生物统计学

BIOSTATISTICS

杜荣骞

32

CHEP
高等教育出版社

Springer
施普林格出版社

图书在版编目(CIP)数据

生物统计学/杜荣骞. -北京:高等教育出版社;
海德堡:施普林格出版社,1999.7(2001 重印)
ISBN 7-04-006955-5

Ⅰ.生… Ⅱ.杜… Ⅲ.生物统计 Ⅳ.Q-332

中国版本图书馆 CIP 数据核字(1999)第 17638 号

生物统计学
杜荣骞

出版发行	高等教育出版社		
社　　址	北京市东城区沙滩后街 55 号	**邮政编码**	100009
电　　话	010-64054588	**传　　真**	010—64014048
网　　址	http://www.hep.edu.cn		
经　　销	新华书店北京发行所		
印　　刷	北京民族印刷厂		
开　　本	787×1092　1/16	**版　　次**	1999 年 7 月第 1 版
印　　张	18.5	**印　　次**	2001 年 2 月第 3 次印刷
字　　数	450 000	**定　　价**	22.50 元

©China Higher Education Press Beijing and Springer-Verlag Heidelberg 1999

版权所有　侵权必究

前　言

生物统计学是现代生物学研究不可缺少的工具。不论是传统学科还是现代分子生物学，时时刻刻都在与数字打交道。为了揭示生物体内在规律或生物与环境之间的关系，都离不开因素分析，特别是多元分析。生物统计学不仅在传统生物学、医学和农学中被广泛应用，而且在新兴的分子生物学研究中也发挥着重要作用。例如，绘制连锁图，特别是绘制人类基因连锁图时，制图函数的获得，Lod Score 的计算以及 DNA 序列同源性分析等都是建立在统计学基础上的。没有良好的统计学基础，这些工作只能知其然，而不能知其所以然，对于工作的深入开展是很不利的。因此，生物统计学已经成为每一个生物科学工作者必备的基础。

这本教材是在 1985 年版本的基础上，广泛征求各方面意见重新编写的。为配合生物学的迅速发展，在内容和编排上做了适当调整，删除了一些不常用的内容，增加了一些必要的基础，如方差分析中均方期望的推演等。近十几年来电脑在我国迅速普及，出现了大量的统计软件。许多过去望而却步的繁重计算工作，现在已变得轻而易举。利用统计软件代替繁重的手工计算，是生物统计学发展的必然趋势。SAS 系统是国际上公认的统计软件，它的包容量大，伸缩性强，在全球范围内被各行各业广泛采用，因此，本书编进了介绍 SAS 软件应用的章节，以满足读者的需要。书内的例题和习题除一部分是编者自己的工作外，很多是从书后所列参考资料中引用的，在这里对原著者深表谢意。为了使例题更具代表性，对其中有些数据做了适当调整，因此，书中例题和习题中的数据只供学习和巩固统计学知识使用，没有真正的科学意义，请广大读者切勿引用。

本书在编写过程中得到了各方面大力支持，四川大学刘天伦先生在内容编排上提出过宝贵建议，本校数学系沈世镒先生，计算机系涂奉生先生曾鼎力相助，生命科学学院王颖老师在资料整理和誊写上做了大量工作，在这里对以上各位先生表示真挚谢意。

在这里需要特别提出的是，美国 SAS 软件研究所上海办事处为本书的编写提供了 SAS 软件和多方支援，为促成本书起了很大作用。编者在这里对上海办事处的关心和支持表示衷心感谢。

编者在编写时虽已尽心竭力，但错误及不当之处仍所难免，敬希读者不吝指出，编者将不胜感谢。

编　者

于南开大学生命科学学院

1999 年 2 月

责任编辑　徐　可　吴雪梅
封面设计　于文燕
责任印制　陈伟光

目　录

试读结束：需要全本请在线购买：　www.ertongbook.com

第一章　统计数据的搜集与整理

§1.1　总体与样本

1.1.1　统计数据的不齐性

人类在生活、生产和科学研究中经常与数据打交道。在对特定的研究对象进行测量、记录并分析所得数据之后，你会发现，既使从同一类对象中所得到的数据仍有大有小、参差不齐。或者说，产生这些数据的个体间存在着广泛变异。

造成生物体变异的原因有很多，概括起来可以分为遗传因素，环境因素及发育噪音(development noise)。遗传因素的影响是显而易见的。就拿身高来说，子女身高直接受父母身高的影响，通常是父母高，子女也高；父母矮，子女也矮。环境因素表现在很多方面。仍以身高为例，包括：食量、蛋白质摄入量、营养成分平衡、维生素、微量元素的获得量、锻炼、劳动强度、睡眠时间、不良嗜好、修养、心理承受力等。我们会发现，既使在遗传与环境因素都得到控制的情况下，个体间仍然存在变异。例如，小麦纯系是经过多代自交得到的，遗传上已经纯合化，个体间遗传成分可以认为是均一的。将自交系的单株后代种植在生长条件都相同的环境中，例如种植在人工气候室中，使用电脑控制肥力、水分、光照、温度、通风等，既使这样，个体间仍存在变异。它们的株高、穗长、穗重、干物重等还会有一定的波动。这种波动的产生是由发育噪音引起的。或者说是由于在个体发育过程中的某些随机因素造成的。如果把影响生物变异的各种遗传因素、形形色色的环境因素以及种种随机因素自由组合起来，其组合数将是一个天文数字。不同个体的组合方式不同，由此造成了生物个体之间的广泛变异。由此可见，变异性是自然界存在的客观规律。

由于个体间的变异，给我们处理数据带来很多困难。例如，考察我国18岁男青年身高，若个体间没有变异，我们随便测量一个人就可以了。然而由于个体间存在着变异，为了测得18岁男青年身高，从理论上讲，应当把全国所有18岁男青年身高都测量一遍，用其平均数来代表身高数值。把所有18岁男青年身高都测量一遍是很难做到的。退一步讲，虽然很难做到，但只要投入足够的人力和财力，还是可以测量出这些数据的。如果要测量所有新生儿体重，则无论如何也拿不到全部数据。因为新生儿不断出生，要想收集到所有新生儿体重，就要不断测量，只要有新生儿出生，测量就不能停止。由此可见，测量全部对象既不现实也不可能。我们只能从全部研究对象中抽出一部分个体来，通过对这一部分个体的研究来推断全体的情况。这就出现了我们下面将要提出的两个概念：总体与样本。

1.1.2　总体与样本

统计学的核心问题是研究如何通过样本推断总体。因此，总体与样本是生物统计学中的两个最基本概念。

总体(population)是我们研究的全部对象。总体又分为**无限总体**(infinite population)和**有限总体**(finite population)。例如，我们要研究在某种条件下生长的小麦的株高，因为无法估计出在这种条件下生长的小麦的数量，可以设想这一总体是无限的。或者研究新生儿体重，因为新生儿是无止

境的,所以这一总体也可以设想是无限的。如果我们要调查一所学校今年新生的身高。这一总体则是有限的。生物统计学中所遇到的总体多数都是无限总体。构成总体的每个成员称为**个体**(individual)。

样本(sample)是总体的一部分,样本内包含的个体数目称为**样本含量**(sample size)。

1.1.3 抽样

从总体中获得样本的过程称为**抽样**(sampling)。抽样的目的,是希望通过对样本的研究,推断其总体。例如,希望由 100 株"三尺三"高粱的株高,推断在这种条件下生长的该品种的株高。这就要求样本应能在最大程度上代表总体的情况。为此,在从总体中抽取样本时,总体中的每一个个体被抽中的机会必须都一样,不能带有偏见。例如,在小麦育种工作中,我们常常希望得到矮秆品种。为了满足个人愿望,在抽样时便多抽矮秆的,这样得到的样本没有代表性,属于偏性抽样,不能代表总体的情况。我们需要的样本应该是一个总体的缩影。为了达到这个目的,就需要用**随机抽样**(random sampling)的方法获得样本。

随机抽样的方法很多,例如抽签、拈阄等。最好的方法是使用随机数字表(见附表 1)进行抽样。现举例说明怎样用随机数字表进行抽样。假设需要从包含 4 728 个个体的总体中,抽出一个含量为 20 的样本。因为个体总数 4 728 是一个四位数,所以总体中每一个个体的编号都应是四位数,即从 0001 号到 4 728 号。第一步,闭上眼睛用铅笔在随机数字表上任意点上一点,假若点到奇数上,就用第一页表;点到偶数上,就用第二页表。第二步,在选定的那一页上,再点一次,决定从哪个字开始。决定了起点以后,开始以四位数字为一节连续读下去,不用考虑数字间的间隙。可以正读、倒读、横向读、纵向读,也可以沿对角线方向读。选出小于等于 4 728 的数字,大于 4 728 的则舍弃,直到取满 20 个数为止。这 20 个数所对应的个体,即为我们选中的样本。

从一有限总体中抽样,可分为**放回式抽样**(sampling with replacement)和**非放回式抽样**(sampling without replacement)。所谓放回式抽样是指:从总体中抽出一个个体,记下它的特征后,放回总体中,再做第二次抽样。这种抽样方式可能会重复抽中某一个体。非放回式抽样是指:从总体中抽出个体后,不再放回。在上述的例子中,若保留重复的随机数字,则为放回式抽样;若舍弃重复的数字,则为非放回式抽样。对于无限总体来说,放回式抽样和非放回式抽样,实际上没有区别。

样本的含量越大越有代表性。但是,太大的样本研究起来是很困难的。因此,样本的含量必须合适。

§1.2 数据类型及频数(率)分布

1.2.1 连续型数据和离散型数据

统计学的最基本工作是收集数据。把原始数据收集上来之后,首先要对数据进行整理并分析这些数据的特性和变化规律。生物统计学中经常遇到的数据有两种类型,一种是连续型数据,另一种是离散型数据。

与某种标准做比较所得到的数据称为**连续型数据**(continuous data),又称为**度量数据**(measurement data)。例如,长度、时间、重量、OD 值、血压值等。这类数据通常是非整数。虽然有时记载的是整数,如身高的厘米数,但是当提高精确度后,总会出现小数。对连续型数据进行分析的方法,通常称为**变量的方法**(method of variable)。

由记录不同类别个体的数目所得到的数据,称为**离散型数据**(discrete data),又称为**计数数据**

(count data)。例如,某一类别动物的头数,具有某一特征的种子粒数,血液中不同类型的细胞数目等。所有这些数据全都是整数,而且不能再细分,也不能进一步提高它们的精确度。对离散型数据进行分析的方法,通常称为**属性的方法**(method of attribute)。

在判断数据的类型之后,就要进一步研究数据的变化规律。描述数据变化规律的最简单方法是将这些数据列成**频数表**(frequency table)或绘成**频数图**(frequency graph),根据频数分布进行研究。

1.2.2 频数(率)表和频数(率)图的编绘

离散型数据及连续型数据的频数表和频数图的编绘方法略有不同,下面各举一例说明。先看离散型数据频数(率)表和频数(率)图的编绘方法。

例 1.1 调查每天出生的 10 名新生儿中,体重超过 3 kg 的人数,共调查 120 d。每天的 10 名新生儿中,体重超过 3 kg 的人数,可能有 11 种情况:1 名也没有,有 1 名,有 2 名,…,10 名都是,如表 1-1 的第一列所示。这一列称为组值(class value)。表 1-1 的第 2 列所记载的是调查结果。

表 1-1 每 10 名新生儿中体重超过 3 kg 的人数的频数(率)表

组值 (体重超过 3 kg 的人数)	频数计算	频 数	频 率
0		0	0.000
1		0	0.000
2		0	0.000
3	一	1	0.008
4	丁	2	0.017
5	正正丁	12	0.100
6	正正正丌	19	0.158
7	正正正正正正正丌	39	0.325
8	正正正正正正丌	34	0.283
9	正正	10	0.083
10	下	3	0.025
总 计		120	0.999

如第 1 天调查的结果,有 6 名超过 3 kg 的,则在组值为 6 的一行做个记号,一般使用"正"字或"卌"号表示。全部调查完毕,累加各行结果,填入频数一栏。或者将各行的结果除以总数而得出频率。把频数或频率按超过 3 kg 的人数的顺序排列起来,便得到了**频数分布**(frequency distribution)或**频率分布**(percentage distribution)。频数表可以比较清楚地描述出数据变化规律。为了更直观地描述数据变化规律,还可以绘成频数图表示(图 1-1)。图 1-1 的横轴表示每 10 名新生儿中,体重超过 3 kg 的人数,纵轴表示每一组的频数。若将纵轴改为频率的话,则得到频率图。频率图与频数图的图形完全一样。

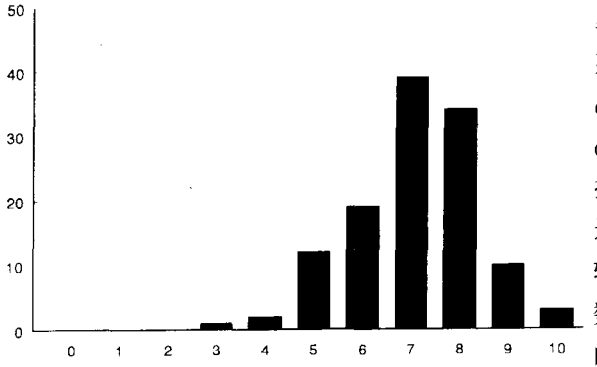

图 1-1 频数图

下面这个例子介绍了连续型数据频数(率)

表和频数(率)图的编绘方法。

表 1-2　"三尺三"株高测量结果

155	153	159	155	150	159	157	159	151	152
159	158	153	153	144	156	150	157	160	150
150	150	160	156	160	155	160	151	157	155
159	161	156	141	156	145	156	153	158	161
157	149	153	153	155	162	154	152	162	155
161	159	161	156	162	151	152	154	157	162
158	155	153	151	157	156	153	147	158	155
148	163	156	163	154	158	152	163	158	154
164	155	156	158	164	148	164	154	157	165
158	166	154	154	157	167	157	159	170	158

例 1.2　表 1-2 列出了某农场在做高粱"三尺三"提纯时所调查的 100 个数据。表中所列出的数据虽然是连续型的,但看上去好象是离散型数据。产生这种误解的原因,是由于株高的单位是"cm",在这种精确度下出现了许多高度相同的植株,当进一步提高精确度后,便很难再找到两个高度相同的植株了。从表 1-2 的原始数据中,除可以找出最大值是 170 cm,最小值是 141 cm 以及估计出它们的平均高度大约在 150~160 cm 之外,很难再看出什么规律来。但是,当我们将表1-2中的数据列成频数表之后,便可以比较清楚地看出这些数据的变化规律。高粱的株高是连续型数据,不是一个个孤立的值。因此,不能像例 1.1 那样制表。连续型数据频数表的制作过程如下:首先将数据分组,一般来说,100 个数据可以分成 8~10 组。根据极差 $R = \max x - \min x = 170 - 140 = 30$,分为 10 组比较合适,每一组的间距刚好是 3 cm。用比较简单的组间距分组,编制频数表比较方便。将分好的组填入表 1-3 的第 1 列。表 1-3 的第 1 列称为**组限**(class limit),组限是根据原始记录中的数值确定的。本实验是以 cm 为单位统计数值的,所以第一组的上限"143" cm 的实际值,可能在大于等于 142.5 cm,小于 143.5 cm 范围内。同样,第一组的下限"141" cm 的实际值,可能在大于等于 140.5 cm,小于 141.5 cm 范围内。因此,这一组的全部实际可能值是在 140.5~143.5 cm 范围内。140.5~143.5 这个范围称为**组界**(class boundary)。对于其他各组,同样可以定出相应的组界。

表 1-3　"三尺三"株高频数(率)表

组限/cm	组界/cm	中　值	频数计算	频　数	频　率
141~143	140.5~143.5	142	一	1	0.01
144~146	143.5~146.5	145	丅	2	0.02
147~149	146.5~149.5	148	𠂇	4	0.04
150~152	149.5~152.5	151	正正下	13	0.13
153~155	152.5~155.5	154	正正正正下	23	0.23
156~158	155.5~158.5	157	正正正正正下	28	0.28
159~161	158.5~161.5	160	正正正	15	0.15
162~164	161.5~164.5	163	正正	10	0.10
165~167	164.5~167.5	166	下	3	0.03
168~170	167.5~170.5	169	一	1	0.01
总　计				100	1.00

中值(midvalue)是每一组的两个组限的平均值,但是也有例外。例如,习惯上通常以"岁"为计算年龄的单位。假若有一组的组限是20～29岁,上限29岁包括这个人可能刚刚29岁,也可能即将进入30岁。所以这一组既包括20岁的,也包括29岁的,共有10个年龄级。因此,中值应是(20+30)÷2=25,而不是(20+29)÷2=24.5。类似这种情况,在计算中值时应特别注意。

频数计算一列,就是将表1-2中的数据"对号入座",最常用的方法是采用"唱票"的方式,一人读,一人填。最后将每一组的频数统计出来,记入频数栏,并计算出频率。在制成频数(率)表以后,连续型数据的频数(率)分布规律便清楚多了。

编制连续型数据的频数(率)表,一般需要以下各步:

①从原始数据表中找出最大值和最小值,并求出极差。

②决定划分的组数,分组数是由数据的多少决定的,在数据较少时,如50～100个数,可以分为7～10组。数据较多时,可分为15～20组。

③根据极差与决定划分的组数,确定组限。

④在频数表中列出全部组限、组界及中值。

⑤将原始数据表中数据,用唱票的方式填入频数表中,计算出各组的频数和频率。

表1-3以表格的形式,描述了高粱品种"三尺三"的株高频数分布。除此之外,还可以用频数图更直观地描述这一分布。下面是三种最常用的频数图。

1. 直方图

在横轴上标明各组的组界,纵轴标明频数。然后以每一组的组界为一个边,相应的频数为另一个边,作矩形,构成**直方图**(histogram)(图1-2)。若纵轴改为频率,则得到频率直方图。直方图又称组织图。频率直方图与频数直方图的图形完全一样。

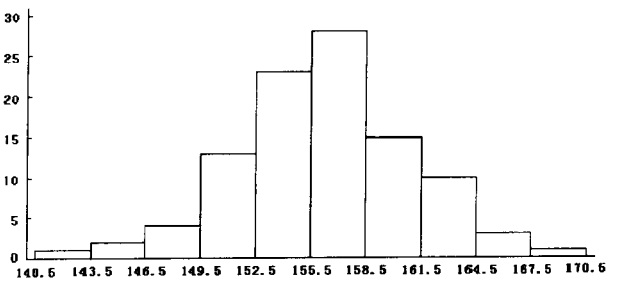

图1-2 "三尺三"株高直方图

2. 多边形图

在横轴上标出各组的中值,纵轴上标出频数(率),在坐标平面内,标出相应的每个点(以中值为横坐标,以该中值对应的频数(率)为纵坐标),用线段连接各点。最低一组非零频数的点,应该直接与相邻的零频数中值相连;最高一组非零频数点,亦应该与相邻的零频数中值点相连。最后得到一个**多边形图**(polygon)(图1-3)。

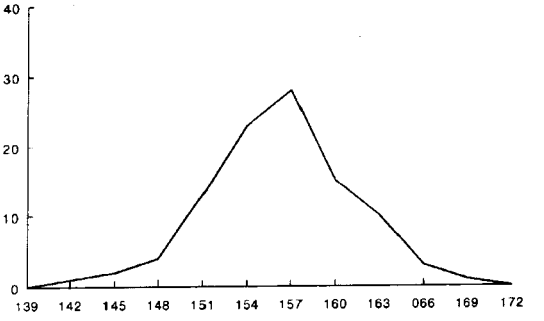

图1-3 "三尺三"株高多边形图

3. 累积频数图

经常使用的第三种频数图,称为**累积频数图**(cumulative frepuency graph)。作图法如下:首先根据表1-3制成**累积频数表**(表1-4)。在横轴上标出各组的中值,纵轴上标出累积频数(率)。在坐标平面内标出相应的点(以中值为横坐标,以该中值对应的累积频数(率)为纵坐标),连接各点,从而得到累积频数(率)图。

试读结束:需要全本请在线购买: www.ertongbook.com

图 1－4 就是根据表 1－4 所绘制的累积频数图。累积频数图与直方图和多边形图描述数据的方式不一样。累积频数图不能表达任何中值的频数，但可以表达某一中值以下的有多少株，以及一定数量的植株在哪一高度之下。例如，株高在 150 cm以下的大约有 15 株，最矮的 20 株大约在 151 cm以下。

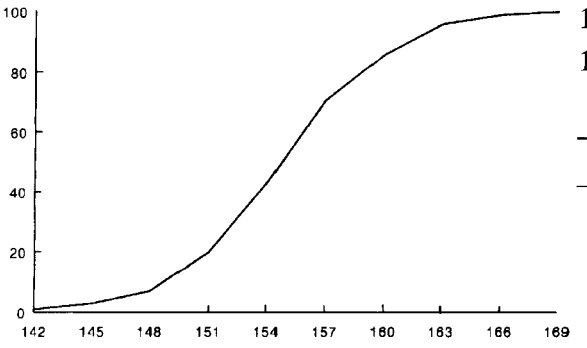

图 1－4　“三尺三”株高累积频数图

表 1－4　“三尺三”株高的累积频数表

中　值	累积频数	中　值	累积频数
142	1	157	71
145	3	160	86
148	7	163	96
151	20	166	99
154	43	169	100

1.2.3　研究频数(率)分布的意义

根据编绘的频数(率)表或频数(率)图，可以明显地看出数据的三个重要特征。

首先，根据频数(率)分布，可以看出数据的集中情况。一般来说，不论是离散型数据，还是连续型数据都有聚集于某一范围内的趋势，常常用**平均值**(average value)表示全部数据的集中点。使用最广泛的**平均**(average)是**算术平均数**(arithmetic mean)、**中位数**(median)和**众数**(mode)。算术平均数使用的最多，它是一群数据的重心所在。第二种平均是中位数，它是在累积频数图中$\frac{1}{2}$总频数位置上的数值。在例 1.2 中，总频数的中点在 50～51 之间。从累积分布图中可以找到，株高大约为155 cm。第三种平均是众数。离散型数据的众数是频数图中频数最高的组值。连续型数据的众数是频数图中频数最高的中值。

其次，从频数(率)表或频数(率)图中，可以直观地看出数据的变异情况：这群数据是集中在平均数附近，还是分散在平均数的两侧。如果数据大部分集中在平均数附近，远离平均数的两侧数据比较少，这样的数据是比较整齐的；若分布在平均数附近的数据与分布在远离平均数的两侧数据相差无几，这样的数据则是比较分散的。

第三，从频数(率)分布图中，还可以看出图形的形状。例如，有些分布从零频数开始平稳地上升，直到最高频数，然后平稳地下降直到零频数。结果得到一个对称的直方图或多边形图。而另一些分布，在上升阶段可能要经过很多步，到达最高频数后，突然下降；或者相反，上升很快，下降很慢。

此外，频数表或频数图还可以显示一些不规则的情况。例如，在一个分布中，出现一个或几个频数突然高出正常频数的情况，这是一种异常分布，可能是由于条件不一致，或由于度量时的失误造成的。当出现这样一些不规则情况时，需要认真研究，尽可能找出原因。

1.2.4　频数(率)分布的不恒定性

用随机抽样的方法，从某一总体中抽取两个含量相同的样本，分别编制出它们的频数分布表，并进行比较。这时会发现，虽然它们都是从同一总体中随机抽取的，但其频数分布并不完全相同

(表 1-5),有时差距还很大。若另外再取一个样本时,可能与前两个又不一样。出现这种现象并不奇怪,是抽样时经常遇到的现象。当用随机抽样方法获得样本时,由于偶然性,有时在一个样本中抽到的数值偏高,而另一个样本中数值偏低,使两个样本的频数分布出现不同。由于样本分布的不恒定性,当用样本去推断总体时,推断的结果也会有所不同。这就需要考察当用某一样本去推断总体时所得结果与真正总体之间有多大失误,或者说所得结果的可信度有多高。为了回答这一问题,首先要对总体分布有所了解。后面第二、三、四章的内容就是围绕总体展开的。

表 1-5 每 10 名行人中男性人数分布表

样本 1		样本 2	
男性人数	频数	男性人数	频数
0	1	0	0
1	2	1	1
2	9	2	6
3	17	3	18
4	27	4	25
5	46	5	40
6	29	6	30
7	12	7	20
8	4	8	9
9	3	9	1
10	0	10	0
总计	150	总计	150

§1.3 样本的几个特征数

频数表和频数图,只能定性地描述一组数据。对于生物统计学来说,这种描述远远不够。为了更客观地描述这些数据,需要借助于以下三种分析工具的帮助。它们是:数据集中点的度量——平均数,数据变异程度的度量——标准差和数据分布的对称程度及陡峭程度的度量——偏斜度和峭度。这些数字是描述样本频率分布特征的,称为样本数字特征或简称为**样本特征数**(sample characteristics)。

1.3.1 平均数

求平均的目的,是为了给出一个数,用这个数来描述由许多数组成的样本。如果样本中所有的数都是一样的,那么平均值就是这个数。若样本中的数不一样,则针对不同的目的可使用不同的平均。在生物统计学中,使用最多的是算术平均数,简称为**平均数**(mean)。样本算术平均数的符号是 $\bar{x}$,读做"x 杠"或"杠 x"。若用 $x_1, x_2, \cdots, x_n$ 表示组成样本的每一个数,则它们的算术平均数为,

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1}$$

或简写为

$$\bar{x} = \frac{\sum x}{n}$$

其中"∑"是求和的符号，n 为样本含量。$i=1$ 是相加数的下限，表示从 x_1 开始相加，"∑"符号上的 n 是相加数的上限，表示一直加到 x_n。

求和符号经常用到，关于它的三个简单的运算法则如下：

$$\sum_{i=1}^{n} c = nc \text{ 或 } \sum_{i=a}^{b} c = (b-a+1)c \quad (c \text{ 为常数}) \tag{1.2}$$

$$\sum_{i=a}^{b} cx_i = c\sum_{i=a}^{b} x_i \quad (c \text{ 为常数}) \tag{1.3}$$

$$\sum_{i=a}^{b} (x_i \pm y_i) = \sum_{i=a}^{b} x_i \pm \sum_{i=a}^{b} y_i \tag{1.4}$$

算术平均数有以下几个基本特性：

① 算术平均数的计算与样本内的每个值都有关，它的大小受每个值的影响。

② 若每个 x_i 都乘以相同的数 k，则平均数亦应乘以 k。

③ 若每个 x_i 都加上相同的数 A，则平均数亦应加上 A。

④ 如果 $\bar{x}_1$ 是 n_1 个数的平均数，$\bar{x}_2$ 是 n_2 个数的平均数，那么全部 n_1+n_2 个数的算术平均数是**加权平均数**(weighted mean)：

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \tag{1.5}$$

⑤ 平均数在理论上和在抽样实践中，还有更多的特性，主要表现在样本与总体的关系上。

第二种平均称为中位数。所谓中位数是指位于有序数列中点上的数。具体计算方法是：将样本中的 n 个数从小到大，或从大到小排列好，位于中间位置上的那个数即为中位数。n 为奇数时，很容易从数列中找出中间位置的数。但当 n 为偶数时，就需将中间位置上的两个数取其算术平均数作为中位数。

中位数有许多特性。算术平均数的第 2、第 3 条性质，中位数也具备，但却不具有第 1 条性质，中位数直接与样本含量 n 有关，而不是与具体的某个数有关。只要中间数值不改变，排列顺序不改变，两翼数字任意改变也不会影响中位数值。中位数不存在算术平均数的第 4 条特性，中位数也没有那么多抽样特性。

第三种平均称为众数。众数是具有最高频数的组值或中值。众数具有算术平均数的第 2、第 3 条特性，但不具有第 1、第 4 条特性。众数主要用来描述频数分布。例如，具有两个分开的高频数分布称为**双众数**(bimodal)。中位数和众数在生物统计学中很少使用。

1.3.2 平均数的计算方法

对于非频数资料可以直接使用(1.1)式进行计算。具体计算方法很简单，这里就不再举例了。

在计算离散型数据的频数资料时，可用下式：

$$\bar{x} = \frac{\sum_{i=1}^{k} (fx)_i}{N} \tag{1.6}$$

其中，x = 组值，f = 频数，N = 总频数，k = 组数，fx 代表 f 和 x 相乘。

下面计算例 1.1 的平均数。根据表 1-1 中的数据列成下表。

x	f	fx
0	0	0
1	0	0
2	0	0
3	1	3
4	2	8
5	12	60
6	19	114
7	39	273
8	34	272
9	10	90
10	3	30
总　计	120	850

由公式(1.6),得

$$\bar{x} = \frac{\sum_{i=1}^{k}(fx)_i}{N} = \frac{850}{120} = 7.08$$

$\bar{x}=7.08$ 的含义是:平均每 10 个新生儿中,大约有 7 个体重超过 3 kg。

计算连续型数据的频数资料时,假定频数表中的中值就是这一组的平均数(实际上只是近似于平均数),中值乘以频数就是这一组的和。再将各组之和相加,并除以总数即得出平均数。

$$\bar{x} = \frac{\sum_{i=1}^{k}(fm)_i}{n} \tag{1.7}$$

其中,m = 中值,f = 频数,n = 总数,k = 组数,fm 代表 f 和 m 相乘。

下面用频数分布法,计算表 1-2 中"三尺三"株高平均数。数据整理结果列在下表中。

m	f	fm
142	1	142
145	2	290
148	4	592
151	13	1 963
154	23	3 542
157	28	4 396
160	15	2 400
163	10	1 630
166	3	498
169	1	169
总　计	100	15 622

根据(1.7)式,得

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k}(fm)_i = \frac{1}{100}\sum_{i=1}^{10}(fm)_i$$

$$= \frac{15\,622}{100} = 156.22$$

1.3.3 标准差

对于数据的变异程度,经常使用的度量方法有三种,它们是:**范围**(range)或称为**极差**、**平均离差**(mean deviation)和**标准离差**(standard deviation)或称为**标准差**。其中最重要的是标准差。

范围是一组数据中的最大值与最小值的差,

$$\text{范围} = \max x - \min x \tag{1.8}$$

例如,有 5 个数:96.4、96.6、97.2、97.4、97.8(mL)。它们的范围(R)

$$R = 97.8 - 96.4 = 1.4 \text{ mL}$$

范围给出的是,在含量为 n 的样本中,两个极端值之间的差。这种表达变异的方法最简单,但也最容易受远离数群的一个数的影响。虽然如此,范围仍然是一个很有用的度量方法,特别是对于一个较小的样本。很难用范围解释一个个别的数与平均数之间的关系。我们常常希望知道一组数中的某个数,是否比平均数低得特别多,或仅仅是一个一般的量,或者其他什么程度。因此我们就需要了解这一组数据与它们的平均数之间的标准的,或平均的不符合程度。一般的做法是,首先求出**离均差**(deviation from average),即求出每个数与它们平均数之间的离差,然后以某种方式表示这种离差的平均值。让我们先看一看表 1-6。

表 1-6 度量离均差的几种方式

x/mL	离均差$(x-\bar{x})$/mL	$\lvert x-\bar{x}\rvert$/mL	$(x-\bar{x})^2$/mL^2
96.6	-0.48	0.48	0.2304
97.2	+0.12	0.12	0.0144
96.4	-0.68	0.68	0.4624
97.4	+0.32	0.32	0.1024
97.8	+0.72	0.72	0.5184
$\bar{x}=97.08$	$\sum=0$	$\sum=2.32$	$\sum=1.3280$

假定我们希望求出离均差 $x-\bar{x}$ 的算术平均数,从表中的最末一行可以看出,离均差的和等于零,其平均数当然也等于零。因此,求离均差的算术平均数是没有意义的,我们必须找出另外的途径来解决这一问题。

一种解决的办法是,求离均差绝对值的和,然后用 n 去除,而得出平均离差(MD)。

$$\text{MD} = \frac{\sum \lvert x - \bar{x} \rvert}{n} \tag{1.9}$$

如表 1-6 中数据的

$$\text{MD} = \frac{2.32}{5} = 0.464 \text{ mL}$$

在表 1-6 的 5 个离均差绝对值中,有 3 个超过该值,有 2 个低于该值。

另一种更常用的解决办法是求样本的标准离差(或称标准差)。将所有的离均差都平方(表 1-6 中的最后一列),然后相加,所得到的和称为**离差平方和**(sum of square of deviations)。习惯上是用样本含量 n 除离差平方和而得到一个平均数。但按照统计学理论,这里不用 n 去平均而用 $n-1$ 去平均(见 **4.1.1**)。当样本很大时,用 n 与 $n-1$ 去平均的差别就不大了,这时也可以用 n。

除得的商称为**样本方差**(sample variance),用符号 s^2 表示。

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{1.10}$$

表 1-6 中所列数据的样本方差

$$s^2 = \frac{1.3280}{5-1} = 0.332\ \text{mL}^2$$

方差 s^2 是离均差平方的平均数。虽然方差在实际应用中用得最广泛,但因它的单位是原始数据单位的平方,所以它还不能直接地指出某个数 x 与平均数之间的偏离究竟达到什么程度。为此,采用标准差 s 做标准衡量 x 与平均数之间的离散程度。标准差有时也记为 SD。

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1.11}$$

或

$$s = \sqrt{s^2}$$

表 1-6 中所列数据的样本标准差为:

$$s = \sqrt{\frac{1.3280}{5-1}} = \sqrt{0.322} = 0.576\ \text{mL}$$

现在的单位与原始数据的单位一致了,于是我们可以用 s 作为标准来度量各个离均差。表 1-6 的 5 个离均差绝对值中有 2 个大于 s,3 个小于 s。

现在我们有两种度量 x 与 $\bar{x}$ 之间平均离散程度的标准,$s = 0.576$ mL,MD = 0.464 mL,前者比后者要大一些,这种情况是比较普遍的。它们是两种不同的度量方法,对问题的解释也不一样。在一个大的样本中,如果数据分布曲线是平滑的且对称的,那么大约有 57% 的数据落在平均数 ±1 个平均离差范围内,而大约有 68% 的数据落在平均数 ± s 范围内。

总之,衡量数据离散程度时,三种方法都可以使用。用抽样理论可以证明:用标准差估计总体离散程度最可靠,平均离差次之。在样本含量 $n > 2$ 时,范围是最不可靠的。因此,我们经常用的是标准差。但是当样本较小(如只有 4~5 个数)而且变异又不是很大时,为了简化起见,也可以用范围。

1.3.4 标准差的计算方法

1. 非频数资料的计算方法

公式(1.11)给出了标准差的一般公式。用这个公式计算时,首先要计算出平均数,再求离均差,当平均数是一个近似值时,会有许多位小数,计算很繁琐,也影响结果的准确性。因此,可将(1.11)式变为另一种形式(见习题 1.4)

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}} \tag{1.12}$$

也可以略去下标,简单地表示为:

试读结束:需要全本请在线购买: www.ertongbook.com