第一章　军事医学统计学概述

第一节　军事医学统计学

一、军事医学统计学的重要性和任务

(一)军事医学统计学的意义和作用

医学统计学(medical statistics)是运用概率论与数理统计的原理及方法，结合医学实际，研究数字资料的收集、整理、分析与推断的一门学科。医学研究的主要对象是人和有关的影响因素。人本身是一种生物，具有生物的一般特性，即生物现象有较大的变异;但人又不同于一般生物,影响人体健康的因素除了自然因素外,还有心理、社会等因素。在人体研究中，生物学研究常用的损害性、破坏性的方法不能使用。由于上述原因，使医学研究变得更复杂，医学数字资料的收集、整理与分析也更困难。随着医学研究的发展,20世纪20年代后逐渐形成了医学统计学。60年代以来，医学研究的迅速发展又促进了医学统计学的迅速发展。电子计算机的发展、应用与普及为医学统计学提供了强有力的工具，使医学统计学进一步向深度和广度发展，成为医学研究中必不可少的手段。

军事医学统计学(military medical statistics)是医学统计学在军队卫生工作中的应用，其原理、方法和基本内容与医学统计学相同，但突出了军队是武装集团的特点。军事医学统计的主要作用是及时准确地反映部队卫生工作情况的各种统计资料和统计分析结果，为各级军政首长和卫勤领导制订卫生工作规划、考核指导工作及科学管理和决策提供统计学依据。

(二)我军医学统计工作的基本内容

军事医学统计学在介绍统计学的基本知识（如统计工作的步骤、常用统计指标的计算、统计表与统计图的绘制等)的基础上,主要包括以下内容：

(1)部队平时卫生统计。这是军队卫生统计中最基本的内容，无论部队、机关、医院、院校，都必须按统一规定的内容和时间完成登记、统计并逐级上报。其中主要有：①反映军队人员健康水平的健康统计；②反映军队人员疾病发生及其危害情况的疾病统计；③反映军队开展预防工作和健康教育情况的卫生防疫统计；④反映军队卫生资源利用和卫生力量状况的卫生业务统计等。军队卫生统计工作已形成一套完整的制度，有严格的规定，并都已纳入军队卫生统计学的内容。

(2)医院卫生统计。其研究对象是医院管理工作中的各种随机现象，透过其偶然性探讨其规律性。医院卫生统计主要内容是：①医院卫生资源统计，目的是了解卫生资源（含人员、设备、经费与信息等）的投入、分配及利用是否合理与有效;②医疗工作数量、质量统计，目的是了解工作量以及工作量与人员配备是否相适应，能否及时做出诊断与

需要全本请访问 www.ertongbook.com

诊断是否正确,治疗是否有效、及时和彻底,抢救危重病人是否成功,护理技术操作是否正规、熟练,服务是否优质等;③医院综合效益统计，分社会效益与经济效益，前者是了解医疗、科研、人才培养以及为部队、为地方群众服务工作的情况,后者是了解医疗费用的收支和对上级拨款的使用情况等。近几年来，医院卫生统计发展很快，在医院信息管理系统和医院统计指标体系的研究与实践中取得了积极有效的成果，统计资料的搜集、整理、分析与信息反馈正逐步实现自动化，从而促进了医院管理的科学化。因此，医院卫生统计的内容已越来越充实。

(3)战时卫生统计。由于战时情况紧急，部队行动快，因而要求统计工作必须内容简练，填写方便，报告及时。其主要内容有:①部队减员统计，以反映指战员因伤、因病减员情况;②战伤救治工作统计，以反映各级救治机构的床位展开、救治效果及伤病员转运等情况;③伤员的伤情统计，以反映负伤的类别、部位及其原因等情况。海、空军及特种兵部队可根据自身特点适当修订或补充部分具体内容，但须报总后卫生部审查备案。此外,平时发生的各种灾害,如地震、洪水、火灾、空难等,其抢救中的卫生统计工作，也可参照战时卫生统计执行。

(三)军医掌握统计知识的重要性

医务人员既是卫生统计工作的执行者，又是受益者。军医掌握统计知识的重要性主要体现在以下几个方面：

(1)医生的正确判断，如患者的病情诊断、治疗方法的选择与预后分析等，在很大程度上取决于统计信息的积累。

(2)临床和卫生防疫中的许多观察数据具有不确定性，如患者的各项检查、化验数据，某一单位不同时期的发病人数等。没有一定的统计学知识就不能正确地解释和评价这些数据。

(3)大量的外界信息,如各种传播媒介、科技文献和情报资料,都伴随有直接或间接的统计信息，没有一定的统计学常识，就无法正确理解。

(4)军队卫生统计的许多第一手资料产生于军医之手，如门诊登记、体格检查记录、病历等。军医应该清楚如何从这些资料中提取出统计信息，并使之发挥更大的作用。

(5)具备一定的统计学知识，能对科研数据进行基本的统计分析，是医学工作者应具备的基本素质。

二、军队医学统计工作的要求

军队医学统计工作的基本要求是准确、完整、及时,并注意保密。

(一)统计数据必须准确

统计数据的准确首先要求原始登记准确，因此要求军医须正确理解医疗文书的各项内容并认真填写；其次是在数据的整理汇总过程中防止出现过失误差。负责统计工作的军医决不能在统计数据上弄虚作假，并且要勇于同篡改统计数字的不良行为作坚决斗争。

(二)统计资料力求完整

统计资料应能反映一个单位不同时期卫生工作全面情况，因此要求登记和统计的项目不能残缺不全。对于历年积累的统计资料，要定期进行整理、汇总，撰写分析报告，

并列入移交。

（三）提供统计资料要及时

统计资料具有很强的时效性，在战时更是如此。各基层单位应按规定时间向上级单位报送统计数字，否则将影响全军统计资料的汇总。此外，许多专题统计调查规定有统一的起始时间，如果延误了时间将影响整个统计调查的工作进度，甚至会影响整个调查结果的代表性和科学性。

（四）提高警惕、保守军事秘密

军队医学统计资料直接或间接地反映了部队的编制、人数、伤亡和疫情等军事秘密，如果麻痹大意或保管不当容易造成失密或泄密。因此，军医要提高警惕，对所掌管的统计资料要切实按保密制度严加保管。因学术交流需公开发表统计资料时，要防止泄密。在对外报告时，单位名称要用代号，尽量避免使用绝对数。某些疫情资料的公布，还须经国家有关部门的批准。

三、军队医学统计工作的步骤

军队医学统计工作的步骤与其它行业的统计工作的步骤基本相同，即先有一个周密、完整的统计设计，然后再按照设计的要求收集资料、整理资料和分析资料。

（一）统计设计

在医学研究设计中，除了医学专业设计外，还需要对数字资料的收集、整理、分析的方法和过程进行必要的设计，称为统计设计。医学研究一般分为调查与实验两类，统计设计也相应地分为调查设计与实验设计。由于临床试验研究的对象主要是病人，既不能像实验研究那样进行随机抽样与分组，也很难严格控制非实验因素，因而临床试验设计与一般实验设计有所不同（详见第八章）。

（二）收集资料

收集资料是根据统计设计的要求采集原始数据的过程。由于原始数据关系到最终统计分析的正确性，而且原始数据中有错误很难事后补救，所以对收集资料的每一个环节必需高度重视。

1.资料的来源

(1)经常性资料

经常性资料（regular data）来自于常规的数据采集系统。在我国，主要是国家、军队规定的统计报告(报表)制度和日常医疗卫生工作登记制度，如出生、死亡登记、疫情报告和统计报表制度 我军平时军队卫生统计报表制度、军医门诊(巡诊)登记制度、医院病案书写和管理制度；战时伤票登记制度等。

国家的法律、军队的纪律、卫生统计常设机构和专(兼)职统计人员为经常性统计资料的准确性提供了有力的保障。我军关于军队卫生统计报表的纪律是：①对报表项目和格式不得擅自修改；②按规定日期和程序上报；③不得伪造数据；④保守机密。

我军团以上卫勤机关设专人分管统计工作，驻军以上医院设有专职统计员或病案统计室，每个军医要自觉接受他们的指导和监督，严格执行各项登记、统计规章制度，为保证各项原始记录的正确、完整做出自己的努力。

(2)专项调查资料

专项调查资料是为特定目的而专门收集的资料。

2.数据的采集方式

数据的采集方式主要有以下三种：①直接使用测量工具，如磅称、体温表、血压计、血气分析仪等；②主观判断，如常见病诊断、临床疗效评价等；③借助测量仪器进行判断，如阅看X线片、红细胞计数等。

从数据质量看，直接用仪器测量的数据受主观因素的影响较小，比较可靠。因此，在设计调查项目时应尽量选择测量指标。如果在调查项目中有主观判断的项目，数据采集人要统一判别或评价标准，必要时进行集体培训。

（三）整理资料

整理资料是将原始记录按统计的要求进行归纳分组的过程。原始记录经整理后，可以将原始记录转变为有分析价值的数字信息，并且容易显现出隐藏在原始记录中的统计规律。此外，在资料的整理过程中，通过逻辑检查和简单计算，如经过简单的归纳和整理，计算均数、率等统计指标，能够及时发现、更正或剔除错误的原始记录，以保证下一步统计分析的正确性（具体内容详见第二、三章）。

（四）分析资料

医学研究的最终目的是通过样本的信息去推断总体的情况，包括两个重要领域：一是参数估计，即用样本统计量来估计总体中相应的参数；二是假设检验，即依据资料性质和所需解决的问题，建立统计假设，然后采用适当的检验方法，根据样本是否支持所作的假设，来决定该假设应当拒绝或不拒绝。

分析资料是在分组的基础上，计算有关统计指标，然后结合各专业的研究背景和业务知识对统计指标的大小和差别做出合理的分析、解释。

分析资料主要有以下三个步骤：

(1)选择正确的统计指标。这里最主要的判别标准就是要根据资料的具体分布情况来正确选择，使计算的统计指标能反映出该资料的特征。

(2) 为便于分析和比较，可将分组计算后的统计指标列在统计表内或作统计图（详见第三章)。

(3) 对经常性资料和普查资料，可直接比较统计指标的大小和组间差别，并做出合理的分析和解释。如果是根据抽样调查资料推论全体观察对象的特征，则需要估计统计指标的抽样误差，并对组间差别作假设检验，有关统计方法将在第七章以后介绍。

第二节　常用名词概念

这里先介绍一些常用的医学统计学的基本概念，从中也可了解到本课程的基本思路和总体框架。

（一）同质与变异

同质（homogeneity），系指事物的性质、影响条件或背景相同或非常相近；变异，系指相同条件下的同类个体之间某一方面发展的不平衡性。例如研究儿童的身体发育，同性别、同年龄的儿童，在统计上称为“同质”观察单位，但是，一批同性别、同年龄的儿童，他们的身高、体重、胸围等却各不相同，互有差异，这就是变异（variation）。又如在同

样条件下接受同一种疗法的某病患者,可认为是‘同质’观察单位,但他们的疗效并不一样,有的被治愈,有的略有好转,还有的反而恶化甚至死亡,这也称作‘变异’。个体变异可表现为定量,如身高的高低、体重的轻重等,亦有表现为定性的,如用某药治疗某种病人后的痊愈、好转、恶化等。

同质观察单位之间的个体变异，是生物的重要特征，是偶然性的表现，是由于生物体内外环境中多种因素的综合影响造成的，其中许多因素是未知的，也是难以控制的。统计研究的是有变异的事物，其任务就是在同质分组的基础上，通过对个体变异的研究，透过偶然现象，反映同质事物的本质特征和规律。

（二）总体与样本

总体（population）是根据研究目的确定的同质的研究对象的全体。更确切地说，是性质相同的所有观察单位某种变量值的集合。例如研究某地某年正常成人的血压值，则研究对象是该年的正常成人，观察单位是每个人，变量是血压，变量值是测得的血压值，该地该年全部正常成人的血压值就构成一个总体。它的同质基础是同一地区、同一年份、同为正常成人。这里的总体只包括有限个观察单位，称为有限总体。有时总体是设想的，如研究高血压患者用某药治疗后的血压，这里总体的同质基础是同为高血压患者，同用某药治疗，包括设想该药治疗的所有高血压患者，其观察单位数显然是不确定的、无限的，称为无限总体。

医学研究中，很多是无限总体，要直接研究总体的情况是不可能的。即使对有限总体来说，若包括的观察单位数过多，直接研究总体既耗费人力、财力，有时也是不可能和不必要的，如成批罐头食品的生产，不可能将所有罐头逐个作食品卫生检查。所以，在实际工作中，常常是从其样本的信息来推断总体的性质。

样本（sample）是从总体中随机抽取的部分观察单位，其实测值构成样本。样本性质的研究是对总体性质研究的基础，如在对某地区成年人血压的研究中，可从该地某年的正常成人中，随机抽取200人，分别测定其血压值，组成样本，计算样本均数，用来估计该地该年正常成人血压的总体均数。这种方法称为抽样研究，这是一种很常用、极其重要的科学研究方法。

要做好抽样研究，必须注意：①样本的可靠性。首先要明确划清总体的同质范围，也就是确定被研究对象的同质基础，包括时间、空间、条件等，如观察某种疫苗预防接种的效果，首先要确定易感人群，接种组和未接种组的界限清楚，未完成全程接种的人如何划分等也要作出规定。②样本的代表性。必须遵循随机抽样的原则，必须保证样本中有足够的观察单位数，即样本含量要足够大。

（三）参数与统计量

参数（parameter）是用来表示总体分布特征的统计数字。统计中常用的总体参数有描述总体中心位置或集中趋势的总体平均数、中位数和众数等；有描述总体离散度的总体标准差、极差和四分位数间距等。研究总体的特征往往是医学统计研究的目的，若能直接观察总体的每一部分无疑是最理想的，但事实上是做不到的，主要因为：①总体中的个体数无穷多；②总体随时间和空间不断变化，如某病患者不断发生、其转归情况也不断变化；③研究方法具有破坏性；④用于某项研究的技术力量、设备条件或资金不足等，限制了人们直接了解总体全貌。故人们只能从总体中抽取一部分具有代表性的

需要全本请访问 www.ertongbook.com

样本进行检测，从而对整个总体进行推断。这也是整个概率统计研究的最基本的思路和方法。

统计量(statistic）是我们对样本做了若干次观察后，依据观察值推出的反映样本分布特性的一些量。例如，为了了解健康成年男子每升血液中白细胞数而进行检验，由所测得的一系列数值算出一个平均数，这个平均数就是一个“样本统计量”。样本标准差和样本率也是统计量。

从研究总体中抽取部分观察单位构成具有代表性的样本，用样本统计量推断总体参数的方法称为抽样研究。抽样研究有两个目的：一是估计总体参数；二是判断某个随机变量是否取自某一预期或假设的总体。它们也是参数估计与统计量关系研究的主要内容。

参数估计的内容包括点估计和区间估计。点估计是用一个样本统计量（点值）作为总体参数的估计值；区间估计是从样本统计量存在抽样误差这一角度出发，把抽样误差引入统计量，来判定的一种有概率意义的误差范围，以对总体参数进行统计。点估计的常用方法有矩法和极大似然法。矩法，即用样本 r 阶矩作为总体 r 阶矩的统计量。极大似然法，即由似然函数确定参数的极大似然估计量。一个良好的参数点估计应具备以下三个基本特性，即无偏性、一致性和有效性。无偏性是指在同一总体中独立随机抽取 n 个样本,能得到 n 个样本统计量，这些样本统计量的平均数即数学期望必须等于总体参数。θ 表示总体参数,$\hat{\theta}$ 表示 θ 的点估计量,无偏统计量 $\hat{\theta}$ 应满足 $E(\hat{\theta})=\theta$。一致性是指当样本含量无限增大时，样本统计量具有趋向总体参数的特性。用数学表达式说明时,可表示为 $P(|\hat{\theta}-\theta|<\varepsilon)=1$,$\varepsilon$ 是任意小的正数，表示误差。这样的样本统计量 $\hat{\theta}$ 称为总体参数的一致性估计量。有效性包括了这样的含义，即有两个无偏估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$,它们的方差分别为 $S_{\hat{\theta}_1}^2$ 和 $S_{\hat{\theta}_2}^2$,如果 $S_{\hat{\theta}_1}^2<S_{\hat{\theta}_2}^2$,就认为 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 对总体参数的估计更有效。

参数的区间估计是基于样本统计量存在抽样误差这一事实，从而用样本均数或样本率去估计总体均数或总体率时，存在着估计误差，这种估计误差也就是不准确度。如果能了解伴有抽样误差的样本均数或样本率围绕总体均数或总体率分布的规律，就可以在实行点估计的同时，把表示抽样误差数值大小的标准误引入估计量，从而计算出某种有特定概率意义的误差范围或称置信区间，用来估计总体均数或总体率。这时，虽然总体均数或总体率的确切值仍属未知，但人们能够根据所计算出的特定概率置信区间说明总体均数或总体率有多大的可能性被包括在内。由于通常对概率置信区间所取的概率值较大(如 95% 或 99%)，故此区间估计比点估计能表达得更充分，更具可信性。

参数与统计量另一重要关系研究是参数假设检验。假设检验是指事先假设某一样本具有某种分布特征(如具有某种参数,或遵从某种分布),然后再检验这个假设是否可信。

在同一总体中，随机抽样获得的样本统计量是否由假设的总体中随机抽得，以及该样本统计量与已知总体间的差别是否可用抽样误差来解释，就应结合这种统计量的抽样分布进行考察。从样本统计量分布的概率范围出发，制定一个抽样误差的界限作为判断标准，这个人为的临界限称为显著性界限。它是以小概率事件作为临界值的，即在大量抽样中，超出该临界点的统计量分布的概率值很小。习惯上，人们通常取 $P=0.05$

作为小概率事件。小概率事件代表了“样本统计量与总体参数偏差很大”这一事实，而这种事件在多次抽样中出现的总概率是很小的，在一次抽样中则几乎不可能发生。研究者将获得的统计量与假设的总体参数值的差别转换为 t 值(或 U 值或其它统计量)，用小概率事件的临界点 $t_{0.05(\nu)}$ 及 $t_{0.025(\nu)}$ 来进行判断，若 $|t|<t_{0.05(\nu)}$(单侧)或 $|t|<t_{0.025(\nu)}$(双侧)，则推断统计量来自假设参数的总体。它们间的差别是由于抽样误差造成的。反之，则推断统计量并非来自假设的总体，也就是说，它们间的差别很难用抽样误差来进行解释。由此可见，统计检验就是带有概率性质的反证法。利用概率论中“小概率事件在一次抽样中几乎不可能出现”的原理对假设作出判断。

(四) 确定性现象与随机现象

在一定条件下，必定会发生或必定不会发生某一种结果的现象，称为确定性现象(decided phenomenon)。其表现结果为两种形式相反的事件：肯定发生某种结果的叫做必然事件，肯定不发生某种结果的叫做不可能事件。例如，在适当温度下经过一定时间孵化，正常受精鸡蛋必然会孵出小鸡来，而石头是不可能孵出小鸡来的。

在同样条件下，可能出现两种或多种结果，而事先不能确定发生哪一种结果的现象，称为随机现象(random phenomenon)。其表现结果叫做随机事件。例如胎儿的性别，可能是男，也可能是女，事先不能肯定，这是一个随机现象，而结果胎儿为男性，则是一个随机事件。随机现象有两个主要特征：①随机性，即它的发生是随机的；②规律性，即它每次发生的可能性的大小是确定的。在医学研究中，大量的是随机现象，通过对某随机现象的研究，找出其规律性。

(五)概率

概率(probability) 的直观意义是描述事件发生可能性的大小的指标。下面，我们将给出用于建立概率理论体系的严密数学定义：设(Ω, F)是可测空间，对每一集 $A\in F$，定义实值函数 $P(A)$，它满足如下三个条件：

(1)对每一 $A\in F$，有 $0\leqslant P(A)\leqslant 1$；

(2)对必然事件 $P(\Omega)=1$；

(3)对任意 $A_i\in F$

$$A_i\cap A_j=\Phi \quad i\neq j$$

恒有 $P(\bigcup_{i=1}A_i)=\sum_{i=1}P(A_i)$ 则称实值集函数 P 为(Ω,F)上的概率，$P(A)$就称为事件 A 的概率。

[例 1-1] 在生男生女问题中，其样本空间为 $\Omega=\{\omega_1,\omega_2\}$，$\omega_1=$(男)，$\omega_2=$(女)，基本事件为 $A_1=\{\omega_1\}$，$A_2=\{\omega_2\}$。已知 $F=\{\Phi,A_1,A_2,\Omega\}$，这里 Φ 表示“既不生男，也不生女”这一事件，Ω 表示“生男或生女”这一事件。

对 F 中的集合，对应一实数 P，定义如下：

$P(\Phi)=0$；$P(A_1)=P(A_2)=\frac{1}{2}$；$P(\Omega)=1$ 显然，P 满足上述定义中的 (1)(2)(3)，因此，P 是(Ω, F)上的概率。

作为上述关于概率定义的特殊情况，一般常用的有古典概率、几何概率。 为便于理解，我们给出古典概率的定义：设计一个试验，其所有可能出现的结果有 n 个，这些结

果是互不相容且等可能的，其中有 f 个结果具有属性 A，则具有属性 A 的事件(简称事件 A，下同)出现的概率记作

$$P(A)=f/n \qquad \text{(公式 1-1)}$$

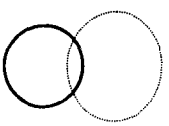
其统计意义是：设在相同条件下，独立重复进行 n 次试验，事件 A 出现 f 次，则称 f/n 为事件 A 出现的频率。当 n 逐渐增大时，频率 f/n 始终在某一常数 P 的左右作微小摆动，就称 P 为事件 A 的概率，记作 $P(A)=P$。在许多实际问题中，当概率不易求得时，只要 n 充分大，可以将频率作为概率的估计值。

由此，必然事件的概率等于 1，不可能事件的概率等于 0，随机事件的概率介于 0 和 1 之间。概率越接近 1，表示事件发生的可能性越大 概率越接近于 0，表示事件发生的可能性越小。

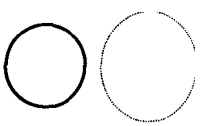
事件之间的相互关系及其示意图如下：



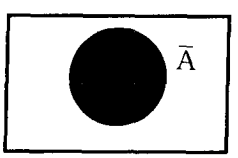
$A+B=$｛事件 A 和 B 至少有一个发生｝



$AB=$｛事件 A 和 B 同时发生｝



A 和 B 互不相容



$\bar{A}=$｛事件 A 的对立事件｝

(1)概率运算法则

加法法则 $P(A+B)=P(A)+P(B)-P(AB)$ (公式 1-2)

如 A 和 B 互不相容

$$P(A+B)=P(A)+P(B) \qquad \text{(公式 1-3)}$$

这一法则可以推广到有限个互不相容的事件。

$$P(A)+P(A)=1 \qquad \text{(公式 1-4)}$$

乘法法则 在事件 A 已发生的条件下，事件 B 发生的概率，就称为事件 B 的条件概率，记作 $P(B/A)$。

对于任意两事件 A 和 B 同时发生的概率为

$$P(AB)=P(A)P(A/B) \qquad \text{(公式 1-5)}$$

$$P(A\dot{B}=P(B)P(B/\dot{A} \qquad \text{(公式 1-6)}$$

如 A(或 B)发生与否并不影响 B(或 A)的概率，就说 A 和 B 相互独立，则事件 AB 的概率等于 A 的概率与 B 的概率之积，即

$$P(AB)=P(A)P(B) \qquad \text{(公式 1-7)}$$

这一法则可推广到有限个相互独立的事件

[例 1-2] 对某地区 2 000 名 60 岁以上老年人的疾病调查中发现：710 人患有高血压，315 人患有糖尿病，150 人同时患有高血压及糖尿病。

则

$A=$｛患有高血压病｝

$B=\{$患有糖尿病$\}$

$A+B=\{$患有高血压或糖尿病$\}$

$AB=\{$既患有高血压又患有糖尿病$\}$

$\bar{A}=\{$未患有高血压病$\}$

由古典概率的定义，且调查数 n 充分大，于是可以将频率作为概率

$P(A)=710/2000=0.355$；

$P(B)=315/2000=0.1575$ ；

$P(AB)=150/2000=0.075$;

$P(A+B)=P(A)+P(B)-P(AB)=0.4375$ ；

$P(\bar{A})=1-P(A)=0.645$ ；

[例 1－3] 据估计，成年人口中约 15% 有高血压，又所有成年人口中约 75% 不觉得血压高。同时估计人口中约 6% 患有高血压而不觉有病。如果一成年人认为自己没有高血压，此人实际有病的概率是多少。

$A=\{$病人不觉有病$\}$

$B=\{$患有此病$\}$

则

$P(A)=0.75$　　$P(B)=0.15$

$P(AB)=0.06$

$P(B/A)=0.06/0.75=0.08$

全概率公式：

$A_i\cap A_j=\Phi,\bigcup_{i=1}^{n}A_i=\Omega,P(A_i)>0$ 则

$$P(B)=\sum P(B/A_i)P(A_i) \qquad \text{（公式 1－8）}$$

Bayes公式：

若　$A_i\cap A_j=\Phi,\bigcup_{i=1}A_i=\Omega,P(A_i)>0$

则在事件 B 出现的条件下 $P(B)>0$，事件 A_i 出现的概率为

$$P(A_i/B)=\frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{n}P(A_i)P(B/A_i)} \qquad \text{（公式 1－9）}$$

[例 1－4] 在对某单位生产的注射器进行检验中发现：在 1 000 只中有 20 只是不合格产品，现做不放回抽取 2 只注射器，问第 2 次抽取到的是不合格产品的概率。

$A=\{$第一次抽到的是不合格品$\}$

$B=\{$第二次抽到的是不合格品$\}$

$A_1=A$　　$A_2=A$

$P(A_1)=20/1000$　　$P(A_2)=980/1000$

$P(B/A_1)=19/1000$　　$P(B/A_2)=20/999$

$P(B)=20/1000\cdot 19/999+980/1000\cdot 20/999$

$=0.0234$

[例 1－5] 某地区根据多年的某种病史分析，将病例分为甲状腺机能低下（H_1）、

正常(H_2)与亢进(H_3)三类,并得:$P(H_1)=0.15, P(H_2)=0.65, P(H_3)=0.20$。若仅按'食欲亢进'指标($A$)而言,根据积累的资料得:$P(A/H_1)=0.08$, $P(A/H_2)=0.10, P(A/H_3)=0.61$。今有某病员新进食量大增,试分别求其甲状腺机能低下、正常与亢进的概率。

$$P(H_1)P(A/H_1)=0.15\times0.08=0.0120$$
$$P(H_2)P(A/H_2)=0.65\times0.10=0.0650$$
$$P(H_3)P(A/H_3)=0.20\times0.61=0.1120$$
$$\sum P(H_i)P(A/H_i)=0.1990$$
$$P(H_1/A)=0.0120/0.1990=0.0603$$
$$P(H_2/A)=0.0650/0.1990=0.3266$$
$$P(H_3/A)=0.1220/0.1990$$

故仅按'食欲亢进'一项指标而言,该病员患甲状腺机能亢进的可能性最大,大约2倍于正常,10倍于低下。

(六)误差

观察值与真值之差称为误差(error)。误差主要有三类:

(1)过失误差 又称粗差,由观察者的错误造成。如记录误差、计算误差或错用分析方法等。这类误差是不允许的,必须消除。

(2)系统误差 由于受试对象、研究者、仪器设备、研究方法、非实验因素影响等原因造成的,有一定倾向性或规律性的误差。如抽样不匀、仪器未校正、操作技术不当、计算方法不完善等。这类误差应通过严格的设计和严密的技术措施尽力消除或减到最小。

(3)随机误差 或称偶然误差,是排出了上述误差后,由于众多无法控制的因素引起的,大小及方向不一的随机变化。随机误差是不可避免的。它服从一定的分布规律,可以用概率统计方法处理。在随机误差中,最重要的是抽样误差。

由于总体中各观察单位间存在个体变异,抽样研究中得到的样本不会完全反映总体的信息。样本统计量与总体参数间存在的差异称为抽样误差。显然,抽样误差愈小,用样本推断总体的精确度就愈高,反之亦然。由于生物的个体变异是客观存在的,因而抽样误差是不可避免的,但抽样误差是可以控制的。抽样误差有一定的规律性,在没有人为或外来因素干扰的条件下,如果样本不是太小,则抽样误差将会时大时小而小多大少,或正或负而正负近似各半,随着抽样次数的增加,各抽样误差的代数和逐渐趋向于"0"。因此,大多数来自样本的平均指标或相对指标都围绕在相应总体指标的上下。在根据样本估计总体时,这是必须正确领会的一个基本概念。

第三节　变量与资料类型

简单地说,变量(variation)即指标,是指观察单位的某项特征,观测所得结果称为变量值或指标值。变量值可以是定量的、定性的与半定量的,相应的资料有计量资料、计数资料与等级资料。

(一)连续型变量与计量资料

身高、体重、胸围和血压等的变量值是定量的，一般有度、量、衡或其它单位。如调查某部队入伍一年的战士身体发育状况，以人为观察单位，测得每个人的身高（cm）、体重(kg)、胸围(cm) 和血压(kPa)等。尽管有的指标的取值是有限的，但为便于理论研究，人们将其视为连续型变量(continued variation) 。一般将这类变量的观察值构成的资料称为计量资料(measurement data)。如表 1－2 的第(3)栏的年龄、第(8)栏的PCNA指数、第(9)栏的生存时间为连续型变量。

表 1－2　胆管癌患者部分指标

编号 (1)	性别 (2)	年龄(岁) (3)	部位 (4)	分化程度 (5)	分期 (6)	肝转移 (7)	PCNA 指数 (8)	生存时间(月) (9)
1	男	61	上	低分化	Ⅰ	阳性	52	14
2	女	58	中	高分化	Ⅱ	阴性	89	20
3	女	63	上	高分化	Ⅳ	阴性	93	19
4	女	71	下	中分化	Ⅱ	阳性	78	5
5	男	59	上	高分化	Ⅲ	阴性	85	35
…	…	…	…	…	…	…	…	…

(二)离散型变量与计数资料

某些变量只能定性地划分成少数几个互相排斥的类型，如表 1－2 第(2)栏的性别变量的取值为“男”“女”，表 1－2 第(7)栏的肝转移变量的取值为“阳性”“阴性”再如血型变量的取值为“O”、“A”、“B”、“AB”等。这类变量称为分类变量或名义变量，如果给各类别适当赋值，这类变量成为假变量。如性别变量 X 与肝转移变量 X 可分别定义为一个二值变量：

$$\text{性别变量 } X=\begin{cases}1 & \text{女性}\\ 2 & \text{男性}\end{cases},\text{肝转移变量 } X=\begin{cases}1 & \text{阳性}\\ 0 & \text{阴性}\end{cases}$$

血型变量 X 可定义为 3 个二值变量。令

$$X_1=\begin{cases}0 & \text{非 O型}\\ 1 & \text{O 型}\end{cases}\qquad X_2=\begin{cases}0 & \text{非 A 型}\\ 1 & \text{A 型}\end{cases}\qquad X_3=\begin{cases}0 & \text{非 B 型}\\ 1 & \text{B 型}\end{cases}$$

表 1－3　血型变量 X 的二值变量

血型	变量		
	X_1	X_2	X_3
O	1	0	0
A	0	1	0
B	0	0	1
AB	0	0	0

血型变量 X 表示为表 1－3。这类只能在几个数中取值的变量称为离散型变量(discontinued data)。离散型变量的频数资料称为计数资料(count data) 。离散型变量中也有的是计量资料，如抗体效价，每户人口数等。

(三)有序变量与等级资料

某些变量的检测结果是半定量的，如表 1－2 的第(4)栏的部位分为上、中、下，第

需要全本请访问 www.ertongbook.com

(5)栏的分化程度分为高分化、中分化、低分化，第(6)栏的分期分为Ⅰ、Ⅱ、Ⅲ、Ⅳ等。这些等级可以取不同的值来描述，但所取的值只反映类别的等级或秩次，并不反映观察单位的确切定量值，这类变量称为有序变量（ordinal variation)。如表1－2的第(4)栏部位变量可取值1、2、3，分别代表上、中、下，第(5)栏的分化程度变量可取值1、2、3，分别代表高分化、中分化、低分化，第(6)栏的分期变量可取值1、2、3、4，分别代表Ⅰ、Ⅱ、Ⅲ、Ⅳ期。有序变量的频数资料称为等级资料（ordinal data)。

在医学实践中，根据分析的需要，各类变量可以互相转化。例如，血红蛋白原属连续型变量，若按血红蛋白正常与异常分成两个类别，就成为离散型变量；若按血红蛋白量的多少分成重度贫血、中度贫血、轻度贫血、血红蛋白正常、血红蛋白增高五个等级，就成为有序变量。再如在多变量分析中，定性指标数量化时，将无病与有病分别取为0与1，或将上述血红蛋白量的五个等级依次取为1,2,…,5，这时计数资料或等级资料就转化为计量资料。但须明确，凡能计量的，应尽可能采用计量资料，因为计量资料可以得到较多的信息。

练习题

1. 军队卫生统计工作的基本要求是什么？
2. 试述统计设计的重要性。
3. 试述统计常用名词间的相互关系。

（易东）

第二章　常用统计指标

统计指标描述了统计数据的分布特征，是统计分析的基本依据。根据样本计算出来的统计指标又称为统计量（statistic）。在医学统计学中常用的统计指标有：总量指标、相对数指标、平均数指标、变异指标。

第一节　相对数指标

一、相对数指标的意义

表 2－1 为甲、乙两部队在某季度的感冒发病情况，其中部队发病人次、未发病人次均是对相应的统计指标进行分类整理后得到的，这些数据我们称为绝对数或总量指标。总量指标主要反映了在一定条件下某种事物的规模或水平，是计划和总结工作的依据。其主要缺点是：①不利于保密，如部队人数、传染病发生数、战斗伤亡人数等是保密的。②不便于比较，如表 2－1 中两部队发病人次的多少并不能说明发病程度的差异，因为两部队的人数不等。为此，在实际工作中需要计算相对数来克服以上缺点。

表 2－1　某季度甲、乙两部队的感冒发生情况

部别	总人数	发病人次	发病率(%)
甲部队	1 834	58	3.16
乙部队	1 313	42	3.20
合　计	3 147	100	3.18

相对数是有联系的两个指标之比。计算相对数的意义在于：一是便于表达事物发生的普遍程度；二是把基数标准化，便于相互比较。如表 2－1 的发病率就是相对数，由发病率可以看出，甲、乙两部队的感冒发病情况严重程度相似。

二、几种常用的相对数

(一)率(rate)

又称强度相对数或频率，它是一定时期或一定范围内某现象的发生数与可能发生该现象的总数之比，反映某现象出现的频率或强度。常以百分率（%）、千分率(‰)、万分率(1／万)、十万分率(1/10 万）等比例基数表示。计算公式为：

$$率 = \frac{某现象的发生数}{可能发生某现象的总数} \times 比例基数 \qquad （公式\ 2-1）$$

公式 2－1 是率的通用计算公式，由此可定义有具体意义的强度相对数，如患病率、

发病率、治愈率、病死率、死亡率、减员率等。

[例 2－1] 如表 2－1 中，甲部队的感冒发病率为：$\frac{58}{1834}\times 100\% = 3.16\%$。

（二）构成比

又称结构相对数或百分比（percentage），表示某现象内部各个构成部分所占的比重或分布，常以百分数表示，计算公式为：

$$\text{构成比}=\frac{\text{某一构成部分的数量}}{\text{各构成部分数量之和}}\times 100\% \qquad \text{（公式 2－2）}$$

构成比是以 100 作为基数，各构成部分所占的比重之和必须为 100%。

[例 2－2] 某年某地区共发生各种疟疾 97 例，其中恶性疟 68 例，间日疟 12 例，三日疟 17 例，则三种疟疾的构成情况如下：

$$\text{恶性疟：}\frac{68}{97}\times 100\% = 70.1\%$$

$$\text{间日疟：}\frac{12}{97}\times 100\% = 12.4\%$$

$$\text{三日疟：}\frac{17}{97}\times 100\% = 17.5\%$$

（三）比

亦称比值（ratio），是甲、乙两个相关指标之比，说明甲为乙的若干倍或百分之几。两个指标可以是同类的，如各地区某种传染病发生数之比；或是不同类但有联系的，如医院的门诊人次数与床位数之比等。通常以某现象的数量为 1 或 100 作基数，看另一现象的数量是多少。计算公式为：

$$\text{比}=\frac{\text{甲现象的数量}}{\text{乙现象的数量}}(\text{或}\times 100\%) \qquad \text{（公式 2－3）}$$

[例 2－3] 某地 1991～1995 年肺癌患者中男性 478 例，女性 136 例，试求当地肺癌患者的性别比值。

$$\text{性别比}=\frac{\text{男}}{\text{女}}=\frac{478}{136}=3.5$$

即男性患者为女性患者的 3.5 倍。该式亦可写作男：女 $=3.5:1$。

比值所对比的两个指标可以是绝对数、相对数或平均数，但二者须一致。常用的比值有流行病学中的相对危险度，人口统计中的男女性别比，医院管理中的医生数与床位数之比等。临床上亦有许多比值指标，如骨髓粒细胞与红细胞之比、血浆白蛋白与球蛋白之比、体液中钠盐与葡萄糖之比等。

（四）动态数列（dynamic series）

又称动态相对数，是同一现象的统计指标在不同时间上的对比，用以反映某现象随时间的变化和发展趋势。计算时，如果统一用某个时间的指标作基数，以各时间的指标分别与之对比，称为定基比；如果依次以前一个时间的指标作基数，以相邻的后一个时间的指标与之对比，称为环比。

[例 2－4] 表 2－2 是某院 1988～1994 年的入院人次数，试描述其变化情况。

表 2－2 中，第(3)栏定基比发展速度是以 1988 年入院人次数作基数（分母）计算的，如：

$$1989\text{ 年的发展速度} = \frac{6906}{6330} \times 100\% = 109.1\%$$

表 2－2　某院入院人次数的变化

年份	入院人次数	发展速度(%)		增长速度(%)	
		定基比	环　比	定基比	环　比
(1)	(2)	(3)	(4)	(5)	(6)
1988	6 330	100.0	100.0	－	－
1989	6 906	109.1	109.1	9.1	9.1
1990	7 135	112.7	103.3	12.7	3.3
1991	7 080	111.8	99.2	11.8	-0.8
1992	7 663	121.1	108.2	21.1	8.2
1993	7 467	118.0	97.4	18.0	-2.6
1994	7 938	125.4	106.3	25.4	6.3

$$1990\text{ 年的发展速度} = \frac{7135}{6330} \times 100\% = 112.7\%$$

第(4)栏环比发展速度是依次以上一年入院人次数为基数计算的，如：

$$1990\text{ 年对 }1989\text{ 年的发展速度} = \frac{7135}{6906} \times 100\% = 103.3\%$$

$$1991\text{ 年对 }1990\text{ 年的发展速度} = \frac{7080}{7135} \times 100\% = 99.2\%$$

第(5)、(6)栏的增长速度分别为第(3)、(4)栏的发展速度减去 100%。

动态数列常用百分点表示环比增长速度的变化。如表 2－2 中 1989 年的环比增长速度为 9.1%，1990 年的环比增长速度为 3.3%，则 1990 年的增长速度比 1989 年的增长速度低 5.8 个百分点，但不能说 1990 年的增长速度比 1989 年低 5.8%；又如 1991 年的环比增长速度为 －0.8%，则 1991 年的增长速度比 1990 年低 4.1 个百分点。

三、应用相对数的注意事项

相对数意义明确，计算简单，应用普遍，但在计算和应用时要注意以下几点：

(1)根据需要正确选择相对数。在相对数中，最容易混淆的是率（强度相对数）与比（结构相对数）。常见错误是以构成比代率，例如表 2－3 是把病人职业构成比误为“发病率”的一例。构成比只能说明各种职业沙眼患者占该病总就诊人数的比例，不能说明各种职业沙眼患病的普遍程度。要说明患病普遍程度，应计算其发病率（或患病率）。

表 2－3　各种职业沙眼患者就诊情况

职业	工人	农民	学生	干部	其他	合计
人数	1792	754	920	320	432	4 218
发病率(%)	42.5	17.9	21.8	7.6	10.2	100.0

临床中根据收治的某病病人数，只能计算病人的年龄构成比、性别构成比、职业构成比等，这些都是结构相对数。

(2)计算观察单位数不等的几个率的平均率时不能将几个率直接相加求平均率。如表 2－4 是某地三个生产队斑釉发生率，不能用（85.57% + 91.59% + 92.86%）/3 = 90.01%的方法求三个队的斑釉平均发生率，而应将三个生产队的总发生数除以总检查人数，即(272/302)×100% = 90.07%。

表 2－4 某地三个生产队的斑釉发生率

生产队	检查人数	斑釉人数	斑釉发生率(%)
甲	97	83	85.57
乙	107	98	91.59
丙	98	91	92.86
合计	302	272	90.07

(3)正确选择分母。计算强度相对数时，分母中的每一观察单位应当都有可能成为分子。

(4)分母应当够大。习惯上分母大于 100 时，所得的相对数说服力较强；分母略小于 100 时,相对数仍有一定价值 分母小于 20 时可靠性较差。如果分母太小，宜用绝对数表示,如“4 例全部治愈”,“3 例中死亡 1 例 ”等。

(5) 应用相对数比较时应注意是否具有可比性，同时还要注意其内部构成是否相同。例如某部队对老战士计算三年累计痢疾发病率，而对新战士只计算本年度痢疾发病率，结果得出“新战士的痢疾发病率低于老战士”的结论。这显然是不正确的。因为计算的时期不同，不具备可比性。如果都计算本年度的发病率就可以比较了。

(6) 根据样本数据计算的强度相对数，要考虑抽样误差的影响。

第二节 平均数指标

一、频数表的编制

(一)频数表的概念

样本中的数据个数 n 称为样本含量。若样本的含量较大，其分布特征可以用频数表(frequency table)来表示,资料可以用频数表来描述。如表 2－5,第(1)栏‘子女数’为各观察单位（每个已婚育龄妇女）的观察值，第(2)栏为相应的观察单位数，亦称频数。列有类似第(1)、(2)栏的统计表称为频数分布表，简称频数表。必要时还可计算出各观察值范围内的相应的频率，用以反映该观察值范围内的观察单位数占总观察单位数的百分比。有时还需要计算第（4)栏的累计频数和第 5)栏的累计频率。

表 2－5　1982 年我国农村已婚育龄妇女按现有子女数的分布

子女数 (1)	妇女数 (2)	频率(%) (3)=(2)/145 525	累计频数 (4)	累计频率(%) (5)=(4)/145 525
0	13 751	9.45	13 751	9.45
1	25 171	17.30	38 922	26.75
2	30 426	20.91	69 348	47.65
3	28 560	19.62	97 908	67.28
4	21 719	14.92	119 627	82.20
5	13 695	9.41	133 322	91.61
6	7 255	4.98	140 577	96.60
7	3 268	2.25	143 845	98.85
8	1 151	0.79	144 996	99.64
9	373	0.26	145 369	99.89
≥10	156	0.11	145 525	100.00
合计	145 525	100.00	—	—

（二）频数表的编制方法

离散型变量与连续型变量在频数表的编制方法上略有差别。对离散型变量，当观察值较少时，可直接列出其可能的观察值及其相应的频数，如表 2－5 当观察值较多时，可以用组段形式表示，如表 2－6。每个组段的起点成为该组段的下限，终点称为上限，上限与下限之差称为组距。如表 2－6 中第一组的下限为 0，上限为 3，组距为 3，余类推。归组后，该组的变量值用该组段的中值代替，称为组中值。如第一组的组中值为 1.5，第二组为 5.5。

表 2－6　204 名轧钢工人白细胞中大单核所占百分比

大单核数(个/每百个细胞)	0～3	4～7	8～11	12～15	16～19	合计
人数	64	92	45	2	1	204

对连续型变量资料，编制频数表的方法比离散型变量资料稍复杂一些。现以表 2－7 的数据说明其具体步骤。

表 2－7　某市 110 名健康男性工人的血红蛋白量(g/L)

118	148	158	163	138	156	134	140	152	132
149	165	160	119	174	137	152	154	140	122
138	148	139	168	160	120	151	121	146	135
156	149	135	148	138	133	150	132	153	140
148	135	146	158	137	153	142	155	114	147
133	127	132	155	134	138	139	131	142	135
145	134	143	128	147	137	142	119	139	131
145	141	150	152	125	147	144	131	162	138
146	121	143	130	141	150	128	142	150	107
123	143	140	140	147	130	141	126	151	149
140	139	140	125	147	129	149	142	154	151

需要全本请访问 www.ertongbook.com