医药统计学

科研设计与数据分析

新世纪出版社

前　言

《医药统计学》是将数理统计学的原理与方法应用于医药学的一门边缘学科。随着科技的迅速发展，医药各科应用统计学设计与处理日益广泛和深化，并推动了医药各科的发展。医药专业的学生及工作者应加强医药统计学的学习，本书正为其学习之需编著，并有如下特点。

1．本书按预防医学系本科生卫生统计学的教学大纲编著。在扼要介绍统计学基本原理的基础上，重点叙述统计方法的应用。着重于使读者学得懂，用得上，并能举一反三，触类旁通。

2．本书作者既经历临床诊疗与医药科研的实践，又担任卫生统计与生物统计的教学工作，因此，能较好地做到医药实践与统计方法的紧密结合。

3．将卫生统计与生物统计融合于一书。统计学原理与方法的共性方面统一编著，各自的专业统计方法独立设章。既便于由一个教研室承担这两门学科的教学工作，又利于学科间的渗透和读者对学习内容的联系，是一种有益的探索。

4．结合本省医药科研开展及读者对象的实际，在介绍基本统计方法的基础上，加强了科研设计、多元分析及医药业务统计的应用。因此，本书适用性广，针对性强，可供临床各科、预防医学、药学中的各级业务、科研、教育及管理等人员，多专业、多层次的需要。

5．为开发思维，帮助读者复习与练习，书末编印了300道各型多选题及43道练习题。

本书由宋士德主编。全书共分21章，宋士德编著第1、2、3、4、5、6、12、14、15、17、18及19章。杜琳编著第7、9、10及13章。陈勇编著第8、11、16、20及21章。最后由宋士德修订统稿。

本院陈少贤讲师参加了本书的部分筹编工作，周丹虹统编了统计实习题，陈倩、李燕芬担任大部分书写与部分计算工作，王英担任绘图工作，对上述教师们的辛勤劳动，谨此致谢。

由于时间仓促及作者的学识所限，书中遗误之处在所难免，恳请使用本书的师生及医药工作者们提供宝贵意见，以便再版时订正。

作　者

于广东医药学院

1988年7月

目　录

第一章　绪　论

1.1　医药统计学的性质

医药统计学是运用概率论和数理统计学（mathematical statistics）的原理与方法，结合医药学工作实际，研究数字资料的搜集、整理、分析和推论的一门应用学科。研究随机变量（random vuriable）的变异性，是医药实践与医药科学研究所必需的重要手段。

1.2　医药统计学的主要内容

1．统计学设计　在实施医药学实践或开展医药学科学研究之前，除作专业上考虑外，还必须从统计学角度进行调查设计、实验设计及临床试验设计。使之能科学地回答所研究的问题，并能用较少的人力、物力和时间取得更多的较可靠的资料。

2．样本指标的计算　根据观测资料的性质，运用恰当的数字模型和统计方法，计算观测值的有关指标，以描述调查或实验结果。如平均水平指标、变异指标、相对数指标及统计量的抽样误差（标准误）等。

3．总体指标的估计（推定）　用样本指标估计总体中相应的统计指标，也称参数估计，以推论总体指标的可能范围。

4．假设检验（检定）　根据资料的性质和所需解决的问题，建立检验假设，然后采用适当的检验方法，根据样本是否支持所作的假设，用求得的统计量观察值相当的概率，与所取检验水准比较，决定该假设予以拒受或不拒受。

5．医药学业务统计　医药学业务较多、常用的如人口统计、计划生育统计、疾病统计、死亡统计、半数效量计算及生物检定等。属专题或专业性统计，它是结合医药学专题运用一定的统计原理与方法所作的统计，有时需采用较特殊的数学模型与统计方法。

6．联系、分类、鉴别、监测及管理等研究　由于学科间的横向联系及其渗透，医药学的多因素性，研究方法日益深化。其中有研究某医药学诸现象间的联系。医药学诸样品或指标间的聚类。疾病计量鉴别诊断。防治方案与药物的筛选。对某些疾病的预报预测，流行病学监督，药品制造和医学检验的质量控制，以及卫生事业科学管理研究等。医药统计学为解决上述问题中的单变量或多变量分析，提供了必要的方法和手段。

1.3　医药统计工作步骤

医药统计工作一般分为设计、收集资料、整理资料及分析资料四个步骤。四者既有先后顺序、又是密切联系的不可分割的整体，任何一个环节发生缺陷，都会影响研究结果的正确性。

1．设计　根据研究目的作出周密的调查设计或实验设计，是关键的一步。

2．收集资料　根据研究设计的目的与要求，用较少的人力、物力、时间，及时收集

正确、完整的原始资料。资料常来源于三个方面。

(1) 统计报表 按国家规定的报表制度，由卫生机构定期逐级上报。但项目较少，填报者对项目填报不尽一致。

(2) 医疗机构日常工作记录与报告卡 如病历、医学检验记录、工作日志及传染病报告卡等。需防止错漏和重复。

(3) 专题调查或实验 为常用的一时性调查方法。常易达到研究目的，但需一定的有关条件。无论哪种资料，都应做到正确性与完整性。

3．整理资料 又称统计归纳。根据研究设计中整理分析计划的要求进行分组与汇总。分组常用划记、分卡等方法，得到变量值分组及其相应的频数，填入整理表。注意纵、横向核对。

4．分析资料 又称统计分析。包含指标的计算、统计图表的绘制及作出推断结论。要求计算正确，图表符合绘制原则，参数估计与假设检验合理，并结合专业作出恰如其分的结论。

1.4 资料类型

统计资料一般分为计量资料与计数资料两大类，等级资料则介于其间。各种资料按分析需要也可相互转换。但不同类型资料应采用不同的统计方法进行分析。

1．计量资料（measurement data） 对各观察单位用定量法即与某标准作比较所得的资料数据。其观测值由其真值和系统误差及随机误差等综合组成的近似值。提高精密度后为非整数，属连续性数据。如身高、体重、浓度、血压、药物剂量及血清中某物质的含量等。描述及推论计量资料的统计量，常用平均数、标准差、t检验、方差分析、相关与回归分析等。

2．计数资料（enumeration data） 将观察单位按某属性或类别分组，然后清点各组发生的频数所得的资料数据。只能取有限个或无限可数个数值。属离散型数据。如粪检虫卵结果分阳性与阴性两组，然后清点阳性组例数与阴性组例数。某人群血型分A、B、AB、O型四组，清点各血型组的人数等。各组有质的区别，不同质的观察单位不能归在同一组。分析计数资料常用率、构成比、u检验及 χ^2 检验等。

3．等级资料（ranked data） 将观察单位按某属性或类别分成有序数量级别组，然后清点各级别组发生的频数所得的资料。如尿蛋白分一、十、卄、卅、等后，清点各等级组例数。疗效分痊愈(1)、显效(2)、好转(3)及无效(4)，级别组后，清点各组例数。各等级组为不同量的属性，归级时不能混淆。分析等级资料常用率、构成比、χ^2 检验、秩和检验及等级相关等。

资料的转换 按分析的需要，计量、计数及等级资料间可相互转换。如上述尿蛋白测定，原属计量资料，将它转换成一、十、…、等级时，便成等级资料。上述粪检结果阳性和阴性可转换为1和0，将疗效取值1、2、3、4，这时原计数资料或等级资料就转换为计量资料。

1.5 数据的整理

1．尾数计算时的修整

若拟修整尾数，则采用4舍6入、5则按其前一位数用“奇进偶舍”法，较“4舍5入”更接近原值．如原数0.625，0.565，0.595，0.645，0.475，合计为2.905．用“奇进偶舍”法合计为2.900．若用4舍5入法合计为2.930．可见前者合计更接近原合计值．

2．整数尾数的简化

据对准确度的要求简化，如某人红血球数为4,754,635个／mm^3，若要求准确度以万为单位，则不满一万之余数，按“奇进偶舍”法简化为4,750,000个，或写为475万个．若以十万为单位，用上法简化为4,800,000个，或用科学记数法为4.8×10^6个．

3．小数后位数的取舍

据所需准确度而定，并非小数位愈多愈好，因实际度量数字后的估计数字，人员（估计）误差较大．如测得血压120.5mmHg，整数为度量数字，小数为估计数字．若准确到个位数，按上法修整为120．最大呼气中期流速若准确到0.1，则可将3.12（L／秒）修整为3.1．若准确度为0.01，则仍为3.12．

4．有效数字

决定数量大小的数字，称有效数字，含实际度量数字与估计数字，表达度量结果．如救41.2有效数字为三位．12.800的有效数字为五位，小数点后的零表示准确到0.001，其误差为0.0005，故其实际可能值为12.7995—12.8005．当观测得组限（class limit）为141～145cm时，其准确度以1cm为单位，误差为0.5cm，故其实际可能组界（class boundary）为140.5～145.5cm．0.00075（7.5×10^{-4}）的有效数字为二位，其前的零取决于采用的单位，用以定位，并非有效数字，如0.00075g可写为0.75mg或750μg．

整数位零的有效性，视原数的准确度而定，如红血球为4,750,000个／mm^3，若准确至个位，则末尾四个零都是有效数字．若准确至万为单位，则475万个为三位有效数，末尾四个零不是有效数字．

1.6 几个基本概念

1．总体与样本（population sample） 总体是指性质相同的研究对象中、所有观察单位某种变量值的集合．如研究某地1985年健康成人的肺活量，则研究对象是该地1985年健康成人，观察单位是其中每个人，变量值为肺活量值，该地1985年全部健康成人的肺活量值构成一个总体．同质基础是同一地区、同一年份、同为健康成人．这总体只包含有限个观察单位，称有限总体．有时总体是设想的，如某法对哮喘的疗效，这总体将包含设想用本法的所有哮喘患者，其观察单位由于未能全被检出或无法全给予相同处理，故称为无限总体或理论总体．

从总体中随机抽取部分观察单位，某种变量值的实测值构成样本．如上述总体中随机抽取200名健康成人的肺活量值构成一个样本．计算样本的有关统计量后，可用于估计总

体相应指标的可信区间。但这种推论是以随机化抽样为前提，以足够的样本个数为条件。

2．误差（error） 统计学上将测得值与真值之差，样本指标（统计量）与其相应的总体指标（参数）之差统称误差。按产生原因与性质主要分为三类。

（1） 系统误差（systematic error）在收集资料过程中，因仪器、试剂、技术、条件或设计等不合要求，造成观测值偏大或偏小，有时偏性呈一定的倾向性，可泛称为系统误差。系统误差影响原始资料的正确性，应力求排除。如已发生，要尽力查明原因，予以校正。但不能靠统计方法解决。

（2） 随机误差（random error）是排除系统误差后尚存的误差，由多种尚无法控制的偶然因素（chance）的影响所引起，是客观存在的、大量因素作用下的综合结果。其值常无固定的倾向。随机误差呈正态分布，故可用概率统计的方法作处理。

（3） 抽样误差（sampling error）从同一总体中随机抽取若干个样品数相等的样本，各样本统计量间有所不同，它同时反映了样本与总体间的差异。这种由于抽样而引起的误差称抽样误差（有时统称随机误差）。这是由于总体中各样品存在个体差异所致。生物变异性是普遍存在的客观现象，故抽样误差不可避免。但它有一定的规律可循，概率论可帮助我们对偶然性与必然性的机遇作出决策。

此外，还有如“过失误差”，必须加以避免。舍入误差与非均匀误差，应尽力减少。

3．概率p（机率、或然率）（Probability） 指某一事件A发生的可能性大小的一个度量。必然发生的事件的概率为1，不可能发生的事件的概率为0，一般地P在0～1之间。P愈接近1，表明发生A的可能性愈大，P愈接近0，表明发生A的可能性愈小。医药学研究中，习惯用$P \leqslant 0.05$或$P \leqslant 0.01$，作为事物间差异有显著意义或有非常显著意义的界值。

4．参数与统计量 参数常指总体指标，在一定条件下为一常数。统计量是指从总体中随机抽取的样本，所构造的随机变量或指标值。或曰几个变量值不包含未知参数的任一函数称为统计量。样本特征数都是统计量。

5．分布与数学期望（期望值）

分布即随机变量X，在各变量值（元素）x上频率（或频数）的排列。

数学期望是对随机变量进行长期大量观测、所得数值的平均值。用E（expected value）表示求期望值。

6．准确度与精密度 准确度是指观测值与其真值的接近程度，反映系统误差的大小、亦即资料的均值（$\bar{x}$）与其真值（总体均值 μ）间的离差（deviation），又称偏差（bias）。精密度是指多次重复观测值x_i与其$\bar{x}$的接近程度，其差值属于随机误差，常用标准差或变异系数表示。统计理论要求排除偏因，实验波动仅容许随机误差引起。首先保证准确度，才有精密度的实际价值，否则，即使多次重复使后者很高，也仅重复错误而无俾益。

第二章　集中趋势（central tendency）

2.1　频数分布（frequency distribution）

1．离散型数据

例2.1　每天测定10例慢阻肺患者肺功能，共测180天，肺功能异常例数组值（class value）的频数分布如表2.1．它扼要地描述了数据的频数分布．有时为了更直观地表达其分布，可绘制频数分布图（纵轴为频数），如图2.1．也可绘频率分布图（纵轴为频率），两者图形相同．

表2.1　每10例慢阻肺中肺功能异常例数的频数（率）分布表

组值r_i（例数）	频数 f_i	频率 p_i
0	0	0.000
1	0	0.000
2	5	0.028
3	12	0.067
4	25	0.139
5	34	0.189
6	42	0.233
7	40	0.222
8	11	0.061
9	7	0.039
10	4	0.022
合计	180	1.000

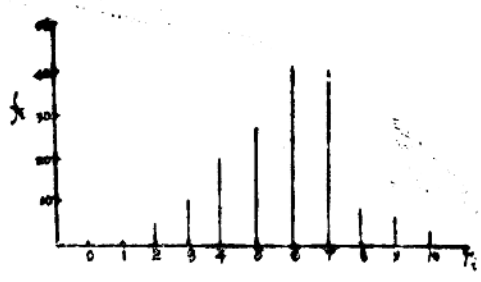

图2.1　频数分布图

2．连续型数据

例2.2　随机抽取100名男大学生测得最大呼气流速（PEFR，L／秒）如下表，编制频率分布表描述其频数分布．

表2.2　PEFR测定结果

6.50	8.10	6.95	3.51	4.25	7.65	6.90	5.62	6.10	7.15
6.55	6.85	4.35	5.35	7.60	8.48	7.85	4.75	7.00	7.44
8.90	3.82	5.70	7.58	5.45	6.15	4.00	8.45	6.35	6.30
9.50	5.10	7.55	6.20	4.58	7.38	8.30	6.35	7.44	5.60
6.58	6.75	4.62	6.36	5.14	5.64	5.85	5.50	4.45	5.25
6.60	8.15	6.40	7.48	6.85	7.88	6.98	7.20	7.38	7.28
8.95	6.80	5.55	5.60	7.85	6.45	3.95	7.25	6.42	6.45
6.68	7.90	7.35	7.22	5.78	5.60	5.46	6.32	5.75	5.20
8.55	6.10	6.15	6.32	6.38	4.55	6.45	4.85	4.95	5.80
6.80	8.25	6.00	5.45	6.25	3.10	5.95	7.45	5.90	5.15

频数分布表的编制　（1）求全距R（range）也称极差R = maxx − minx = 9.50 − 3.10 = 6.40.（2）求组距i（class interval）. 先定组（段）数k，以8－15组为宜. 或用下式估计

$$k = 1.87(n-1)^{\frac{2}{5}} \qquad (2.1)$$

式中n为样本含量（sample size）或样品数. 例2.2代入式2.1得k = 12组. i = R／k = 6.4／12 = 0.53 ≈ 0.5.（3）列表归纳：列出组限即该组的下限（lower limit）及上限（upper limit）（通常不标出）. 末组的上限应含最大值（maxx）. 用划记或分卡法归纳各组段的频数f，见表2.3. 通常用中值m（midvalue）代表该组段，m =（上限＋下限）／2. 但年龄组段的上限因已接近下组的下限，故中值为两组下限和的均值，如10－19岁，20－29岁，则第一组的m =（10＋20）／2 = 15（岁）. 并非m =（10＋19）／2 = 14.5（岁）.

表2.3　100名男大学生PEFR频数（率）分布表

组限	中值（m）	频数（f_i）	累计频数（f_c）	累计频率%（f_c%）
3.0～	3.25	1	1	1
3.5～	3.75	3	4	4
4.0～	4.24	4	8	8
4.5～	4.75	6	14	14
5.0～	5.25	9	23	23
5.5～	5.75	14	37	37
6.0～	6.25	20	57	57
6.5～	6.75	13	70	70
7.0～	7.25	12	82	82
7.5～	7.75	8	90	90
8.0～	8.25	6	96	96
8.5～	8.75	3	99	99
9.0～9.5	9.25	1	100	100
合计	—	100	—	—

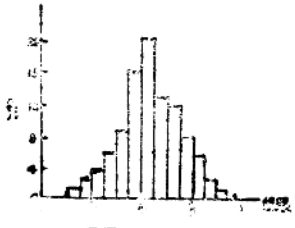

图22　PEFR频数直方图

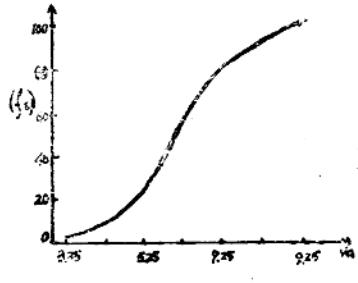

图23　PEFR累计频数图

还可用频数直方图（图2.2）或频率直方图，累计频数图（图2.3）或累计频率图更直观地描述其分布，或估计其频数及中值．后两者是将中值置于横轴，纵轴标累计频数或累计频率绘制而得．直方图是用各矩形面积代表各组段的频数（或频率）．累计频数图则可表达某值上、下各有多少人，或多少人在某中值以上或以下等，如PEFR在6.25（L／秒）以下约有57人，10人约在7.75以上等．

2.2 平均数（mean）

描述一组观察值的平均水平或称集中趋势，为常用的位置特征数．

1．算术均数$\bar{x}$（arithmetic mean） 简称均数或均值．适用于对称或近似对称分布，尤其正态分布资料．

（1） 直接法

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n}\ (\text{或简写为}\frac{\sum x_i}{n}\text{或}\frac{\sum x}{n}),\qquad (2.2)$$

式中x_i（i＝1，2，…，n）为第i个计量观察值（变量值）．计数变量值常称发生数，用r_i表示．Σ为求和符号．$\bar{x}$为样本均数，为对其总体均数μ的估计值．

例2.3 求表2.2第5列10人PEFR的均值．

本例为来自正态分布的连续型随机数据．代入式2.2 $\bar{x}=4.25+7.60+\cdots+6.25/10=6.013$（L／秒）

例2.4 随机抽样10人一年内感冒次数r_i依次为3、2、3、5、1、4、2、3、2、5，求其均数．

本例为离散型数据，设近似对称分布．仿式2.2，$\bar{r}=\sum_{i=1}^{n}r_i/n$．代入$\bar{x}=30/10=3$（次）．

（2） 加权法

连续型数据时
$$\bar{x}=\sum_{i=1}^{k}f_i m_i/n \qquad (2.3)$$

例2.5 求近似正态分布的计量资料，表2.3PEFR的均值．

代入式2.3得 $\bar{x}=\frac{1}{100}(1\times3.25+3\times3.75+\cdots+1\times9.25)=6.34$(L／秒)．

离散型数据时
$$\bar{r}=\sum_{i=1}^{k}f_i r_i/n \qquad (2.4)$$

例2.6 呈近似对称分布的计数资料表2.1，求平均每10例慢阻肺患者中肺功能异常的例数．代入式2.4得

$$\bar{r}=\frac{1}{180}(0\times0+0\times1+5\times2+\cdots+4\times10)=5.77\text{（例）}.$$

2．几何均数G（geometric mean） 适用于等比或近似等比级数资料及对数正态

分布资料。

（1） 直接法 为n个观察值乘积的n次方。

$$G=\sqrt[n]{x_1\cdot x_2\cdots x_n}\quad , \tag{2.5}$$

对数式

$$G=\lg^{-1}\left(\sum_{i=1}^{n}\lg x_i/n\right)\quad . \tag{2.6}$$

例2.7 6名儿童接种麻疹疫苗后抗体滴度为1∶4、1∶8、1∶16、1∶128、1∶128、1∶512，求平均滴度。代入式2.5或2.6得

$$G=\sqrt[6]{4\times8\times16\times128\times128\times512}=40.32,$$

或

$$G=\lg^{-1}\left(\frac{\lg4+\lg8+\lg16+\lg128+\lg128+\lg512}{6}\right)$$

$$=\lg^{-1}1.60=40.27.$$

故血凝抑制抗体平均滴度为1∶40 。

（2） 加权法

$$G=\lg^{-1}\left(\frac{\sum_{i=1}^{k}f_i\lg x_i}{\sum f}\right)\quad . \tag{2.7}$$

应用于相同x的频数较多时，将x_i转换为$\lg x_i$，再乘f_i，k组求和平均，其反对数即为G。通常列出频数表计算统计数 $\sum_{i=1}^{k}f_i\lg x_i$后，代入式2.7求得。

（3） 对数正态分布 变量值间相差数大的偏态分布数据，常可通过对数转换成对数正态分布，如表2.4。

表2.4 200例慢阻肺血清IgM（mg／100cc）频数表

IgM (x_i)	0～	5～	10～	15～	20～	25～	30～	35～	40～	45～	50～	55～60
频数 (f_i)	6	48	43	36	28	13	14	4	4	1	2	1

上表为正偏态，将原数据x转换为$x'=\lg x$，成对数正态分布。当x中有小值或零时，可用$x'=\lg(x+1)$或$x'=\lg(x+k)$，k为常数。负偏态数据的对数转换选用$x'=\lg(k-x)$。表2.4作对数转换后，经正态性D检验，其总体变量值服从对数正态分布，故其集中点宜用几何均数。

例2.8 求表2.4资料的均数。用式2.7，式中的x_i为上表内IgM组限的中值。

$$G=\lg^{-1}\left(\frac{6(\lg2.5)+48(\lg7.5)+\cdots+1(\lg57.5)}{200}\right)=14.1\ (\text{mg}/100\text{cc})\quad .$$

在描述集中趋势的指标中，众数M_0(mode) 是最简单的一种。它是一组资料中，频数（或频率）最多的观测值或该组限的中值。

3．中位数M及百分位数P_x（median & percentile） 百分位数P_x为位置指标，表示x%位次的变量值。中位数M为50%位次的变量值，即P_{50}，故M为一特定的百分位数。常用于偏态分布（无代换性正态）、未知分布或分布末端无界的资料，描述其集中点

及离散度，还可用于确定正常值范围．由于两端变量值未予计算，故代表性不及均数．当n增大时，M及P_x渐趋稳定．

（1） 直接法 变量值按大小（常从小到大）顺排后，位次居中的变量值即为M．

例2.9 某病患者7例，该病潜伏期天数（设未知分布）顺排为3、3、4、6、7、10、13、求M．

7例（n为奇数）居中的位次为4，其对应的变量值即M，M＝6（天）．

例2.10 承例2.9，若第8例潜伏期为16天，求M．

8例(n为偶数)居中的位次为4与5，其对应的变量值为6与7天，故$M=\frac{6+7}{2}=6.5$(天)．

（2） 频数表法 先编制频数表，并计算累计频数及累计频率%，然后按下式计算．

$$P_x=L+\frac{i}{f_x}(n\cdot x\%-f_c) \quad . \qquad (2.8)$$

式中P_x为x%位次的变量值．L、i及f_x为P_x所在组的下限、组距及频数．f_c为P_x前一组的累计频数．f_c对应于该组上限．取f_c%略大于P_x的组作有关计算．

表2.5 120名慢性混合型哮喘最大呼气中期流速（MMF，L／秒）频数表

组限(MMF)	1.2～	1.3～	1.4～	1.5～	1.6～	1.7～	1.8～	1.9～	2.0～	2.1～2.2
频数(f_i)	28	26	23	18	10	7	4	2	1	1
累计频数	28	54	77	95	105	112	116	118	119	120
累计频率%	23	45	64	79	87.5	93	96.7	98	99	100

上表呈右偏态（右尾长）分布，无适当的代换性正态，故以M及P_x描述其分布特征．

例2.11 承表2.5，计算M（即P_{50}）P_5 P_{25}及P_{75}．代入式2.8得

$$M=1.4+\frac{0.1}{23}(120\times\frac{50}{100}-54)=1.43(L／秒)．$$

$$P_5=1.2+\frac{0.1}{28}(120\times\frac{5}{100}-0)=1.22(L／秒)．$$

$$P_{25}=1.3+\frac{0.1}{26}(120\times\frac{25}{100}-28)=1.33(L／秒)．$$

$$P_{75}=1.5+\frac{0.1}{18}(120\times\frac{75}{100}-77)=1.57(L／秒)．$$

4．角均数（mean angle） 也称平均角．图形分布（circular distribution）无确切的零点，数值大小的意义是特定的．通过三角函数变换使原数据转换为线性．用$\bar{a}$表示其集中点．

（1） 时点资料

例2.12 10例混合型哮喘发作时间及数据转换如表2.6，求平均发作时间．