万)	23.5	32.1	21.2	43.1	20.8	52.4	21.7	18.3	26.4	21.9	16.6	44.6	17.7

用统计表来表达疾病的地区分布的优点是可以表达得很精确，缺点是并没有完全把疾病的空间分布表达出来，不能一目了然。统计地图是表达疾病空间聚集现象的最简单方法。把表6资料画成附图，就可更清楚、直观地看到疾病的聚集现象。



附图　某年13个地区居民动脉硬化性心脏病死亡率

二、病例的空间分布：如果要研究某疾病在小范围内的空间分布，由于人口的居住密度不同，所以就不能以划方格取等面积的地域为单位按 Poisson 分布原理对病例分布的随机性作简单的分析。这里介绍一种不必考虑人口分布的统计分析方法。

假定某乡有n名病例，要研究是否有空间聚集性。先从该乡同样范围内的非病例中随机抽取m人作为对照，在n+m名个体中，反复随机抽取含量为n的样本，在每一个含量为n的样本中，总共可组合成C_n^2=n(n−1)/2对，分别测量每对个体间的距离，清点此距离小于某个指定值（如1公里）的对子数r，从这些反复抽取的随机样本中，可得r的分布。对n名病例也可清点距离小于该个指定值的对子数r_D，以r_D与r的分布相比，看r_D是否落在r分布的小概率区域内。一些统计学者称此为排列检验(permutation test)。感兴趣的读者可参阅Lloyd和Roberts 1973年在Brit.J.Prev.Soc.Med.发表的论文及Smith和 Pike 1974年在同一杂志发表的论文。

时间和空间的聚集性

在某些情况下，流行病学者希望同时研究某种疾病的时间和空间的聚集性。这里以Knox在1964年发表的资料为例，介绍一种相当简便的方法。

Knox在英格兰某一规定区域某一规定时间范围内调查了96例儿童白血病患者。规定两例相距不满1公里时称为空间的“近”，否则称为“远”；两例发病时间间隔不满60天时称为时间的“近”，否则称为“远”。由于有96例患者，所以可以有C_{96}^2=(96×95)÷2=4560种组合的对子，调查结果得表7。

表7　儿童白血病时间及空间聚集性检验用表
(Knox, 1964)

		空间		合计
		近	远	
时间	近	5	147	152
	远	20	4388	4408
合计		25	4535	4560

对表7资料不宜用2×2表的χ^2检验。如果不论在时间方面还是在空间方面，“近”的比例都很低，那就可近似地用Poisson分布原理来作检验。发病若与时间和空间的分布无关，那么两者都“近”的期望数为(25×152)/4560=0.83。把0.83看作poisson分布的均数λ，可判断表7左上角的频数5的出现是否为小概率事件。以λ=0.83，计算两者都近的对子数x为0，1，2，3，4，5，……的概率分别为：

$$P_0=e^{-0.83}=0.4360$$

$$P_1=0.4360\times0.83=0.3619$$

$$P_2=0.3619\times0.83/2=0.1502$$

$$P_3=0.1502\times0.83/3=0.0416$$

$$P_4=0.0416\times0.83/4=0.0086$$

$$P_5=0.0086\times0.83/5=0.0014$$

$$P_{(x\geq6)}=1-(P_0+P_1+P_2+P_3+P_4+P_5)$$
$$=0.0003$$

由于$P_{(x\geq5)}=0.0014+0.0003=0.0017$，所以认为表7左上格观察频数出现5是一种小概率事件。结论是认为有时间和空间的聚集性。

家庭聚集性

疾病的家庭聚集性表现为不同家庭发病率的差异，一些家庭的某病发病率高于其他家庭。造成家庭聚集性的原因可以是：疾病的遗传性或传染性，家庭成员的相似生活条件或行为。然而，现实社会中各种因素的同时存在往往会使家庭聚集性的研究工作复杂化。

如果调查和收集资料的方法不同，那么统计分析方法也不同。调查方法主要可分为两类。

A类：通过家庭作调查。在调查或收集资料时，以家庭为基本单位。可以调查某一地区的全部家庭，也可以从全部家庭中随机抽取一部分。

B类：通过病例作调查。在调查中，通过病例来收集他们所在家庭的资料。如果研究者打算估计病例的近亲中患病的比例，那么人们常把这些最初确定作为线索的病例称为先证者（Probands或Propositi）或指示病例（index case）。这种资料的一个明显特点是没有把无病例的家庭包括在内。通过病例作调查还可以分为下面几种情况。

1.完全分析：在某些情况下，一个地区的全部病例都是先证者，不管一个家庭内有几名病例（先证者），分析中只用一次。有时，可以从户籍登记部门或其他登记部门获得先证者名单，一个家庭只会出现一次（如6个月内死胎死产登记等）。

2.单一分析：由于抽样比例很小，即病例中的很小一部分作为先证者。所以在所有家庭中也只有很小一部分列为调查对象。这样，同一家庭被调查2次或2次以上已不大可能，可略而不计。

3.不完全有重复分析：通过一些病例（先证者）来调查他们的家庭时，有可能在同一家庭里的几个病例都是先证者，因此这一部分家庭在收集资料时会重复数次。

按调查方法的不同而把资料作上述区分是很重要的。因为对不同的资料应该用不同的统计方法。首先，要假定家庭人口数为常数n，每名个体发病的概率相等。在A和B（1）两种情况下，病人数为r的家庭数近似于二项分布，但在B（1）情况下缺少r＝0的家庭数（Fr）。在B（2）情况下，有r个患者的家庭的概率与r成比例，所以有r个患者的家庭数在理论上与rPr成比例（Pr是二项分布中阳性数取值为r时的概率）。在B（3）情况下，显然会使患者多的家庭被抽样的可能性大。

由上述分析可见，对疾病的家庭聚集性调查和分析方法是比较复杂的。由于某些研究人员不熟悉其正确的调查统计方法，所以经常出现一些差错，下面是两个误用的例子。

例1：某研究者对90名M病患者（作为先证者或指示病例）及90名对照分别调查了他们的家庭成员，得如下资料（表8）。

作者的结论是：病例组家庭成员中(包括本人)M病发生率为8.07%(105÷1301)，而对照组为0.47%（6÷1289），说明M病的发生有明显的家庭聚集现象（$\chi^2=91.3$，$P<0.001$）。

由于本例中作者把至少有一个病例的家庭作为病例组家庭，而指示病例又仍算作此家庭内的病例数之中，显然就会使病例组家庭内发病率（或患病率）增高，如果把两组中的指示病例与对照者都删去，即在总人数（1301与1289）及病例组的病例数（105）中各扣除90，可得四格表(表9）。

表8　　M病的病例-对照研究结果

家庭内病例数	病例组			对照组		
	户数	病例数	总人数	户数	病例数	总人数
0	0	0	0	86	0	1215
1	78	78	1134	3	3	60
2	11	22	158	0	0	0
3	0	0	0	1	3	14
4	0	0	0	0	0	0
5	1	5	9	0	0	0
合计	90	105	1301	90	6	1289

表9　　对表8与资料作修改后的四格表

	病例组家属	对照组家属
病例数	15	6
非病例数	1196	1193
合　计	1211	1199

$\chi^2=3.80$　　$P>0.05$

没有理由说病例组家属与对照组家属M病发生率（或患病率）不同。与原来的结论相反。

例2：某研究者以100例某肿瘤病例为先证者（指示病例），在他们的近邻（无肿瘤者）中随机取100人为对照，分析两组的家属（不包括指示病例或对照本人）中该肿瘤的发病率，结果病例组家属该肿瘤发生率高于对照组家属，所以认为该肿瘤有家庭聚集性。

对于这个问题，许多统计学者已有过专门的研究。正确的方法如下：如果某病在某地诸家庭中分布确属二项分布，那么全部病例的家属（一家有2个病例，家属要重复计算）和全部非病例的家属（同一家庭内的非病例也作为家属）的患病率是相等的。这里的非病例也包括指示病例家庭内的非病例。如按此方法调查分析，以5口之家为例，理论上可得表10模式。

表10　　5口之家作病例-对照研究的统计模型

病例数	家庭数	指示病例数	家属数	家属中病例数	家属非病例数	对*照数	家属数	家庭中病例数	家属非病例数
X	F	f=xF	4f	(x−1)f	(5−x)f	(5−x)F	4(5−x)F	X(5−x)F	(5−x)(5−x−1)F
0	q^5N	0	0	0	0	$5q^5N$	$20q^5N$	0	$20q^5N$
1	$5q^4pN$	$5q^4pN$	$20q^4pN$	0	$20q^4pN$	$20q^4pN$	$80q^4pN$	$20q^4pN$	$60q^4pN$
2	$10q^3p^2N$	$20p^3p^2N$	$80q^3p^2N$	$20q^3p^2N$	$60q^3p^2N$	$30q^3p^2N$	$120q^3p^2N$	$60q^3p^2N$	$60q^3p^2N$
3	$10q^2p^3N$	$30q^2p^3N$	$120q^2p^3N$	$60q^2p^3N$	$60q^2p^3N$	$20q^2p^3N$	$80q^2p^3N$	$60q^2p^3N$	$20q^2p^3N$
4	$5q\ p^4N$	$20qp^4N$	$80qp^4N$	$60qp^4N$	$20qp^4N$	$5qp^4N$	$20qp^4N$	$20qp^4N$	0
5	p^5N	$5p^5N$	$20p^5N$	$20p^5N$	0	0	0	0	0
合计	N	5pN	20pN	$20p^2N$	20qpN	5qN	20qN	20qpN	$20q^2N$

* 对照数=各家庭内非病人数总和

N表示某地区总的家庭数，p表示总患病率，q=1−p。

病例家属患病率=$20p^2N/20pN=p$

对照家属患病率=$20qpN/20qN=p$

由此可见，按这样的方式作调查分析时，在二项分布的假设下，病例与对照的家属患病率是相等的。特别应引起注意之点是：如果不把病例的家属也看作对照之总体的一部分，就会使对照家属的患病率偏低，从而会导致不正确的结论。

一起由病死马内脏引起鼠伤寒沙门氏菌食物中毒传染的调查报告

呼和浩特市卫生防疫站*　柴崇山　唐恩辉　徐素云

本市土左旗某村一村民的自养马于6月6日因发热、腹泻而死。畜主当日即剥皮、取其内脏煮食，并且有生熟刀案不分的情况。此两日后，畜主及其亲友相继发病，共66例。

流行病学调查：调查24户进食者全部发病，未食病死马内脏共49户，发病3户。进食者6月8日突然发病，6月9日达到高峰，11日停止流行，共计54例，发病率96.4%（54/56），未进食者249人，发病12例（与中毒病人有密切接触史，如有4个婴儿吃患者的母奶），发病率4.8%。

临床表现：潜伏期7～60小时，平均16小时，多为头痛、头晕、发热38～40℃，寒战、恶心、呕吐、腹痛、腹泻，多者日泻10余次，少数里急后重，周身关节疼痛，个别病人出现嗜睡、抽搐等症状。病程2～6天，预后良好，无死亡病例。

实验室检验：对马肠、肝和熟内脏经增菌、分离培养和双糖铁，均检出可疑沙门氏菌。三种标本培养出的菌株其生化反应相同。用沙门氏菌因子血清做凝集试验，AFO多价、单因子O4及Hi凝集均为阳性，盐水对照无自凝现象，可确定为鼠伤寒沙门氏菌。血清免疫学试验：食物中毒后20天，取当地10名病人和6名正常人血清，与食物中检出的细菌作凝集反应，其凝集效价病人明显高于正常人。

讨论：本次有12人未吃病死马内脏，由于与患者有密切接触史而感染，发生与中毒患者相同的症状，这与许多报道本菌中毒不进食者无发病，且停食处理后再无新发病例出现有所不同。据此我们认为人与人之间是可以传染的，这在采取预防措施时是不可忽视的。

* 邮政编码 010020