流行病学常用的统计方法

Ⅱ.定群研究的设计与统计分析

辽宁省卫生防疫站 章扬熙

定群研究（cohort study），又称队列研究或群组研究，是分析性流行病学研究方法之一。这种研究是把特定人群区分为暴露于某因素和非暴露于某因素两个组，或按暴露水平分成若干个组，然后观察与比较各组发病频率（或死亡频率）的差异，从而考察因素对发病（或死亡）的影响有无及影响程度的大小。研究因素可以是危险因素，也可以是保护因素。这是一种从因到果的研究，多数的定群研究是从现在的因来追踪将来的果（发病或死亡），所以称前瞻性定群研究；若研究是基于过去的因追查现在的果（发病或死亡），则属于回顾性定群研究。以下就定群研究设计中的若干问题及常用的统计分析方法作一介绍。

定群研究设计的若干问题

1.策略的选择：定群研究能直接考察因素与发病（或死亡）的联系和联系程度，结果较为可靠；因素的作用可分若干个暴露水平，便于计算“剂量-反应关系”。所以，通常在分析性流行病学的研究中，先作病例对照研究，当找到有明显意义的因素后，再作定群研究。由于定群研究是从因到果的研究，对因素单一、果出的快、结得多的场合尤为适宜。

2.暴露组的选择：由于所研究的因素在人群中多早已存在，故可根据因素的情况来选择暴露于该因素的人群。比如，职业因素可选择职业人群。暴露水平要明确规定分级标准或进行定量，把暴露人群分为若干组，但最好要有高度暴露于某因素的组，以利考察因素与疾病的关系。暴露剂量既要考虑每日（或次）平均暴露剂量，也要考虑接触天数（次数）的积累暴露剂量，可用二者的综合指标，比如铀矿中放射性物质剂量，用工作能级月（WLM）来表示，并以此来分析剂量反应关系。

3.非暴露组的选择：与暴露组比较的对照，可以设非暴露组、内对照或人群对照等。非暴露组除不暴露于研究因素外，在其他主要特征如年龄、性别、住地、民族 等均与暴露组一致，以利对比。有时定群研究的对象是一个群体，通过调查暴露因素，把整个群体按暴露水平等级分成若干组，其中非暴露的组或暴露水平最低的组则为该群体的内对照组，内对照组除研究因素外，在其他主要特征方面可能与暴露组不一致，可进行分层分析或多因素分析来解决。在职业病流行病学研究中，常用人群对照，二者年龄、性别的差异，可通过比较二者的标化率或计算预期发病（死亡）数来解决。在定群研究中，考察某因素与某病的联系同时，也可考察该因素与其他病的联系是否存在，如果该因素仅与某病有联系，而与其他病无联系，说明二者相关的特异性，从而增强了判断的依据。若设立多种对照且结果一致，将更有说服力。无论非暴露组，还是暴露组，研究对象宜选不流动，与研究密切合作者，以减少在随访中的失访情况与偏性。

4.样本的大小：定群研究的样本大小可用下式（公式1）计算

$$N=\frac{(K\alpha\sqrt{2\bar{P}\bar{Q}}+K\beta\sqrt{P_1Q_1+P_2Q_2})^2}{(P_2-P_1)^2}$$

这公式与病例对照研究的样本大小公式形式相同，但P_1、P_2的含意不同。这里P_1、P_2分别为暴露组与非暴露组的发病率。$Q_1=1-P_1$。$Q_2=1-P_2$，$\bar{P}$为两组发病率的平均值，即$\bar{P}=(P_1+P_2)/2$，$\bar{Q}=1-\bar{P}$，如已知非暴露组的发病率P_2及相对危险度RR，暴露组的发病率P_1可用下式（公式2）求得

$$P_1=RR\times P_2$$

α、β分别为第一类误差和第二类误差，K为标准正态分布的横坐标，Kα为双侧时，每侧尾面积为α/2，Kα为单侧时，单尾面积为α。Kβ则皆为单侧。

〔例1〕 为了考察母亲孕期使用雌激素与所生子女患先天性心脏病的联系，拟进行定群研究。估计非暴露组的发病概率$P_2=0.002$，即每1000活产中有2例先天性心脏病，使用雌激素引起所生子女患该病的相对危险度RR=2，设$\alpha=0.05$，$\beta=0.10$，求样

本大小N。

已知$P_2=0.002$，RR＝2，代入公式2，得

$P_1=RR\times P_2=2\times0.002=0.004$

$Q_1=1-P_1=0.996$

再求$\bar{P}=\frac{0.002+0.004}{2}=0.003$

$\bar{Q}=1-\bar{P}=0.997$

按双侧面积，查标准正态曲线下的面积表(或t值表中df＝∞时的t值)，知$K\alpha=K_{0.05}=1.96$，按单侧时，$K\beta=K_{0.1}=1.282$，将有关数据代入公式1，得

$$N=\frac{\left(1.96\sqrt{2\times0.003\times0.997}+1.282\sqrt{0.002\times0.998+0.004\times0.996}\right)^2}{(0.004-0.002)^2}$$

$=15716$

即暴露组与非暴露组各需约15720人。用定群研究所需样本通常比病例对照研究要大得多，这是因为定群研究中各组的发病率比病例对照研究中各组的暴露比例小得多的缘故。

定群研究的统计分析

1.定群研究资料的模式：对定群研究所得资料的整理，依两组研究对象在随访观察期间有无变化而具有两种不同的模式。如果两组例数在观察期间没有变化，则可用各组的例数作分母计算其发病率，这种发病率又叫累积发病率（或累积死亡率）。其资料整理表格如表1。

表1　定群研究累积发病率资料整理表

组　别	发病	未发病	合计	发病率
暴露组	ai	bi	N1i	ai／N1i
非暴露组	ci	di	Noi	ci／Noi
合　计	M1i	Moi	Ti	M1i／Ti

如果两组例数在随访观察期间，由于失访、死亡、新补充等原因不断变化，这就需要计算人时发病率（或人时死亡率），时间可用日、周、旬、月、季和年为单位，但通常为年。这种率又称发病密度。其资料整理表格如表2。

表2　定群研究发病密度资料整理表

组别	发病	人时数	发病率
暴露组	ai	N1i	ai/N1i
非暴露组	ci	Noi	ci/Noi
合计	M1i	Ti	M1i/Ti

应用人时发病率来分析通常比累积发病率为好，但计算较繁。

2.人时的计算：人时的概念是人×时间，比如2个人观察了4年，则为2×4＝8人年。计算人年的方法，常用的有以个人为单位来计算人年和应用寿命表计算人年等方法。

(1) 以个人为单位来计算观察人年数：比如，某人1915年4月9日生，观察从1941年11月11日开始，到1953年12月1日停止。共观察了12.06人年。如果按年龄组段25岁～、30岁～、35岁～等来计算每年龄组段的观察人年数，则为25岁～组，观察期间为41.11.11～45.4.8计3.41人年，30岁～组，为45.4.9～50.4.8，计5人年，35岁～组为50.4.9～53.12.1，计3.65人年。这种方法计算人年比较精确，但较繁，可用电子计算机来完成。

(2) 应用寿命表法来计算人年：在定群研究中，观察期间的研究对象在不断变动，有失访、死亡、新补充的情况，这时若用寿命表法来计算人年，较为方便。

〔例2〕某定群研究中，观察期限为12个月，分别计算暴露组与非暴露组的人年数，其中暴露组的观察情况如表3的第(1)～(5)栏，试求该组的人年数。

本例为按月观察，人时为人月。观察人月数L_x用下式（公式3）计算

$$L_x=l_x+\frac{1}{2}(N_x-D_x-W_x)$$

下一个月的月初人数l_{x+1}用下式（公式4）计算

$$l_{x+1}=l_x+N_x-D_x-W_x$$

本例第一个观察月的观察人月数L_1为

$$L_1=1100+\frac{1}{2}(72-4+15)=1126.5$$

第二个月的月初人数l_2为

$l_2=1100+72-4-15=1153$

余类推。把各观察月的人月数相加，得总观察人月数为14146人月，折算成人年为14146÷12＝1178.83人年。

3.两组发病率的假设检验：研究因素与研究的疾病是否有联系，需作假设检验，常用的方法为χ检验（或称U检验、Z检验）。用下式（公式5）计算

$$\chi=\frac{a-M_1N_1/T}{\sqrt{M_1N_1N_0/T^2}}$$

〔例3〕　对日本广岛原子弹爆炸幸存者中的白血病发病情况进行了16年的前瞻性定群研究，结果如

表3　　应用寿命法计算人年数

观察时间（月数）	月初人数	月内进入人数	月内发病人数	月内离去人数	观察人月数
x	lx	Nx	Dx	Wx	Lx
(1)	(2)	(3)	(4)	(5)	(6)
1	1100	72	4	15	1126.5
2	1153	18	3	11	1155
3	1157	20	2	7	1162.5
4	1168	14	2	5	1171.5
5	1175	11	1	8	1176
6	1177	8	0	3	1179.5
7	1182	7	1	5	1182.5
8	1183	10	0	2	1187
9	1191	10	1	4	1193.5
10	1196	11	2	5	1198
11	1200	15	1	6	1204
12	1208	12	1	7	1210
合计			18		14146人月
					(1178.83人年)

表4，试问两组发病率的差别有无显著意义？

检验假设为两组总体发病率相等，$\pi_1=\pi_2$，然后求χ值，把有关数值代入公式5，得

$$\chi=\frac{61-88\times306500/1221500}{\sqrt{88\times306500\times915000/1221500^2}}=9.57$$

χ的界值同U的界值，查U值表，$U_{0.05}=1.96$，$U_{0.01}=2.58$，所得$\chi>U_{0.01}$，所以$P<0.01$，拒绝

表4　　广岛原子弹爆炸后幸存者中白血病的发病率

组别	患者数	人年数	发病率(/10万)
暴露组	61	306500	19.90
对照组	27	915000	2.95
合计	88	1221500	7.20

检验假设，说明暴露组与对照组白血病的发病率有差别，暴露组高于对照组。

4.危险度分析：当χ检验有显著性差别后，可计算相对危险度（RR）、特异危险度（AR）和特异危险度的百分比（AR%），分别用下式（公式6～8）

$RR=Ie/I\mu=(a/N_1)/(c/N_0)$

$AR=Ie-I\mu=a/N_1-c/N_0$

$$AR\%=\frac{Ie-I\mu}{Ie}\times100\%=\frac{RR-1}{RR}\times100\%$$

上式中Ie为暴露组的发病率，Iμ为非暴露组（对照组）的发病率。

以上所得的是样本统计指标，对总体相对危险度的区间估计，用下式（公式9）求

$RR(1\pm z/\chi)$

对总体特异危险度的区间估计，用下式(公式10)求　$AR(1\pm Z/\chi)$

当求90%可信限时，Z=1.645，求95%可信限时，Z=1.96。

〔例4〕　求例3的RR、AR、AR%及其总体值的95%可信限。

$RR=Ie/I\mu=19.90/2.95=6.75$

$AR=Ie-I\mu=19.90-2.95=16.95$

$$AR\%=\frac{RR-1}{RR}\times100\%=85.19\%$$

这结果说明，暴露于原子弹爆炸后幸存者患白血病的危险为对照组的6.75倍，在暴露组发病率19.90/10万中有16.95/10万归因于暴露于原子弹爆炸，因暴露于原子弹爆炸造成的发病率占整个发病率的85.19%。总体RR的95%可信限用公式9求得为

$6.75^{(1\pm1.96/9.57)}=4.57\sim9.98$

总体AR的95%可信限用公式10求得为

$16.95(1\pm1.96/9.57)=13.48\sim20.42$

应用总体RR的95%可信限代入公式8得总体AR%的95%可信限为

$$\frac{4.57-1}{4.57}\times100\%\sim\frac{9.98-1}{9.98}\times100\%$$

$=78.12\%\sim89.98\%$

5.人群特异危险度（PAR）分析，特异危险度（AR）不考虑人群暴露于危险因素的比例，如果人群中暴露于该危险因素的比例很小，尽管AR很大，人群因该危险因素而患病的也不会多，所以需计算人群特异危险度（PAR）及人群特异危险度百分比（PAR%），所用公式依次为公式11及公式12

$PAR=It-Iu$

$$PAR\%=\frac{It-I\mu}{Iu}\times100\%=\frac{Pe(RR-1)}{Pe(RR-1)+1}\times100\%$$

式中It为全人群某病的发病率（或死亡率），Iu为非暴露人群某病发病率（或死亡率），Pe为人群中暴露于某因素的比例。（注，$It=PeIe+(1-Pe)Iu$）

PAR%的标准误应用下式(公式13)计算

$$S_{PAR\%}=\sqrt{\frac{CT[A(N_0-C)(T-C)+(N_1-A)C^2]}{M_1^3N_0^3}}$$

总体PAR%的95%的可信限用下式(公式14)计算

$$PAR\% \pm 1.96 \times S_{PAR\%}$$

〔例5〕仍以例3为例，求PAR、PAR%及其总体值的95%可信限

$$Pe=\frac{N_1}{T}=306500/1221500=0.2509$$

$$It=7.20/10万 \quad Iu=2.95/10万$$

$$PAR=It-Iu=7.20/10万-2.95/10万$$
$$=4.25/10万$$

说明人群中原子弹爆炸所致的白血病发病率为4.25/10万 $PAR\%=\frac{It-Iu}{It}\times100\%=\frac{7.20-2.95}{7.20}\times100\%=59.03\%$，说明人群中的白血病患者有59.03%是由于原子弹爆炸的后果。

$$S_{PAR\%}=\sqrt{\frac{27\times1221500\,[61\times914973\times(1221500-27)-306439\times27^2}{88^3\times915000^3}}$$
$$=0.0656$$

总体PAR%的95%可信限为

$$0.5903\pm1.96\times0.0656=0.4617\sim0.7189$$
$$=46.17\%\sim71.89\%$$

6.剂量反应关系的分析：在定群研究中，把暴露水平计量或分为等级来分组，如果剂量与反应呈正相关，则进一步增强了这种联系。

〔例6〕对日本广岛原子弹爆炸幸存者，按T_{65}照射剂量分组，进行16年前瞻定群研究来观察白血病的发病情况，结果如表5，试进行分析。

表5 不同T_{65}照射剂量的幸存者中白血病的发病率(/10万)

T_{65}总剂量范围(拉德)	白血病例数	暴露人年数	发病率(/10万)	RR	AR	AR%
<5	27	915000	3.0	1.0	0	0
5～	8	156000	5.1	1.7	2.1	41
20～	14	67000	20.9	7.0	17.9	86
50～	7	38300	18.3	6.1	15.3	84
100～	10	24100	41.5	13.8	38.5	93
200～	5	9000	55.6	18.5	52.6	95
≥300	17	12100	140.5	46.8	137.5	98
合计	88	1221500	7.2		4.2	58

从表5可以明显看出，无论RR、AR、AR%均与T_{65}总剂量呈正相关，从而进一步支持原子弹的爆炸后果可引起幸存者发生白血病的论点。

7.分层分析：定群研究中采用内对照、人群对照时，暴露组与对照组在年龄、性别等方面往往不均衡，这时要作按年龄、性别的分层分析，以下结合实例介绍分析方法。

〔例7〕 Doll和Hill以英国男医师为对象，应用定群研究观察吸烟与冠心病死亡的关系，按年龄分5层所得结果如表6，试进行分层分析。

表6 吸烟组与不吸烟组冠心病死亡情况

年龄组	吸烟组 死亡数	吸烟组 人年数	吸烟组 死亡率‰	不吸烟组 死亡数	不吸烟组 人年数	不吸烟组 死亡率‰	RRi
35～	32	52407	0.61	2	18790	0.11	5.55
45～	104	43248	2.40	12	10673	1.12	2.14
55～	206	38612	7.20	28	5710	4.90	1.47
65～	186	12663	14.69	28	2585	10.83	1.36
75～	102	5317	19.18	31	1462	21.20	0.05
合计	630	142247	4.43	101	39220	2.58	1.72

首先对两组死亡率进行假设检验，所用公式(公式15)为

$$\chi=\frac{\Sigma ai-\Sigma(N_1iM_1i/Ti)}{\sqrt{\Sigma(M_1iN_1iN_0i/Ti^2)}}$$

本例，检验假设为两组率相等，用上式求χ得

$$\chi=\frac{(32+\cdots+102)-(\frac{52407\times34}{71197}+\cdots+\frac{5317\times133}{6779})}{\sqrt{\frac{34\times52407\times18790}{71197^2}+\cdots+\frac{133\times5317\times1462}{6779^2}}}$$
$$=3.319$$

$U_{0.01}=2.58$，$\chi>2.58$，$P<0.01$，拒绝检验假设，说明两组死亡率有差别。

应用公式6求各层RRi，结果列于表5的第末栏。再用下式(公式16)求总相对危险度

$$RRa=\frac{\Sigma(aiN_0i/Ti)}{\Sigma(ciN_1i/Ti)}$$

本例，

$$RRa=\frac{\frac{32\times18790}{71197}+\cdots+\frac{102\times1462}{6779}}{\frac{2\times52407}{71197}+\cdots+\frac{31\times5317}{6779}}=1.425$$

再求RR的渐近极大似然估计值RR_{m1}，用下式(公式17)求得，用迭代法，初值可用RRa，结果当RR=1.426时，满足公式17的条件。故$RR_{m1}=1.426$。

$$\Sigma ai-RR\Sigma\frac{M_1i}{RR+N_0i/N_1i}=0 \quad (公式17)$$

应用公式9，求总体RRml的95%可信限，得

$1.426^{1\pm 1.96/3.319}=1.156\sim 1.758$

各层的RRi是否一致呢？可用下式(公式18)进行非均匀性检验。检验假设为各层总体RRi相等。

$$\chi^2_{HET}=-2\Sigma\{aiLN\left[\frac{RRM_1i}{ai(RR+N_0i/N_1i)}\right]+CiLN\left[\left(\frac{M_1i}{Ci}\right)\left(1-\frac{RR}{RR+N_0i/N_1i}\right)\right]\}$$

本例，$\chi^2_{HET}=12.132$，自由度=层数$-1=5-1=4$，查χ^2值表，$\chi^2_{0.05,4}=9.49$，$\chi^2_{0.01,4}=13.28$ $P<0.05$，拒绝检验假设，说明各层的相对危险度差别有显著性，不宜求总的相对危险度，还是以各层分别分析为好。35岁组RR最高，RR有随年龄下降的趋势。

8.标准化死亡率比(SMR)与标准化死亡比例比(SPMR)：在职业病流行病学的定群研究中，常采用人群对照，应用间接法进行标准化，以全人群各年龄组的某病死亡率为标准，以某职业(或高发区)人群的各年龄组观察人年数乘以全人群该年龄组某病的标准死亡率求得预期死亡数之和$\Sigma E(ai)$作为分母，以某职业(或高发区)人群在观察期间某病实际死亡数(Σai)为分子所得之比值乘100即为标准化死亡率比(SMR)，用下式(公式19)计算。

$$SMR=\frac{\Sigma ai}{\Sigma E(ai)}\times 100$$

如果不是用全人群某病各年龄组的死亡率乘以某职业人群各年龄组的人年数求E(ai)，而是用全人群某病死亡数占全死因死亡数的构成比乘以某职业人群全死因死亡数来求E(ai)，仍用公式19所得的结果则为标准化死亡比例比(SPMR)。

〔例8〕 某工厂1980～1984年肺癌死亡情况如表7，以全人群为对照，求SMR来考察该厂肺癌死亡是否较全人群为高？

$$SMR=\frac{23}{13.9}\times 100=165.5$$

这个结果说明，该工厂的肺癌死亡数为按年龄用全人群该病死亡率的标准所求得的预期肺癌死亡数的165.5%，比全人群高65.5%。

〔例9〕 某煤矿工人1980～1984年结核病死亡情况如表8，以全人群为对照，求SPMR来考察该矿工人结核病死亡是否较全人群为高？

表7 某工厂肺癌的标准化死亡率比的计算

年龄组(岁)	全人群肺癌死亡率(/10万)	该厂观察人年数	预期肺癌死亡数	该厂肺癌实际死亡数
(1)	(2)	(3)	(4)=(2)×(3)	(5)
20～	0.7	37897	0.3	2
30～	4.4	16342	0.7	2
40～	16.8	9446	1.6	3
50～	57.8	5496	3.2	5
60～	174.3	2519	4.4	6
70～	344.3	1071	3.7	5
合计	—	—	13.9	23

表8 某矿工人结核病的标准化死亡比例比的计算

年龄组(岁)	全人群中结核病死亡占全死因的构成比(%)	观察期间该矿工人死亡数	工人结核病预期死亡数	工人结核病实际死亡数
(1)	(2)	(3)	(4)*	(5)
20～	3.74	49	1.83	4
30～	2.86	41	1.17	3
40～	3.55	62	2.20	5
50～	2.07	140	2.90	6
60～	2.09	157	3.28	4
70～	1.00	283	2.83	5
合计	—	—	14.21	27

*(4)=(2)×(3)

$$SPMR=\frac{27}{14.21}\times 100=190.01$$

这个结果说明，该矿工人结核病的实际死亡数为按年龄用全人群中该病占全死因构成比的标准所计算出的预期死亡数的190.01%，比全人群高90.01%，所以需加强该矿工人结核病的防治工作。