Chin J Epidemiol, June 1998, Vol. 19, No. 3

· 系列讲座 ·

肿瘤流行病学研究资料的统计分析

项永兵

按语: 面对21世纪疾病对人类健康的挑战, 流行病学所担当的角色越来越重要。流行病学做为一门方法学, 在探讨慢性疾病病因与预防方面作出了很大贡献。同时, 其学科自身也得到了逐步发展和完善。这其中生物统计方法在推动流行病学发展方面起了相当大的作用。实际上, 两者是相互促进、共同发展的。

国内在研究统计新方法及其在流行病学方面的应用一直是薄弱环节。所以如何建立起有我国自己特色的流行病学学科, 需要国内同仁艰苦的努力。中华流行病学杂志不仅积极支持我国流行病学事业的发展, 也极力推进统计方法在流行病学研究中的应用。

本次承蒙中华流行病学杂志之邀, 撰写有关肿瘤流行病学研究资料的统计分析方法的文章。内容暂分下述六讲: 发病死亡率的分析和比较、生存率置信区间的估计和比较、调整生存率估计和生存期比较、生存资料单变量分析中危险率比估计、相对生存率的估计、相对生存率的统计学检验(参考文献均备索)。由于本人水平有限, 文中有不当之处请国内同道批评指正。

第一讲: 发病死亡率的分析和比较

通过肿瘤发病死亡资料的积累, 可以为分析流行病学研究提供许多有价值的线索, 还可为社区疾病防治、卫生行政决策部门提供基础数据和决策依据。因此, 本文将根据国际癌症研究中心(IARC)有关专著[1~3], 系统介绍一些发病死亡资料的统计方法。

一、统计方法:

1. 标准化率的估计[1~5]: 用 n_i 表示 i 年龄组病例数, p_i 表示相应年龄组的平均人口数, 若 a_i 为年龄别率, 则年龄调整标准化率(ASR)为

$$ASR = \frac{\sum_{i=1}^{18}(a_i \times w_i)}{\sum_{i=1}^{18} w_i} \qquad (1)$$

式中 w_i 为作为标准的年龄组人口数。(1)式是假设年龄分组为18组(5岁一组)情况下 ASR 估计公式。ASR 通常情况下简称标化率。(1)式是直接标化法计算标化率的公式。如果只利用35～64岁年龄组资料计算标化率, 称为截缩标化率。当某些年龄组病例数太少, 或病例数不明时, 可采用间接标化法, 估计标准化发病率比 SIR(发病资料), 或标准化死亡率比 SMR(死亡资料)。以发病为例, 先计算年龄组期望发病数, 即 $e_i = a_{is} \times p_i / 100000$, a_{is} 为选为标准的年龄组发病率, 则 SIR 估计为

$$SIR = \frac{\sum_{i=1}^{18} n_i}{\sum_{i=1}^{18} e_i} \times 100, \qquad (2)$$

其中 n_i 为年龄组实际发病例数。

2. 置信区间[2~5]: 当 ASR 是用直接法估算时, 它的(100− α)%置信区间(CI)可由下式求得

$$ASR \pm Z_{\alpha/2} \sqrt{\hat{Var}(ASR)} \qquad (3)$$

式中 $Z_{\alpha/2}$ 是标准正态分位数。例如求95% CI 时, 其取值为1.96。$\hat{Var}(ASR)$ 是 ASR 的方差, 可分别在二项($Binomial$)分布或泊松($Poisson$)分布的假设情况下估计, 即

$$\hat{Var}(ASR) = \frac{\sum_{i=1}^{18}\{a_i \times w_i \times (100000 - a_i)/p_i\}}{(\sum_{i=1}^{18} w_i)^2} \text{或}$$

作者单位: 上海市肿瘤研究所流行病学教研室 200032

$\hat{Var}(ASR)=\frac{\sum_{i=1}^{18}(a_i\times w_i\times 100000/p_i\}}{(\sum_{i=1}^{18}w_i)^2}$，其中 a_i、w_i、p_i 分别为按 5 岁一组分组的某疾病年龄别率、标准人口的年龄组人口数、相应时期当地年龄组人口数。

若采用间接法计算标化指标，以 SIR 为例，其 $(100-\alpha)\%CI$ 同样可用(3)式。但方差估计公式不同，见下式

$$\hat{Var}(SIR)=\frac{\sum_{i=1}^{18}n_i}{(\sum_{i=1}^{18}e_i)^2}=\frac{\sum_{i=1}^{18}n_i}{\{\sum_{i=1}^{18}a_{is}\times p_i/100000\}^2},$$

式中 n_i、e_i 为 i 年龄组的实际及期望发病例数，p_i 同上。a_{is} 则为研究者所选标准人口的年龄别率。也可用下述两式建立 SIR 置信区间的下限和上限。

$$\frac{\{\sum_{i=1}^{18}n_i-(Z_{\alpha/2}\times 0.5)\}^2}{\sum_{i=1}^{18}e_i}\text{和}\frac{\{\sum_{i=1}^{18}n_i+(Z_{\alpha/2}\times 0.5)\}^2}{\sum_{i=1}^{18}e_i} \quad (4)$$

3. 统计学检验[2~5]：直接法标化率的统计学检验步骤是，先求两个欲比较的标化率(如 ASR_1 和 ASR_2)的比值；再估计该比值的 95% 或 99% 置信区间。若该比值置信区间不包括 1，则可在 0.05 或 0.01 的水平上拒绝无效假设(两个标化率相等)，从而认为 ASR_1 和 ASR_2 之间存在差别。两标化率比的 $(100-\alpha)\%CI$ [2,4~6]用下式表达

$$(ASR_1/ASR_2)^{(1\pm Z_{\alpha/2}/X)}, \quad (5)$$

式中 $X=\frac{(ASR_1-ASR_2)}{\sqrt{\hat{Var}(ASR_1)-\hat{Var}(ASR_2)}}$，$\hat{Var}(ASR_1)$ 和 $\hat{Var}(ASR_2)$ 是 ASR_1 和 ASR_2 的方差，算法同上。间接法标化率的检验可直接通过其 95% 或 99% CI 是否包括 100 来判断 SIR 或 SMR 有无统计学意义。

4. 其他指标[2~5,7]：累积危险率(CR)和累积危险度($Risk_C$)也是常用分析指标。前者定义为从出生到 64 岁或 74 岁每岁的发病专率相加的和，后者则是指假定无其它死因存在的情况下，一个个体在某一年龄期间(如 0～64 岁)发生某种肿瘤的危险性大小。对应于 0～64 岁和 0～74 岁的累积危险率为

$$CR_{0\sim64}=\sum_{i=1}^{13}(5\times a_i)\text{及}\ CR_{0\sim74}=\sum_{i=1}^{15}(5\times a_i)。 \quad (6)$$

若零岁组分为两组，即 0～岁和 1～4 岁，则用下两式估计

$$CR_{0\sim64}=1\times a_1+4\times a_2+\sum_{i=3}^{13}(5\times a_i)\text{及}\ CR_{0\sim74}=1\times a_1+4\times a_2+\sum_{i=3}^{15}(5\times a_i)。 \quad (7)$$

基于泊松($Poisson$)假设下累积危险率方差的估计公式为

$$\hat{Var}(CR)=\sum_{i=1}^{18}\frac{(t_i^2\times a_i)}{p_i},\ t_i=5,\ i=1,2,\cdots,18。$$

有了累积危险率 CR，可很方便地估计累积危险度 $Risk_C$，见下式

$$Risk_C=100\times[1-\exp(CR/100)]。 \quad (8)$$

类似于 CR 估计，年龄组不同时对(8)式稍作变动。

二、实例说明：

表 1 是某地 1974～1978 年和 1984～1988 年两个时期男性食管癌的年龄别发病率、当地人口资料及世界标准人口数。年龄分组为 18 组，即 5 岁一组。由公式(1)可以很方便地求得两个时期的标化率，见表格底部。分别为 25.7/10 万、13.6/10 万。以第一时期的标化率 ASR_1 为例，我们可以估计它的 95% 及 99% 置信区间(CI)。可以看出，两种假设下的估计值相等(精确到 2 位小数)，详见表 2。

若对两个标化率进行统计学比较，可以通过求它们比值的 95% 或 99% 置信区间来判断两者差异是否有统计学意义。本例两个 ASR 比值为 1.89，差异有高度统计学意义。结合实际数值，可以认为前一时期该地男性食管癌的发病率高于后一时期。

表 4 是某地两个年份男性食管癌的年龄别发病率和人口资料。笔者利用该数据说明间接法标化率的估计及统计学检验。其中，选为标准的年龄别发病率是 1972 年的资料。我们的目的是计算 1985 年食管癌的标准化发病率比(SIR)，以及统计学检验。结果列于表 5。1985 年食管癌的实际发病例数是 504，而按着作为标准的 1972 年的发病率估计的期望发病例数为 1 312 例，则 SIR 为 38.43。表中同时列出了它的 95% 及 99% 置信区间。统计学检验则是根据 SIR 的置信区间来判断。表 5 中 95% 及 99% CI 均不包括 100，则结果是 1985 年观察到的食管癌发病率水平与当地 1972 年的食管癌发病水平之间的差异，有高度统计学意义。

表 1　某地男性食管癌年龄别发病率及人口资料(1974～1978 年和 1984～1988 年)

年龄组	1974～1978 年		1984～1988 年		世界标准人口
i～	a_i	p_i	a_i	p_i	w_i
0～	0. 0	533396	0. 0	1201788	12000
5～	0. 0	652926	0. 0	684380	10000
10～	0. 0	1295876	0. 0	638338	9000
15～	0. 0	1935496	0. 0	946342	9000
20～	0. 1	1422882	0. 0	1772897	8000
25～	0. 2	1338166	0. 2	2477796	8000
30～	0. 3	891558	0. 4	2048129	6000
35～	3. 3	817243	1. 2	1346615	6000
40～	6. 6	984647	2. 2	810005	6000
45～	17. 1	1090815	4. 1	868136	6000
50～	38. 7	945887	12. 6	1210914	5000
55～	65. 3	753809	30. 0	1049711	4000
60～	109. 5	556960	64. 0	835546	4000
65～	167. 1	407561	86. 6	612918	3000
70～	226. 2	251958	127. 3	408277	2000
75～	280. 6	121176	154. 4	226716	1000
80～	290. 7	43005	190. 5	95194	500
85+	261. 4	11476	212. 0	28958	500
标化率	ASR_1= 25. 7/ 100000		ASR_2= 13. 6/ 100000		

表 2　直接法标化率置信区间(CI) 的估计(/10 万)

标化率(ASR):	25. 7	
二项分布假设下:	95% CI:	24. 97～26. 43
	99% CI:	24. 74～26. 66
泊松分布假设下:	95% CI:	24. 97～26. 43
	99% CI:	24. 74～26. 66

表 3　两个标化率差异的统计学检验

标化率比:	ASR_1/ASR_2= 1. 89
标化率比的 95% CI:	二项分布区间: 1. 78～2. 00
	泊松分布区间: 1. 78～2. 00
标化率比的 99% CI:	二项分布区间: 1. 75～2. 04
	泊松分布区间: 1. 75～2. 04
统计学检验:	P< 0. 01

表 4　某地两个年份男性食管癌的年龄别发病率和人口资料

年龄组	1972 年	1985 年	
i～	a_i	n_i	p_i
0～	0. 0	0	253321
5～	0. 0	0	135354
10～	0. 0	0	119356
15～	0. 0	0	177667
20～	1. 1	0	333062
25～	0. 6	1	494749
30～	2. 2	1	421254
35～	5. 6	4	284591
40～	13. 4	4	163565
45～	27. 5	6	159177
50～	49. 6	28	240691
55～	91. 9	52	211643
60～	130. 7	116	168407
65～	229. 6	77	123180
70～	255. 5	81	82972
75～	280. 6	75	46070
80～	222. 7	47	19608
85+	203. 5	12	6066

表 5　间接法标化率估计及统计学检验

实际发病例数:	504
置信区间:	95% : 460. 96～548. 96
	99% : 447. 74～563. 58
期望发病例数:	1311. 617
标化率比(SIR):	38. 43
置信区间:	95% : 35. 14～41. 85
	99% : 34. 14～42. 97
统计学检验:	P< 0. 01

表 6 是利用表 2 中的 1974～1978 年的食管癌数据说明累积危险率和累积危险度的估计。表中分别列出 0～64 岁和 0～74 岁两种方式的估计值及其 95% 和 99% 置信区间。从表中数字可以看出, 累积危险率和累积危险度非常接近。所以通常情况下, 可以用累积危险率代替累积危险度。为了直观地表达累积危险率或累积危险度指标的含义, 可以用图表示。具体可分别计算出从出生至各个年龄组的累积危险率, 再作图即可。

表 6　累积危险率及累积危险度的估计

(用表 2 的 1974～1978 年数据)

累积危险率:	0～64 岁(%):	1. 2055
	置信区间:	95% : 1. 1559～1. 2551
		99% : 1. 1403～1. 2707
	0～74 岁(%):	3. 1720
	置信区间:	95% : 3. 0700～3. 2740
		99% : 3. 0378～3. 3062
累积危险度:	0～64 岁(%):	1. 1983
	0～74 岁(%):	3. 1222

三、小结:

肿瘤基础资料的积累和统计分析在肿瘤流行病学研究中占有非常重要的地位。曾有学者统计国外

癌症流行病学研究方面的杂志, 约 65% 的论文是描述流行病学研究结果, 或与其有关的研究材料。这些方面在我国是一个薄弱环节, 应该引起人们的重视。以肿瘤登记资料为例, 目前国内仅有上海、天津和江苏启东三地的发病资料纳入了国际癌症研究中心(IARC)五大洲肿瘤发病率汇编中。

本文介绍了有关分析癌症描述流行病学研究资料的一些统计方法。标准化率的计算及分析在国内很普遍, 但在估计其置信区间及进行统计学检验方面所做的工作较少。常常偏重于直观性比较。对于长期积累的慢性病发病死亡资料统计分析, 人们还经常借助于一些统计模型。例如年龄-时期-队列回归模型(*Age-Period-Cohort Model*), 简称 *APC* 模型[9,10]。它可以综合定量地评价年龄、时期、队列等因素在疾病发生死亡过程中的作用。但国内的具体应用很少。

(收稿: 1997-12-27)

河南省中学生 1996 年非结核分枝杆菌感染情况调查

高三友　李登旭　要玉霞　彭义利

为了解河南省中学生非结核分枝杆菌的感染情况, 我们对河南省中学生非结核分枝杆菌感染情况进行了调查, 结果如下。

一、材料与方法:

1. 调查对象为河南省各地 23 所中学生计 3204 人(男性 1685 人, 女性 1519 人), 无卡痕者 1933 人(男 998、女 935); 年龄 12～17 周岁。

2. 结核菌素由北京生物制品研究所提供, 批号为 9601。依照 WHO 推荐的 Mantoux 法分别于受试对象的左、右前臂掌侧皮内各注射 5TU PPD-B、2TU PPD-T, 72 小时后查验反应。

3. 标准: PPD-T 以反应硬结平均直径≥6mm 为阳性, PPD-B 以反应硬结平均直径≥4mm 为结素反应阳性; 无卡痕受试对象中 PPD-T 反应阴性但 PPD-B 反应阳性者, 即判其为非结核分枝杆菌感染者。

二、结果:

1. 结素反应及非结核分枝杆菌感染率: PPD-T 反应阳性率为 9.39%(301/3 204), PPD-B 反应阳性率为 7.49%(240/3 204); 非结核分枝杆菌感染率为 5.95%(115/1 933)。男性 PPD-B 阳性率为 8.31%(140/1 685), 女性 PPD-B 阳性率为 6.58%(100/1519); 其中男性非结核分枝杆菌感染率为 6.51%(65/998), 女性非结核分枝杆菌感染率为 5.34%(50/935), 差异无显著性(χ^2= 1.17, P> 0.05)。

2. 不同年龄感染率: 调查发现, 各年龄组非结核分枝杆菌感染率存在一定的差异, 12 岁组和 17 岁组感染率偏低, 但各年龄组之间差异无显著性(χ^2= 4.48, P> 0.05), 可能与样本例数偏少有关。

3. 不同地区非结核分枝杆菌感染情况: 本次调查发现, 我省非结核分枝杆菌感染率地区分布不均衡, 具有显著的地区差异。

三、讨论: 结素试验是调查非结核分枝杆菌感染的常用方法, 但由于非结核分枝杆菌与结核分枝杆菌间存在交叉免疫反应, 因此判断非结核分枝杆菌感染以无卡痕人群中 PPD-T 反应阴性, 但 PPD-B 反应阳性为标准。选择 PPD-T 阴性反应人群可以排除结核杆菌自然感染的影响。PPD-B 的阳性反应标准, 有报道认为 PPD-B≥4mm, 具有较好的灵敏度及特异度, 此次调查我们采用了该标准。

1990 年全国结核病流行病学抽样调查显示, 我国人群中非结核分枝杆菌平均感染率为 15.35%, 随年龄增长而上升, 其中 10～19 岁组感染率为 10.41%, 至 60 岁开始下降。本次调查发现, 河南省中学生人群非结核分枝杆菌感染率为 5.95%, 与之相比, 低于全国同年龄段平均感染水平。由于年龄分组较细, 各组间未发现显著性差异。本次调查还发现, 河南省中学生人群非结核分枝杆菌感染地区分布不平衡具有明显的地区差异, 原因有待进一步探讨。

(收稿: 1998-02-13)

作者单位: 河南省结核病防治研究所 郑州 450003