HAROLD A. KAHN

An Introduction to Epidemiologic Methods

New York Oxford
OXFORD UNIVERSITY PRESS
1983

流行病学方法门入

吕宝成 等译
韩向午 等校

华北煤炭医学院
1987

译者前言

近十几年来，飞速发展的流行病学，正在研究疾病、健康和制订卫生政策中发挥重要的作用。与此同时，也有更多的人认识到：流行病学方法在目前的医学科学研究中已成为不可缺少的部分。现在，我们翻译出版这本专门讨论现代基本流行病学方法的专著——《流行病学方法入门》，对于我国广大卫生防疫工作者、流行病学教学科研人员，以及对流行病学有兴趣、在实践中又需要学习流行病学方法的临床和管理人员将会有所裨益。

本书系原著者Harold A.Kahn教授根据他在流行病学教学中积累的丰富经验而编写的。此书内容新颖，论述精辟，深入浅出，颇具特色。书中所包括的流行病学基本方法和公式演算，均以实例详尽说明，使读者易于领会。鉴于我国目前有关流行病学方法的专著较少，故该书即可做为初学流行病学方法的入门阶梯，也是专业人员一本很好的参考书，对于广大师生无疑更是良师益友。

据本书牛津大学1983年英文版介绍，这本《流行病学方法入门》是《流行病学基础》(Foundation to Epidemiologic Methods, A. M. Lilienfeld 原著，1980年二版；1977年一版，俞焕文译，上海科学技术出版社，1981）的姐妹篇。由于《流行病学基础》系本书的基础课程，故读者学习本书时，应参考有关这方面的专业书籍。

应该说明：尽管流行病学方法各异，但流行病学方法所探讨的病因（或危险因素）最终要为疾病的防制服务。因

此，流行病学决不只限于方法学的研究，而是要从方法学着手，解决防制问题，像全球消灭天花那样，始终不渝地为控制和消灭疾病而努力。

在此书翻译出版之际，我们要向热情支持本书做为内发教材出版的河北省出版管理部门，鼓励本书出版的院系领导，以及为本书译校付出艰苦劳动的同志们表示衷心的感谢。

由于译校者的水平所限，本书谬误之处在所难免，尚希读者不吝指正。读者对本书的意见和要求，欢迎寄给本书的联系人：华北煤炭医学院流行病学教研室吕宝成副教授。（地址：河北省唐山市建设路）

译校者　韩向午
　　吕宝成

1986年12月于唐山

原著者序

本书是根据著者在美国约翰斯·霍普金斯大学卫生学和公共卫生学院讲授流行病学方法课的经验编写的。学习此课的学生，一般是医师、护士、营养师，以及其他对预防医学或公共卫生这一学科感兴趣的专业人员，他们希望学习如何进行流行病学研究或至少希望学习如何评价他人所做的类似研究。值得注意的是，这些学生普遍在定量分析方法上没有受过良好的训练或在此方面造诣不深。因此，本书的基调与我的教学经验一致，即尽可能将强调这一问题的实质详细阐述，并望通过数学运算来说明，这样做要比列出一个神秘、隐晦的纲领要好的多，不管这一纲领是如何严谨。关于严谨，这里的解释与牛津大学一个学生说的一致，在一次考试后问他，当时他是怎样证明二项式定理时，他欣然答到：他没有能力证明这一定理，只是使它似乎合乎道理（J.R.Newman，The World of Mathematics.New York:Simon and Schuster，1956）。在一些地方我已注意包括“人人均知”的琐碎内容。但很多地方未这样做。关于本书的内容，可能有的读者觉得太容易了，但这样会使其他人感到易于理解。

“流行病学原理”和至少一学期“生物统计学”是基础课程，这在霍普金斯大学是学习“流行病学方法”的前提条件，尽管写一本流行病学方法的书，不牵涉流行病学原理是不可能的，但我力求这样做。在更多的情况中，原理和方法是彼此纠缠，因而本书的读者会感到在基本原理方面给予了多余的讲解。对此我不表示谦意，因为我相信它不会有害。

尽管如此，因为本书的目的是一本方法学的书，所以很少包括或根本没有提到为什么采用某种方法，例如，对一个特定的调查要选用回顾性还是前瞻性调查，或者为什么要估算比数比或生存概率等。对解决这些以及有关的问题，必然要将基本课程重复，可我没有这样做。

由于在学习本课前得到的生物统计学训练很少，许多学生不得不吃力地把方法学课程中提出的概念与已有的统计学知识结合起来。这提示复习部分是客观需要的，这部分已列入第一部分。

凡是实用的数字举例说明学生能理解的都已包括。并强调用不同方法可以获得近似数字结果的相同的地方。

总之，本书所遵循的观点是：尽管所有涉及流行病学研究都需要熟知资料汇总和分析，但他们并不需要变为统计人员。

本书是我在耶路撒冷的Hebrew大学担任流行病学莱迪戴维斯客座教授时写的。George Comstock、Sidney Culer、Fred Ederer、Evic Peretz、Nathan Mantel和James Schlesslman诸位阅读了手稿并且提出了有益的意见、建议和批评，为此，我愿意表示感谢。对我的编辑Abraham Lilienfeld最慷慨的帮助值得特别感谢。

我要感谢已故的英国皇家医学会会员Ronald A·Fisher先生的著作权执行人、感谢皇家医学会会员Frank Yates和伦敦朗曼集团，允许从他们的《生物学、农业和医学研究用统计表》（1974年，第6版）一书中翻印随机数字。我还要感谢Jack Medalie和Uri Goldbourt，允许使用尚未出版的犹太人缺血性心脏病研究资料，并且感谢Paolo Pasquini、Lawrence Gould和Stanley Schor允许使用未出版的Merck Sh-

arp和Dohme研究实验室的资料。

最后，我愿意感谢我的妻子伦诺尔为我打印手稿，以及在很多方面的支持和鼓励。

H. A. K

美国加里弗尼亚州

Loma Linda

1982年8月

（吕宝成译　韩向午校）

目　录

1

精选基础统计复习

虽然，几乎本书的所有读者，均已完成了一门以上统计学入门课程，但我的经验告诉我：在词汇、符号、初等代数、均数和方差的公式，以及阐明X^2检验方面的简短复习，对于许多学生将是有益的。在本章的复习中，也包括了对于即将讨论的一些流行病学方法所必须了解的知识。这一章不是简要的全面的复习统计学入门课程，并且也不要打算照这样来用。

1.1 词 汇

总体或**整体**：总体常常需要调查者来很好的确定，由称为**参数**的一些简单常数经常能有效地描述总体。例如，我们可以1942～1945年服役的全部美军的收缩期血压值的均值和标准差为例。1942—1945年这个人群的血压值构成了一个总体，这个数值的均值和标准差是参数，假如已经知道这些参数，就能用于描述这个特殊的总体。在这个例子中，由于总体值几乎无疑是非正态分布，故均数和标准差将不能完善地描述该总体。

血压值的一个随机**样本**能从1942～1945年服役登记中得到，并可从样本中计算样本均数和标准差。这两个值称为**统计量**并且一般用在估计相应的**参数**。

报告的收缩期血压是**离散型**变量，例如140、126和138。

然而，这个变量，至少在概念上是**连续性变量**并且能假设在它的范围中的任一值，例如126、359021……

我们将常常提及一种特殊类型的变量，即**特征**。特征只有两种型式——患病或健康，男性或女性，工作或不工作等——并且只取值为1或0。

无效假设（Ho）：无效假设是指所比较的统计量例如样本均数是由于从同一总体随机抽样的结果，因此，这些统计量之间的任何差异，均是由于概率所致。无效假设检验易发生两个类型的误差。**Ⅰ类误差**是当无效假设在事实上是真实的情况下排除无效假设而产生的误差。**Ⅱ类误差**是当无效假设在事实上是错误的，未能排除无效假设而产生的误差。然而，假如无效假设是真的，则Ⅱ型误差是不存在的和没有意义的。同样，假如无效假设是错误的，则Ⅰ型误差是没有意义的。由于我们不知道从总体中抽样的实际情况（即知道无效假设是否真实），因而，我们无疑希望所设计的，能限制这两类错误的研究。因此，假如无效假设是真实的，我们需要限制Ⅰ型误差在α值，其一般规定α为0.01或0.05。假如无效假设是错误的，和某一备择假设（Ha）是真实的，我们需要限制Ⅱ型误差为β值，一般规定β值为0.10左右。当Ha真实时，检验结果排除Ho的把握度等于1—β。

1.2 符号和初等代数

字母：X_1、X_2、…X_j、…X_N表示总体N中第1、第2、…第j、以及第N个个体的变量值。

字母：x_1、x_2、…x_j、…、x_n表示样本大小为n的样本中第1、第2、…第j、以及第n个个体的变量值。

符号：$\sum_{i=1}^{N} X_i$

表示所有X_j值从X_1到X_N的总和。注意：在意义明确时我们将使用$\sum X_j$或甚至使用$\sum X$来代替完整的符号$\sum_{j=1}^{N}$，以表示所有X值的总和。

下列是用在统计分析上的其他常用符号。

$|X|$：X的绝对值，即$|7|$或$|-7|=7$

E（X）：X的期望值，相当于每个X值乘以它的概率，然后将所有乘积值累计相加，也即X的均值

$\cong$：大约等于

$>$：大于（a＞b，表示a比b大）

$\geqslant$：大于或等于

$<$：小于（a＜b，表示a比b小）

$\leqslant$：小于或等于

注意：许多学生的难点在于是否小于用＜表示还是用＞，应该注意起点指向的是小值：如$7<10$和$5>2$

假如一个二次方程式是标准型：

$a^2+bx+c=0$

于是：$X=\dfrac{-b\pm\sqrt{b^2-4ac}}{2a}$

假如$10^x=y$，于是x是y的对数，以10为底

假如$e^x=y$，于是x是y的对数和y是x的反对数，二者均以e为底

字母e表示自然对数的底，可以视为当x无限增大时，$\left(1+\dfrac{1}{x}\right)^x$这个式子的极限值。举例说明：假如$x=2$，则$\left(1+\dfrac{1}{2}\right)^2=(1.50)^2=2.250$；假如$x=100$，则

$\left(1+\frac{1}{100}\right)^{100} = (1.01)^{100} = 2.705$，至小数三位，e=2.718。也就是说既使x = 1000000或更大时，$\left(1+\frac{1}{x}\right)^{x}$ 将小于2.719。以后，我们提到的都是用e为底的对数。

同样，l_{ny} = y的对数，用e为底

$$l_n(xy) = l_{nx} + l_{ny}$$

$$l_n(x/y) = l_{nx} - l_{ny}$$

$$e^0 = e^{x-x} = e^x e^{-x} = 1$$

$$e^{-x} = 1/e^x$$

$$x^{1/2} = \sqrt{x}$$

1.3 均值和方差计算

以下给予的方程及相应的含义在流行病学分析中经常发生，因此，读者应熟悉它们。

$$\sum_{j=1}^{N} x_j/N = \mu \qquad (1-1)$$

在这里，μ表示总体的均值

$$\sum_{j=1}^{N} (X_j-\mu)^2/N = \mathrm{Var}(X) \qquad (1-2)$$

在这里，Var（X）表示总体方差

$$\sum_{j=1}^{n} X_j/n = \bar{x} \qquad (1-3)$$

这里 $\bar{x}$ 表示样本的均值。假如样本是随机抽取的，$\bar{x}$ 的期望值则是全部可能被抽取的样本的均值等于μ。符号为：

$$E(\bar{x}) = \mu \qquad (1\ 4)$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1) = \hat{var} \qquad (1-5)$$

这里$\hat{var}$（X）是依据样本资料计算的var（X）的估计值。注意：当我们讨论两个样本的统计量和参数时或为了明确，必要时我们将使用发音符号（^）标明样本统计量，无此标志时标明总体参数。因为参数没有方差，我们能在表示变量统计量或标准差统计量上面省略发音符号。对这些习惯有两个例外。首先，对二项分布的参数和统计量，不是用P和$\hat{P}$来所研究特征的比值，而是分别用P和p表示相应的总体率（参数）和样本率（统计量），其次，在统计表中每一格内的例数，用大写字母——A、B、C…表示人群变量，而样本数将用小写字母——a、b、c…表示。

$$var(\bar{x}) \cong \frac{var(X)}{n} \quad (除非 \frac{n}{N} > 0.10) \qquad (1-6)$$

$$SE(\bar{x}) \cong \left[\frac{var(X)}{n}\right]^{\frac{1}{2}} \quad (除非 \frac{n}{N} > 0.10) \qquad (1-7)$$

SE（$\bar{x}$）是样本均值的标准误。它平常由$\left[\frac{\hat{var}(X)^{\frac{1}{2}}}{n}\right]$来估计，这时，我们写成$\hat{SE}$（$\bar{x}$）。样本均值或来自一个样本为n的其他样本统计量的变异性，与被抽样总体抽取的所有可能样本为n的样本统计量的假设分布的离散度有关。样本统计量的标准误，事实上即是这个假设分布的标准差。

1.4 变量函数的方差公式

下面关于简单函数的方差公式是很有用的。

假设X和Y是独立的

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) \quad (1-8)$$

假设X和Y是独立的

$$\text{var}(X-Y) = \text{var}(X) + \text{var}(Y) \quad (1-9)$$

当K是一个与抽样误差无关的常数时

$$\text{var}(KX) = K^2\text{var}(X) \quad (1-10)$$

1.5 分组资料的均数和方差公式

经常把资料分成多组，例如血压分为120～129，130～139……以下定义使用于：

X_i = 已分配到i组段的每一成员的数值，一般取i组段的组中值

f_i = i组段的频数

m = 组数

于是，$\sum_{i=1}^{m} f_i x_i / \sum_{i=1}^{m} f_i \cong \mu$ (1—11)

此近似值的好与差取决于各组实际数值的分布。

同理，$\frac{\sum_{i=1}^{m} f_i x_i^2}{\sum_{i=1}^{m} f_i} - \left(\frac{\sum_{i=1}^{m} f_i x_i}{\sum_{i=1}^{m} f_i} \right)^2 \cong \text{var}(X)$ (1—12)

1.6 特征资料的均数和方差公式

在全部X_i值为1或为0（一般表示疾病的有无或一些其他特征的有无）的总体中，假设N个值的比为1是P，那么1和0的频数分别是NP和N（1—P），从1—11和1—12公式可以得出：

$$\mu = \frac{NP(1) + N(1-P)(0)}{NP + N(1-P)} = P \quad (1-13)$$

$$\mathrm{var}(X)=\frac{(NP)(1^2)+N(1-P)(0^2)}{NP+N(1-P)}$$

$$-\left(\frac{NP(1)+N(1-P)(0)}{NP+N(1-P)}\right)^2$$

$$\mathrm{var}(X)=\frac{NP}{N}-\left(\frac{NP}{N}\right)^2=P-P^2=P(1-P)\quad(1-14)$$

此处，由于两个X_i值完全代表了总体中相应的两个数值，分组资料公式可给予完全正确的结果。因此，P是均值并且P（1－P）是1和0总体的方差。假如我们取n中的一个样本，用公式1—6得出样本均值的方差为：

$$\mathrm{var}(\bar{x})\cong\frac{\mathrm{var}(X)}{n}=\frac{P(1-P)}{n}\qquad(1-15)$$

对于特征资料，一般以P代替$\bar{x}$做为样本均数，因此，上述结果可写为：

$$\mathrm{var}(p)\cong\frac{P(1-P)}{n}\qquad(1-16)$$

我们很少知道P，一般用p来估计它。这导至常用的公式：

$$\hat{\mathrm{var}}(p)\cong\frac{p(1-p)}{n}\qquad(1-17)$$

有时，我们的兴趣并不集中在总体特征的样本均值而是在于样本中为1的数目。在此情况下，要考虑的统计量不是p而是np，np是在样本中有某一特征的个体总数。利用计算$\hat{\mathrm{var}}(p)$的公式1—17和计算与一常数相乘变量的方差公式1—10，我们得到：

$$\hat{\mathrm{var}}(np)\cong\frac{n^2p(1-p)}{n}=np(1-p)\quad(1-18)$$

一般将（1－P）和（1－p）分别写成Q和q。

1.7 可信限

当计算了样本均数的标准误，并且样本大或者抽样总体近似正态分布时，总体均数的可信限（CL)可以下式运算：

$$95\%\,CL = \bar{x} \pm 1.96\hat{SE}(\bar{x}) \qquad (1—19)$$

$$99\%\,CL = \bar{x} \pm 2.58\hat{SE}(\bar{x}) \qquad (1—20)$$

注意：已给出的公式1—19和1—20并不正确，甚至对于正态分布的大样本亦是如此。为了达到完全正确，我们应该以$SE(\bar{x})$取代$\hat{SE}(\bar{x})$。然而，对大样本来说这种差异是很小的。对正态整体的小样本来说，1.96和2.58这两个值应该用t检验的t值表[1]变为相应的值。对于非正态整体的小样本，可以请教统计学工作者。

使用上面给予的可信限的理由是（1）来自正态分布总体的所有可能样本的样本均数分布是正态的和（2）不论总体样本变量是否正态分布基于大样本所有可能的样本的均值分布似乎都是正态的。我们估计的样本均值的标准误是所有可能样本均值假设为正态分布标准差的一个估计值。在正态分布中，仅有5％的值大于1.96个标准差离开期望值。由于$E(\bar{x}) = \mu$，我们可以说，只有5％的可能样本均值大于± 1.96个$SE(\bar{x})$而远离μ。我们不知道$SE(\bar{x})$，但可使用从样本资料得到的$\hat{SE}(\bar{x})$代替之。对于大样本，这是完全可以令人满意的。

假设样本量足够大或者总体呈正态分布，我们可以从$\bar{x}$加或减$1.96\hat{SE}(\bar{x})$并且包括μ。只有假如我们的$\bar{x}$值是5％的离开μ值大于$\pm 1.96\hat{SE}(\bar{x})$的一个值时，这个范围才不包括μ。因此，$\bar{x} \pm 1.96\hat{SE}(\bar{x})$是μ的95％可信区

间。同理，99%可信区间使用$\bar{x} \pm 2.58\widehat{SE}(\bar{x})$。

1.8 卡方公式

大多数的流行病学学生熟悉的x^2（卡方）公式是：（观察频数-期望频数）^{2}被期望频数相除的商，从所有相互关联的全部频数所得的商数相加累计：

$$X^2 = \sum \frac{(\text{f观察值} - \text{f期望值})^2}{\text{f期望值}}$$

相互关系的各个频数不是可以自由变化的，如果由于一些观察频数大于期望频数，另外一些观察频数一定小于期望频数。然而，一些学生不知道上列公式只是x^2基本公式中一个特殊情况，这在单变量中，卡方可表达为其期望值的离差：

$$X_1^2 = \frac{[x - E(x)]^2}{\text{var}(x)} \qquad (1-21)$$

在这里，角码1指的是自由度为1。在K个变量的情况下，要加上一个总计符号，即：

$$X^2(k-1) = \sum_{i=1}^{K} \frac{(x_i - \bar{x})^2}{\text{var}(x)} \qquad (1-22)$$

（吕宝成译　袁聚祥校）