流行病学方法导论

〔美〕H·A·卡恩著
文万青 谭红专译
文师吾 校
吴彭年 审

译者前

流行病学的发展异常迅猛，流行病学的原理和方法正向其它医学学科渗透，给医学科学研究不断提供新的有效手段。除了专业流行病学工作者必须掌握流行病学原理和方法外，从事其它医学学科研究的人员也应对流行病学原理和方法有所了解。翻译本书的目的，便是为广大基层流行病学工作者以及希望对流行病学方法有所了解的其它专业人员介绍一本优秀的入门读物。

本书是美国霍普金斯大学的流行病学专家 Harold A. Kahn 先生编著的，着重介绍了流行病学研究中的一些常用方法，尤其对资料分析的方法作了重点叙述。内容包括随机抽样方法；比值比、相对危险度和特异危险度及其可信限的计算；资料校正和混杂控制方法；生存概率和寿命表的计算及其应用等。除了这些一般常用的方法外，还对一些多变量方法（如多元回归，判别函数，多元 Logistic 函数等）作了深入浅出的详细论述。本书虽然主要讲方法，但正如作者在序言中写的那样，讲方法离不开讲原理，因此本书对流行病学的一些基本原理也作了简单的说明，对容易混淆的几个基本概念（如混杂作用，偏倚，效应修饰作用等）作了精辟的论述。作者善于用浅显易懂的数字实例阐明较为抽象的数学原理，对数学不熟悉的读者也完全能够理解。本书适于广大基层流行病学工作者，临床医师，大专院校学生和其它有兴趣的读者阅读。本书内容新颖，重点突出，颇具特色，不失为一本良好的学习流行病学方法的入门书。

在本书出版过程中得到湖南省卫生防疫站李庆俊站长和欧阳玉梅科长的大力支持，谨致谢忱。

由于译者水平所限，时间仓促，匆匆脱稿，译文中的不妥之处甚至错误在所难免，衷心希望得到专家的指正和读者的帮助。

译　者

1986年6月于湖南医学院

序　言

此书是根据本人在霍普金斯大学公共卫生学院讲授流行病学方法这一课程的经验编写成的。参加本课程学习的有医生、护士、营养师以及一些被预防医学或公共卫生学所吸引，想要了解怎样开展流行病学调查或者至少知道怎样评价别人的调查结果的其它专业人员。值得注意的是，这些人在数量方法上未曾受过良好的训练。所以本书的格调与我那次教学的经验保持一致，着重介绍基本概念，并通过容易理解而且严密精确的数字运算予以阐明，避免意义含糊的提纲式解说。说到严密，这里的解释就好像牛津大学某学生所说的那样。一次考试后，当问及他是如何证明二项式定理的时，他说：当定理未能得到证明时，结果便是似是而非的①。本书包括了一些“众所周知”的内容。但大部分内容并非如此。我想把本书的难易程度控制在部分读者认为太浅显而大多数读者亦能理解的水平上。

在霍普金斯大学学习流行病学方法这门课程之前应先学习基础课程，流行病学原理以及至少一个学期的生物统计学。不参考流行病学原理，要想写一本关于流行病学方法的教材几乎是不可能的，但我还是试着写了。在很多情况下，原理与方法互相渗透，以至读者会感到本书在基本原理方面写得太多。但我并不为此而不安，我相信这是无妨的。然而，本书的目的是讲方法，故几乎不解释为什么某次调查要采用回顾性的或者是前瞻性的研究方法，也不解释为什么要计算比值比和生存概率等等。为解答此类问题，需要完整复习基础课程的内容，本书没有如此赘述。

很多学生在作为学习本门课程的先决条件之一的生物统计学方面训练有限，因此他们很希望把方法学课程中学习的概念与他们已有的统计学知识结合起来。这促使我写了第一章，作为统计学的复习。

本书尽可能用学生易于理解的数字实例来说明，并强调用不同的方法进行计算可以得到相似的数值结果的情况。

总之，本书的基本观点是：所有从事流行病学研究的工作者都必须理解资料整理和分析的基本原则，但不必成为统计学家。本书是我在耶路撒冷的希伯来大学担任 Lady Davis 的流行病学客座教授期间写成的。George Comstock，Sidney Cutler，Fred Ederer，Eric Peretz，Nathan Mantel 和 James Schlesslman 阅读了手稿并提出了一些很好的意见、建议和批评。在此，我表示衷心感谢。特别对编辑 Abraham Lilienfeld 给予的热情帮助谨致谢意。

感谢已故的皇家学会会员 Ronald A. Fisher 先生的遗嘱执行人，感谢皇家学会会员 Frank Yates 并感谢朗曼公司允许从他们的著作《生物学、农业和医学研究用统计表》（1974年第六版）中翻印随机数字。还要感谢 Jack Medalie 和 Uri Goldbourt 允许从他们的犹太人缺血性心脏病研究中引用未发表的资料。感谢 Paolo Pasquini，Lawrence Gould 和 Stanley Schor 允许从他们的 Merck Sharp 和 Dohme 研究室中引用未发表的资料。

最后，我要感谢我的妻子 Lenora 为我打印手稿以及在各个方面给予的支持和鼓励。

Loma Linda, California

1982年8月

H.A.K.

（文万青译　文师吾校）

① J.R. Newman，数学世界，New York：Simon Schuster，1956

目　　录

第一章　几个统计学基本概念的复习……（1）
第一节　名词解释……（1）
第二节　数学符号和初等代数……（2）
第三节　均数和方差计算公式……（4）
第四节　变量函数的方差公式……（5）
第五节　成组资料的均数和方差公式……（6）
第六节　定性资料的均数和方差公式……（6）
第七节　可信限……（7）
第八节　卡方公式……（8）
第二章　随机抽样……（10）
第一节　单纯随机抽样……（10）
第二节　分层随机抽样……（16）
第三节　系统抽样……（22）
第四节　整群抽样……（25）
第五节　样本大小……（27）
第三章　相对危险度和比值比……（42）
第一节　相对危险度……（42）
第二节　比值比……（47）
第三节　回顾性调查资料的相对危险度……（52）
等四节　比值比的可信限……（52）
第五节　相对危险度的可信限……（55）
第六节　在样本大小计算中比值比的应用……（60）
第四章　特异危险度……（64）

第五章　非多变量模型的资料校正……(73)
第一节　混杂……(73)
第二节　直接校正……(74)
第三节　间接校正……(82)
第四节　2×2表中的混杂变量……(90)
第五节　校正比值比的可信限……(100)
第六节　多重配比对照……(109)
第六章　应用多元线性回归和多元 Logistic 函数的校正……(116)
第一节　一元线性回归的复习……(116)
第二节　多元线性回归系数……(116)
第三节　多元回归方法的基本假设……(120)
第四节　分类变量的多元回归……(123)
第五节　判别函数……(128)
第六节　多元 Logistic 函数……(129)
第七节　利用多变量函数分层……(139)
第七章　纵向研究：寿命表……(141)
第一节　假设与计算……(141)
第二节　病因寿命表……(153)
第三节　抽样误差和显著性检验……(157)
第四节　人口统计寿命表……(169)
第八章　纵向研究：人年……(172)
第一节　假设与计算……(172)
第二节　抽样误差和显著性检验……(184)
参考文献……(188)
索　引……(194)

第一章　几个统计学基本概念的复习

虽然本书的大多数读者可能都学过一学期或者几学期的统计学入门课程，但经验告诉我，简要复习一些术语，数学符号，初等代数，均数和方差的计算公式以及卡方的定义等，对许多学生定会有益的。本章复习的内容对即将讨论的许多流行病学方法的理解是十分必要的。本章不打算对统计学入门课程作完整复习。

第一节　名词解释

总体：研究者总是希望对总体下定义。一些称为参数的简单常数能够有效地描述总体。我们以 1942 年到 1945 年期间美国军人收缩期血压值的均数和方差为例。此期间这群人的血压值构成一个总体。这些血压值的均数和标准差即为参数，如果这些参数已知，就可以用来描述该总体。但在这个例子中，由于总体值不呈正态分布，故均数和标准差不能很好地描述该总体。

从 1942 年到 1945 年的服役记录中，可以抽得血压值的一个随机样本，并计算出样本的均数和标准差。这两个数值称为统计量，一般用于估计相应的参数。

收缩期血压是用离散值的形式报告的变量，例如：140，126，138．但至少在定义上，这个变量属于连续变量。它们可以假定是在一定范围内的任意值，如 126.359021…。

我们将要经常提到的一类变量是定性变量。其性质以两分的形式存在——如疾病或健康，男性或女性，服役或未服役

等——取值为1和0.

无效假设(H_0)：无效假设是指相互比较的统计量（如样本均数）是从同一总体随机抽样的结果。所以，这些统计量之间的任何差别都是由于机率所致。无效假设检验包括两类误差。第一类误差即当无效假设为真时，拒绝无效假设而产生的误差；第二类误差即当无效假设为假时，接受无效假设而产生的误差。但是，如果无效假设为真，则第二类误差就不存在或无意义。同理，如果无效假设为假时，则第一类误差没有意义。因为我们不知道从中抽样的总体的实际情况（即不知道无效假设是否为真），故需要通过研究设计来限制两类误差。所以，如果无效假设为真，我们需要限制第一类误差在α，通常规定α为0.01或0.05. 如果无效假设为假，则某备择假设（H_a）为真，我们需要限制第二类误差在β，通常规定β为0.10左右。当H_a为真时，检验结果拒绝H_0的把握度等于$1-\beta$.

第二节　数学符号和初等代数

字母：

$$X_1,\ X_2,\cdots,X_j,\cdots,X_N$$

分别代表总体N中第1，第2，…，第j，…，和第N个个体的变量值。

字母：

$$x_1,\ x_2,\cdots,x_j,\cdots,x_n$$

分别代表样本n中第1，第2，…，第j，…，和第n个个体的变量值。

数学符号：

$$\sum_{j=1}^{N} X_j$$

代表从X_1到X_N所有X_i值的总和。注意在意义清楚时，我们将用ΣX_i甚至ΣX代替复杂的符号$\sum_{i=1}^{N} X_i$表示所有X值的总和。

以下是在统计分析中经常用到的其它一些数学符号：

$|x|$　　x的绝对值，如$|7|$或$|-7|=7$

$E(X)$　X的期望值，等于每个X值乘以它的概率，然后将所有乘积值累加，也就是X的均数。

$\cong$　　约等于

$>$　　大于($a>b$，表示a大于b)

$\geqslant$　　大于或等于

$<$　　小于（$a<b$，表示a小于b)

$\leqslant$　　小于或等于

注意：难以记住小于号是“$<$“还是”$>$”的学生应注意符号尖角指向小数。如$7<10$和$5>2$

如果二次方程的标准形式是：

$$ax^2+bx+c=0$$

那么

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

如果$10^x=y$，那么x是以10为底y的对数。

如果$e^x=y$，那么x是y的对数，y是x的反对数，两者均以e为底。

字母e代表自然对数的底，当x无限增大时，可用式$\left(1+\frac{1}{x}\right)^x$的极限值来定义$e$．例如：当$x=2$，则$\left(1+\frac{1}{2}\right)^2=(1.50)^2=$

$=2.250$；当 $x=100$，则 $\left(1+\frac{1}{100}\right)^{100}=(1.01)^2=2.705$．取三位小数时，$e=2.718$．即使 $x=1000000$ 甚至更大时，$\left(1+\frac{1}{x}\right)^x$ 仍将小于 2.719．以后我们提到的都是以 e 为底的对数。

也有：

$\ln y=$以 e 为底，y 的对数

$\ln(xy)=\ln x+\ln y$

$\ln(x/y)=\ln x-\ln y$

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

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

$x^{\frac{1}{b}}=\sqrt[b]{x}$

第三节　均数和方差计算公式

下面给出的等式及其相应的含义在流行病学分析中常会碰到，所以读者应当熟悉它们。

$$\sum_{j=1}^{N} X_j/N=\mu \tag{1-1}$$

这里 μ 代表总体均数。

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

这里 $\mathrm{var}(X)$ 代表总体方差。

$$\sum_{j=1}^{n} x_j/n=\bar{x} \tag{1-3}$$

这里 $\bar{x}$ 代表样本均数。如果样本是随机抽取的，那么 $\bar{x}$ 的期望值，即所有可能样本均数的平均值等于 μ，表示为

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

$$\sum_{j=1}^{n}(x_j-\bar{x})^2/(n-1)=\widehat{\mathrm{var}}(X) \tag{1-5}$$

这里 $\hat{\text{var}}(X)$是根据样本资料计算的 var(X)的估计值。注意：当我们同时讨论样本统计量和总体参数时，为了明确起见，必要时用符号(^)表示样本统计量，无此符号则表示总体参数。这是由于参数没有变异，我们可以省去在表达统计量的方差或标准误时置于统计量上方的那个符号。对于这些习惯，有两个例外。其一，对二项分布的参数和统计量，不是用P和$\hat{P}$表示所研究现象的率，而是用P和p，分别表示相应的总体率（参数）和样本率（统计量）。其二，在表格中，每一行的例数，其总体数用大写字母——A，B，C，……表示，样本数用小写字母——a，b，c，……表示。

$$\text{var}(\bar{x})\cong\frac{\text{var}(X)}{n}\left(\text{除非}\frac{n}{N}>0.10\right) \tag{1-6}$$

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

$SE(\bar{x})$是样本均数的标准误。常用$\left[\frac{\hat{\text{var}}(X)}{n}\right]^{\frac{1}{2}}$估计之。这时要写成$\hat{SE}(\bar{x})$。样本均数或从一个样本$n$算出的其它统计量的变异性，与从被抽样总体抽取的所有可能样本n的样本统计量假想分布的离散度有关，样本统计量的标准误事实上就是这一假想分布的标准差。

第四节　变量函数的方差公式

下列对于简单函数的方差的计算公式是很有用的。

$$\text{var}(X+Y)=\text{var}(X)+\text{var}(Y)，\text{如果}X\text{与}Y\text{独立} \tag{1-8}$$

$$\text{var}(X-Y)=\text{var}(X)+\text{var}(Y)，\text{如果}X\text{与}Y\text{独立} \tag{1-9}$$

当K是一个与抽样误差无关的常数时，

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

第五节　成组资料的均数和方差公式

资料常被分成多组，例如血压可能按120～129，130～139，…等组段分组。本书要用到下列定义：

X_i＝分配给 i 组每一成员的数值，常取该组段的组中值。

f_i＝ i 组的频数

m＝组数

于是：

$$\sum_{i=1}^{m} f_i X_i \Big/ \sum_{i=1}^{m} f_i \cong \mu \qquad (1\text{-}11)$$

该近似值是好是坏，取决于各组实际数值的分布。

同理：

$$\frac{\sum_{i=1}^{m} f_i X_i^2}{\sum_{i=1}^{m} f_i} - \left(\frac{\sum_{i=1}^{m} f_i X_i}{\sum_{i=1}^{m} f_i} \right)^2 \cong \mathrm{var}(X) \qquad (1\text{-}12)$$

第六节　定性资料的均数和方差公式

在所有 X_i 值或为1或为0（通常表示疾病或某些其它特征的有或无）的总体中，令 N 中取值为1的比率为 P，那么取值为1和取值为0的频数分别是 NP 和 $N(1-P)$. 从式1-11和1-12得出：

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

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

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

在这里，因为两个 X_i 值能够完全代表总体中相应的两个数值，所以，成组资料的计算公式可以给出完全正确的结果。P 是均数，$P(1-P)$是取值为 1 和 0 的总体的方差。如果我们抽取一个例数为 n 的样本，从式 1-6 得到样本均数的方差是：

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

对于定性资料，常用 p 代替 $\bar{x}$ 作为样本均数，这样可将上式写成：

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

我们很少知道P，常用 p 估计之，这样便导出一个常用的公式：

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

有时，研究者的主要兴趣并不在定性资料的样本均数，而是样本中数值为 1 的个数，如果这样，所要考虑的样本统计量不是 p 而是 np，np 是样本中具有某特征的个体数。利用计算 $\hat{\mathrm{var}}(p)$的式 1-17 和计算与一常数相乘后变量的方差的式1-10，我们得出：

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

通常把$(1-P)$和$(1-p)$分别写成Q和q.

第七节　可信限

计算出样本均数的标准误后，当样本够大或者抽样总体近似正态分布时，总体均数的可信限(CL)可按下式计算：

$$95\%CL=\bar{x}\pm1.96\hat{SE}(\bar{x}) \qquad (1\text{-}19)$$

$$99\%CL=\bar{x}\pm2.58\hat{SE}(\bar{x}) \qquad (1\text{-}20)$$

注意：式1-19和1-20并不精确，甚至对从正态分布抽取的大样本亦如此。为了使两等式精确，应该用$SE(\bar{x})$代替$\hat{SE}(\bar{x})$，但是，对于大样本，差异是微小的。对于小样本，1.96和2.58应该变成t值表中相应的值[1]. 对于来自非正态分布的小样本，最好征求统计学家的意见。

采用上述可信限的理由是：1.来源于正态分布总体的所有可能样本均数的分布是正态分布。2.不论总体中变量是否呈正态分布，所有可能的大样本均数的分布是近似正态的。样本均数的估计标准误便是所有可能样本均数假想分布的标准差的一个估计值。分布中仅5%的值大于±1.96标准差而远离期望值。因为$E(\bar{x})=\mu$，我们可以说，仅5%的可能样本均数值大于$\pm1.96SE(\bar{x})$而远离μ。我们不知道$SE(\bar{x})$，但可用来自样本资料的$\hat{SE}(\bar{x})$代替。对于大样本，这是完全令人满意的。

假设样本足够大或者总体呈正态，我们认为从$\bar{x}$加上和减去$1.96\hat{SE}(\bar{x})$后的范围包括了总体均数μ. 仅当$\bar{x}$的值是5%的离μ值$\pm1.96\hat{SE}(\bar{x})$之外的数值之一时，这个范围才不包括μ. 因而，$\bar{x}\pm1.96\hat{SE}(\bar{x})$是$\mu$的95%可信区间。同理，99%的可信区间是$\bar{x}\pm2.58\hat{SE}(\bar{x})$.

第八节　卡方公式

大多数学习流行病学的学生都熟悉χ^2（卡方）公式。卡方是（观察频数—期望频数）^{2}除以期望频数所得的商。将从所有相互关联的频数所得的这些商累加得：

$$\chi^2=\sum\frac{(f_{观察}-f_{期望})^2}{f_{期望}}$$

相互关联的各个频数并不是都可以自由变化的。因为如果某些

观察频数大于期望频数，则另外一些观察频数一定小于期望频数。有些学生不知道上式只是卡方基本公式中的一个特例，对于单变量来说，卡方可表达为与其期望值的离差：

$$\chi_1^2=\frac{[x-E(x)]^2}{\mathrm{var}(x)} \tag{1-21}$$

这里，脚注 1 指的是自由度为 1，有K个变量时，须添加一个总和符号。

$$\chi_{(K-1)}^2=\sum_{i=1}^{K}\frac{(x_i-\bar{x})^2}{\mathrm{var}(x)} \tag{1-22}$$

（文万青译，文师吾校）

第二章　随机抽样

本书只探讨在抽样过程中包括随机因素的样本。由自愿者组成或由抽样者根据某个体是否合适而决定的非随机样本用来对总体参数作估计，某效果可能好也可能差，但没有客观的方法来判断。象“我们有95%的信心认为高密度脂蛋白中含胆固醇的平均百分数在17到19之间。”这种说法仅适于随机样本。

第一节　单纯随机抽样

单纯随机抽样理论上很简单[2]，但实际上并非如此。理论上，人们有一张包含N个元素（通常指个体）的清单，给从1到N的各个元素按顺序地赋予一个数字，然后利用随机数字表，从N中抽出n个组成样本。不熟悉随机数字表[3，4，5]的读者，请参阅表2-1[6]。使用该表的原理很简单，例如：假设抽样总体共有8059人，从中抽取300人作为样本。注意无论怎样排列，随机数字表都可提供一个连续的四位数。一种方法是采用四位数一列。表2-1中，从左上角开始，将得到0347，9774，1676，等数字。到达该表的底端时，另取一列四位数：4373，2467，等等。使用该表的另一方法是一位数一列，从每列中选出四位数。用此方法，表2-1的前三个数字是0911，5186，3512。

为了避免每次抽出相同样本，要有改变起始点的方法，这毋需赘述。例如，打开一本书，随意翻到一页可能是270和271页。用页数的头二个数字作为起始点，故从随机数字表的第27个四位数开始。表2-1中，用四位数一列的方法，第27