

生物统计附试验设计

全国高等农业院校试用教材

生物统计附试验设计

贵州农学院主编

畜牧专业用

农业出版社

主 编：贵州农学院 俞渭江
编写人：内蒙古农牧学院 曹守漠
吉林农业大学 王滋润
沈阳农学院 杜荣臻
山西农学院 谢文采
四川农学院 明道绪
江苏农学院 于汉周

全国高等农业院校试用教材

生物统计附试验设计

贵州农学院主编

农业出版社出版（北京朝内大街130号）
新华书店北京发行所发行 云南新华印刷厂印刷

787×1092 毫米 16 开本 15.25印张 240千字
1980年10月第1版· 1980年10月昆明第1次印刷
印数 1—11,500册
统一书号 13144·221 定价 1.60 元

前 言

本书是为高等农业院校畜牧专业而编写的。

根据畜牧专业对本门课程内容的要求，以及适应加快实现我国农业现代化和不断发展畜牧业新技术的需要，在内容方面着重于基本理论、基本技能和基本方法的讲授，力求由浅入深，循序渐进。并在各章后附有习题，作为学生进行课内课外作业练习之用。它对熟悉和掌握本门课程是必要的。

本教材为了达到、保持课程系统性以及加强基础理论的要求，对概率知识与理论分布作为专门一章，各院校可根据本专业《高等数学》的讲授情况，酌情增减。考虑到教材是教学中的一项基本建设，既需满足目前需要，又应兼顾今后发展，所以在教材中，还安排了一些选学内容，如：1.差异显著性检验的非参数法，2.方差分析中的基本假定和数据变换及同质性检验，3.复相关等节，并标以“*”号，以资区别，各校在进行教学时，可视具体情况，自行处理。由于本门课程系数理统计方法在畜牧科学中的应用，考虑到学生已有一定的高等数学基础。所以在本教材中对某些公式的数学推导和原理作了一定的介绍，以利于启发学生的独立思考和培养分析问题的能力。

本书引用了一些国内、外文献和资料，初稿承兄弟院校生物统计课教师审查讨论，提出修改意见，修改稿形成后，又蒙赵仁镛教授审校，并提出宝贵意见，为此深表感谢！

由于编写时间短促和编写人员业务水平有限，错误和欠妥之处，欢迎同志们提出批评指正。

编 者

一九七九年六月

目 录

前 言	
第一章 概论	1
第一节 生物统计与试验设计的概念	1
第二节 本课程的主要内容	1
第三节 本学科发展概况	2
第四节 常用统计术语	4
习 题	5
第二章 数据的整理	6
第一节 数量资料的分类	7
第二节 原始数据的检查与核对	8
第三节 数据的分组	9
第四节 次数分布表与次数分布图	11
习 题	13
第三章 平均数、标准差及变异系数	14
第一节 平均数	14
第二节 标准差	19
第三节 变异系数	25
习 题	26
第四章 概率与理论分布	27
第一节 事件与概率	27
第二节 概率相加、相乘定理	28
第三节 理论分布	31
习 题	42
第五章 均数差异显著性检验—— t 检验	43
第一节 显著性检验的意义	43
第二节 显著性检验的基本原理	44
第三节 样本均数与总体均数差异显著性检验—— t 检验	51
第四节 两样本均数差异显著性检验—— t 检验	52
第五节 两个百分数资料差异显著性检验	56
第六节 平均数的置信区间	57
*第七节 非参数检验法	58
习 题	62

第六章 χ^2 (卡方) 检验	64
第一节 χ^2 检验的意义与原理	64
第二节 适合性检验	67
第三节 独立性检验	70
习 题	75
第七章 方差分析	76
第一节 方差分析的意义	76
第二节 方差分析的基本原理	77
第三节 方差分析的基本方法	87
第四节 方差组分与组内相关	104
*第五节 方差分析的基本假定和数据转换	108
*第六节 方差同质性的检验	110
习 题	111
第八章 直线相关和直线回归	114
第一节 直线相关	114
第二节 直线回归	121
第三节 相关与回归的关系	127
第四节 直线相关及直线回归的应用	130
习 题	131
第九章 复回归与复相关	132
第一节 偏回归与复回归	132
*第二节 偏相关	138
*第三节 复相关	140
习 题	143
第十章 协方差分析	144
第一节 协方差分析的意义	144
第二节 协方差分析的作用	144
第三节 协方差分析的方法	145
习 题	154
第十一章 抽样的原理和方法	155
第一节 抽样的基本概念	155
第二节 抽样误差的估计	155
第三节 样本的含量	157
第四节 试验家畜头数的确定	158
第五节 抽样的方法	161
第六节 抽样方案的拟订	162
习 题	163
第十二章 试验设计	164

目 录

第一节	畜牧试验的任务与要求	164
第二节	畜牧试验的种类	165
第三节	试验计划、方案的拟定	166
第四节	试验设计方法	168
第五节	正交试验设计	182
	习 题	194
附录 1	通用202型手电两用计算机使用说明	196
附录 2	希腊字母名称	201
附录 3	两个样本平均数的差 $\bar{x}_1 - \bar{x}_2$ 抽样分布的推证	202
附表 1a	正态分布表	204
附表 1b	K_α 值表	205
附表 2	正态分布的密度函数表	206
附表 3	t 值表 (两尾)	207
附表 4	符号检验表	208
附表 5	秩和检验表	209
附表 6	χ^2 值表 (一尾)	211
附表 7	5% 及 1% F 值表 (一尾)	212
附表 8a	5% q 值表 (两尾)	218
附表 8b	1% q 值表 (两尾)	219
附表 9	百分数反正弦 ($\sin^{-1}\sqrt{n}$) 转换表	220
附表 10	SSR 表 (邓肯氏新复极差值表)	223
附表 11	r 及 R 的显著数值表	225
附表 12	随机数字表	226
附表 13	常用正交表	228
附表 14	数值 1—400 的平方及平方根表	234

注：* 符号表示为选学内容。

第一章 概 论

第一节 生物统计与试验设计的概念

生物统计是数理统计在生物科学中的应用。它是用数理统计的原理和方法来分析和解释生物界各种现象和数量资料的科学。在生物科学研究的过程中，逐渐与数学的发展结合起来形成了本门学科。因此，它是应用数学的一个分枝，属于生物数学的范畴。

人们在从事生物科学研究时，往往是通过某事物的一部分（样本），来估计事物全体（总体）的特征的，目的是为了由样本推断总体，从特殊推导一般，对所研究的总体，作出合乎逻辑的推论，得到对客观事物本质的和规律性的认识。例如，在畜牧业生产中，中国黑白花奶牛305天产奶量资料之庞大，几乎不可能全部观察记载，只能采用抽样的方法，计算样本的统计量。用样本的统计量，来估计总体的参数。

在生物科学研究中，我们所期望知道的是总体，而不是样本，可是在具体实践中，我们所得到的却是样本而不是总体。因此，生物统计在某种意义上来说，它就是研究生物科学中以样本来推断总体的一门学科。

试验设计是指试验研究工作未进行前应用生物统计原理，来制订合理的试验方案，以及正确地选择试验动物、分组与对比，使我们可以利用较少的人力、物力和时间，获得多而可靠的信息，得出科学的结论。

第二节 本课程的主要内容

开设本门课程的目的，是为了给畜牧专业的专业基础课和专业课，特别是为遗传学、家畜育种学和家畜饲养学打好基础。

畜牧生产和科学研究中所常见的变异量，都是服从于一定的理论分布的。所谓理论分布是指某一性状中，每一部分个体出现的次数占总体的百分率。就以奶牛产奶量来讲，它是服从于两头小，中间大，即最高、最低产奶量的奶牛均较少，中等产奶量的奶牛较多的一种正态分布。它的特征之一，是呈现连续性变异。例如，家畜的体重，奶牛的产奶量，猪的膘厚等均属此类。但也有一些性状是属于非连续性变异，呈现离散型变异，如家畜的性比例，非公即母；在药物试验的治愈或无效，它是服从于二项分布的，而研究这些分布都要应用统计分析的原理。

从统计方法的主要内容来讲，大致可分为以下三个方面。

一、资料之整理及其基本分析

生物统计的基本特点，是以样本推断总体。如果在大样本情况下，就需要通过统计资料初步整理之后，再从资料中计算出三个主要的统计量，即平均数、标准差及标准

误。用以来根据样本推断总体的特征，一是资料的集中性，以平均数来表示；二是资料的离中性，以标准差来表示；三是衡量平均数的可靠性——用标准误大小来表示。

二、显著性检验

(一) 平均数间差异的比较 在进行畜牧生产或科学研究工作中，经常会遇到两组或两组以上数据的平均数。那么，它们之间进行比较，就有一个有无显著差异的问题，或者说孰优孰劣的问题。平均数之间的比较，要通过一定的统计方法，并且要注意抽样是否合理，否则它们将没有比较的基础。例如，有两批孵化期不同的两个月龄雏鸡的体重，甲组（3月8日出壳）共61只，平均体重为269.3克，乙组（4月15日出壳）共73只，平均体重为275.6克，试问这两批不同孵化时间，雏鸡平均体重之间的差异是由于不同孵化时期所造成，还是由于其他偶然因素所引起？这就需要应用均数差异显著性检验的统计分析方法，才能作出较可靠的判断，不致受某些偶然性因素所掩盖。

(二) χ^2 检验——属性的统计分析 有许多性状不能直接用测量的方法来加以衡量，一般称之为属性性状。例如，在猪的杂交试验中，毛色的黑白、性别中的公与母，以及药物试验的治愈或无效。均可以应用属性统计的方法，通过对具有相同属性的计数来分析。

(三) 方差分析 方差分析又名变量分析。目的是为了进行多个平均数间的比较。它的原理是应用数学方法，把试验中总变异剖分为由不同变异原因所形成的各种变异并进行显著性检验与多重比较。

三、相关与回归

研究两个变量之间相互关系的密切程度，称为相关。以相关系数来表示。例如，黄牛胸围与体重存在着一定程度的相关。胸围越大，其体重可能越大。我们就用相关系数来表示两者间的相关程度。

回归是指两个或两个以上的变量存在着从属关系，即一个变量（ x ）变化时，引起另一变量（ y ）的相应变化，它们的从属关系可以用回归分析的方法进行研究。进行回归分析，可以根据实际数据建立回归方程式，用以对某些指标进行预测和预报。例如，用胸围，体长来估计体重。或根据亲代生产力预测、预报后代的育种值等。

本门课程的另一主要任务是讨论试验设计的原理和方法。所谓试验设计，主要是指如何选择试验动物，进行合理的分组等，其目的是为了尽量减少和控制试验误差，并对试验误差作出无偏的估计。主要的设计方法有：1.完全随机设计；2.配对设计；3.随机单位组设计；4.拉丁方设计；5.正交设计等。为了使试验结果成为有用而可靠的科学资料，在开始试验之前，认真地进行试验设计是非常必要的。

第三节 本学科发展概况

生物统计是一门新兴的学科。十九世纪末叶，1870年英国遗传学家高尔顿（F. Galton）通过研究人类体高的遗传，认为子女的体高与父母的体高有着直接关系，发现子

女的体高与他们父母平均的体高有回归的趋向。这就是在数理统计中，回归这个术语的由来，因而后人推崇他为生物统计的创始人。

正态分布对研究生物统计的理论是十分重要的，它早在1733年被迪—摩弗来 (De-Moivre) 发现，而被后来的高斯 (Gauss) 所完成，因此，有人称它为高斯分布。

1899年皮尔逊 (K. Pearson) 提出了一个测量实际数与预计数 (或理论数) 之间的偏离度的指数卡方 (χ^2)，它在属性统计分析有着广泛的应用。例如，在遗传上孟德尔豌豆杂交试验，高豌豆品种与矮豌豆品种杂交之后，子一代均为高豌豆，高株这个性状属于显性，而矮株属于隐性，当子一代自交时，它的后代数预计是应该有高3:矮1，但实际后代数，是否符合3:1，需用 χ^2 进行检验。这种方法在畜牧兽医科研方面至今还在广泛应用着。

古斯特 (W.S. Gosset, 1876—1937) 是K. 皮尔逊的学生，他对样本标准差等分布作了不少研究工作，并且于1908年用“学生氏”的笔名将《t检验》发表于《生物统计学报》。t检验已成为当代生物统计工作基本工具之一。

1923年英国的费雪尔 (R.A. Fisher) 第一个把变异来源不同的均方 (S^2) 比值称为F值，当F值大于理论上5%概率水准的F值时，该项变异来源的必然性效应就从偶然性变量中分析了出来。这个分析方法，被称为方差分析。在生物统计中，方差分析应用是很广的。特别是在他发表了《试验研究工作中的统计方法》的专著后，对推动和促进农业科学、生物学和遗传学的研究和发展，起了一定的奠基作用。

纳耶曼 (J. Neyman) 和E.S. 皮尔逊分别在1936年和1938年提出一种统计假设检验学说，对促进理论研究以及对试验研究作出结论具有实用价值。

我国早在三十年代，生物统计与田间试验已成为农学系必修课程，最早出版的有王绶编著的《实用生物统计法》(1935年出版，1953年再版)。在畜牧方面应用则比较迟。四十年代在遗传学中，作为研究变异的统计方法。在饲养学中则进行饲养试验设计与分析。尔后，在兽医方面，药物试验也开始应用。进入七十年代以来，由于人工授精及液氮低温冷冻精液新技术的开展，推动和促进了乳牛育种科学研究。例如，公牛育种值的估计，母牛综合指数的计算，以及各类遗传参数：遗传力 (h^2)、重复力 (r_e)、遗传相关 (r_{mv}) 的计算，都需要应用生物统计中相关与回归，和系统分组的方差分析，从而使生物统计在畜牧业上的应用日益广泛起来。

试验设计是在本世纪初提出来的，三十年代应用于农业科学实验，因此“田间试验技术”成为一门专门学科，以后扩大应用于生物学、医学和工业上，也发展应用到畜牧科学上来，特别以牧草试验更为重要。近年来又提出了安排和分析多因素试验的正交试验法。这个方法在工农业方面的试验研究中，也相继获得了可喜的成果。最近，回归分析方法在工农业生产和科学实验中大量应用，如在进行数据处理时，寻求经验公式，探索新工艺、新配方，以及某些性状的预测预报，都积累了不少新经验，获得了许多新成果。特别是正交、回归试验设计是一个新的研究方向。各种类型电子计算机的发明使用，使运算技术出现新的跃进，变得日益精确和迅速。在畜牧科学现代化的进程中，应用统计方法以及先进的试验设计，来进行分析、研究，在我们今后工作中将越来越显得重要。

第四节 常用统计术语

一、总体与样本

总体是指研究对象的全体，它是客观存在的，但是我们无法一一调查得到，因而只能采用抽样的方法，抽取一定数量的个体，作为统计的依据。被抽出的若干个体称为样本。从样本计算出来的统计量，如平均数和标准差，来对这个总体在一定可靠程度上作出推断。在畜牧科学研究中所取样本，一般以样本中的个体数在30个以下为小样本，30个以上为大样本。

二、随机抽样

抽样是为了通过所抽得的样本来推断总体。那么，如果专抽产量高的，或者把容易抽到的个体作为样本，那就很难作出无偏估计。为了避免主观和偏见，在统计学中，采用随机抽样方法。所谓随机，就是每一个个体被抽作为样本的机会是相等的。拈阄就是一种随机抽样的方式。以后在抽样一章中还要专门讨论。

三、变数与变异数列

统计学上把每一个观察值称为变数。把所有变数整理成为一个由小到大的序列，称为变异数列。

四、参数与统计量

从总体中计算所得的数值，如总体平均数，总体标准差称为参数，而从样本中计算所得的数值称为统计量，它是总体参数的估计值。

五、机误与错误

由样本推断总体，往往不可避免地会产生一部分误差，这一部分误差是由各种无法控制的随机因素所引起，我们把它叫做机误或试验误差。另一种情况是由于粗心大意所引起，如抄错数字，算错结果，这些叫做错误，这种错误在工作中只要我们认真、细心是完全可以避免的。

六、准确性与精确性

统计工作是用样本的统计数来推断总体的参数。我们用统计数接近参数真值的程度，来衡量统计数“准确性”的高低。用样本中各个变数间变异程度的大小，来衡量该样本“精确性”的高低。因此，准确性就不等于精确性。准确性是说明测定值对真值的符合程度大小，而精确性却是多次测定值的变异程度。

生物统计是建立在生物学和统计学（或数学）两个基础之上的。如果生物学本身的理论建立在不充分的基础上，即使再准确地计算也毫无意义；反之，正确的理论也会由

于不精确的计算而导致错误的结论。可靠的判断方法是通过实践来检验的。因此，在畜牧科学研究中，在做出结论之后，也还必须再在生产上加以验证才行。同样，我们也不能把生物统计方法，生搬硬套，轻率地应用到社会研究方面，如果轻率的套用，往往会得出错误的结论。

习 题

1. 生物统计在本专业上有哪些用处？
2. 试举例说明常数与变数，参数与统计数，总体与样本的区别。
3. 练习计算机的使用方法。

第二章 数据的整理

从畜牧生产和科学试验中，取得的大量原始数据，是在某种具体条件下，对某种具体现象的观察结果，这些数据是一些零星的、孤立的现象，如要从其中揭露出它们的内部联系和规律性，必须先通过整理和分析。

数据的整理分析，就是要把大量复杂的数据进行整理归类，使其系统化，便于统计分析，从而得出正确的科学结论。

数据整理的方法，首先就是按照一定的标志，把记载数据分门别类地分成若干部分，把同一现象、同一类型的数据进行合并，使它们与其他现象、其他类型区别开来。另外，在数据整理时，要注意数据的完整性、真实性和准确性。对原始数据要细致地检查与核对，对那些个别极大和极小的数值要反复核实，力求确实可靠。数据整理的作用，就是按不同的标志把数据的特征反映出来，以便于进一步运用各种统计方法进行计算，来研究它们的规律性和相互关系。

表 2—1 70头成年母猪的繁殖力* (单位：头)

7	10	12	10	11	13	10
8	12	10	9	11	10	7
11	11	10	11	13	10	10
14	10	11	12	11	9	13
10	10	8	10	14	11	12
12	11	10	12	13	11	12
11	9	8	9	10	8	13
10	12	10	9	11	9	10
10	8	11	11	13	9	11
9	10	13	10	11	11	9

* 母猪的繁殖力指一窝产仔数。

原始数据的整理，其结果需要用数字来表明，这时，可将整理的数据制成表格。如样本较小可制成“依次表”，即将变数按数值的大小依次排列起来，形成一个由小到大的数字表，称为“依次表”。

例如表 2—1 中，70头成年母猪的繁殖力的记录，在未加整理以前只是一堆数字，看不出资料的任何意义。

如将表 2—1 进行整理可制成一个依次表(表 2—2)。

表 2—2 70头成年母猪的繁殖力依次表(单位：头)

7	7	8	8	8	8	8
9	9	9	9	9	9	9
9	9	10	10	10	10	10
10	10	10	10	10	10	10
10	10	10	10	10	10	10
10	11	11	11	11	11	11
11	11	11	11	11	11	11
11	11	11	11	12	12	12
12	12	12	12	12	13	13
13	13	13	13	13	14	14

从依次表中，可以看出母猪繁殖力的变异情况；即产仔数最高为14头，最低为7头。变异范围为7—14头。然后制成次数分布表，如表 2—3。由次数分布表还可以绘成次数分布图。当数据中变数较多时，制成“依次表”则较麻烦，可不作“依次表”。

表 2—3 70头成年母猪繁殖力的次数分布表

产仔头数 (x)	母猪头数 (划线计数)	次数 (f) 频数
7	//	2
8	///	5
9	///	9
10	///	20
11	///	17
12	///	8
13	///	7
14	//	2
总 和		70

第一节 数量资料的分类

资料的分类是统计归纳的基础，若不进行分类，大量的原始资料就不能系统化、规格化，只有根据科学原理来分类，才能使资料正确地反映出事物的本质和规律。

在调查或试验中，所观察、测量的数据必须加以记载。记载的方式，按观察或测量对象的性质，可分为计量资料与计数资料。

一、计量资料

计量资料是通过直接计量而得来的，即用度量衡等计量工具直接测定的。其数据是用长度、容积、重量来表示（如体高、产奶量、体重、产毛量等）。所测定的各个变数不一定是整数，在整数之间可以出现任何值。如测定奶牛的产奶量在3000公斤至6000公斤之间。可以有很多个变数，这些变数可以是整数，也可以是小数。在统计学中把这种资料称为计量资料，也称连续性变数资料。

二、计数资料

计数资料是指用计数方式而得来的。在这类资料中，每一个变数必须以整数来表示，两整数间的数值是不连续的，因此不具有小数。计数方式表示的数据是以1为单位。如猪的头数、产仔数、成活数、鸡的产蛋数、牛和猪的乳头数等，这些变数都是以个数来表示，各变数是不连续的。对这种资料称为计数资料，也称间断性变数资料。

计数资料亦包括质量性状数量化的资料。所谓质量性状，是指一些能观察到而不易直接测量的性状，如颜色、性别、生死等，对于质量性状的分析，须先将质量性状数量化。其方法有下列两种。

(一) 统计次数法 根据某一质量性状的类别统计其次数，以次数作为质量性状的数据。在分组统计时，可按质量性状的类别进行分组，再统计各组出现的次数。例如，

在研究猪毛色遗传时，其分类统计表如下：

表 2—4 杂一代白毛黑斑猪自交后的毛色分离情况

性状分组	毛 色			总 计
	白	黑	花	
次 数 (<i>f</i>)	332	96	22	450
次数百分率(%)	73.78	21.33	4.89	100

(二) 评分法 对某一质量性状，因其类别不同，分别给予评分。例如，研究绵羊的油汗色泽遗传时，可将五种油汗色泽分别给以不同的分数，以便统计分析，其评分统计表如下：

表 2—5 某种公羊85头后代的油汗色泽评分统计表

性状类别	油 汗 色 泽					总 计
	深 黄	黄	浅 黄	乳 白	白	
评 分	1	2	3	4	5	
次 数 (<i>f</i>)	5	8	16	21	35	85

第二节 原始数据的检查与核对

搜集的原始数据在整理之前，首先要对全部数据进行检查与核对，分辨真伪，力求完整、真实和准确。

对数据检查与核对，应注意以下三方面问题。

一、数据本身差错

记录不全（丢失、损坏、遗漏）、记载错误（笔误、虚构）以及测量工具不准、测量技术不熟练等原因所造成的错误。因此对一些特殊数值（如极大或极小的），应反复核实，以保证计算的正确与可靠。

二、取样差错

取样不全或过少或非随机取样（取样时加入人的因素，把合乎意图的留下，把不合乎意图的删掉），都会影响统计的准确性。

三、数据不合理的合并

当整理数据时，对公畜与母畜、健康与疾病、妊娠与空怀、品种不同、年龄不同、胎次不同、营养不同、时间与年度不同、试验因素不同等数据进行了不合理的合并，在数据检查与核对时，就应加以纠正。

数据的检查与核对，在统计处理工作中是一项非常重要的工作。只有经过检查与核对的数据才能进行统计分析，真实地反映出调查和试验的客观情况。

第三节 数据的分组

调查或试验所得到的数据，经检查与核对后，根据数据的多少确定是否分组。当数据不多时（在30个变数以下的小样本）不必分组，可直接进行统计分析。当数据较多时，一般在30个变数以上的大样本需要将数据分成若干组，以便统计分析。经过分组归类后，制成较有规则的次数分布表，并根据次数分布表，作出次数分布图。

数据分组的方法。

一、间断性变数资料的整理与分组

间断性变数资料的整理与分组基本是采用单项式分组法。它的特点是用样本变数的自然值进行分组，每组均用一个变数值来表示。分组时，可将数据中，每个变数分别归入相应的组内，然后制成次数分布表（见表2—3）。

由整理所得的次数分布表，可以了解数据的集中和变异情况，并便于进一步计算与分析。对于变数较多而变异范围较大的间断性变数资料，若以每一个变数值为一组，则组数太多而每组内包含的变数太少，这样就不易表现出数据的规律性。这种资料可扩大为以几个变数为一组，这样可以适当减少组数，对数据进一步计算分析也较方便。

二、连续性变数资料的整理与分组

连续性变数资料的整理与分组是采用组距式分组法。组距式分组法，在分组前需要确定全距、组数、组距、组中值和组限，然后将每个变数值分别归入相应的组内。兹以200头大白母猪（经产）的仔猪一月窝重的资料为例，说明其方法与步骤。

（一）求全距 全距是数据内变数最大值与最小值之差。它是整个样本的变异幅度。由表2—6中，可以看出大白母猪的仔猪一月窝重，最高为118.5公斤，最低为11.6公斤，因此，全距为 $118.5 - 11.6 = 106.9$ （公斤）。

（二）确定组数 将全部变数按其数值大小分成若干组。组数的多少是根据变数的个数及组距的大小而定的。组数与组距关系极为密切，如增多组数就会使组距变小；减小组数则组距增大。另外，分组越多所求得的统计量越精确，但不便于计算；如分组过少，则计算出的统计量的精确性较差。一般组数的确定，可参考表2—7。

按表2—7要求，200头大白母猪的仔猪一月窝重的资料可初步确定分成14组。

（三）确定组距 每一组内的范围称为组距。分组时要求各组的距离相等。连续性变数用公式所求得的组距不一定是整数，为了便于计算可以采用整数作为组距。组距符号以“ i ”表示。

组距的大小是根据全距与组数的关系决定的，

$$\text{组距} = \frac{\text{全距}}{\text{组数}}$$

将表 2—6 资料的数据代入：

$$\text{组距} = \frac{106.9}{14} = 7.6 \approx 8$$

表 2—6 200 头大白母猪的仔猪一月窝重的资料

(单位：公斤)

47.2	67.7	89.5	108.5	53.0	31.5	60.7	84.0	62.2	68.2
57.1	51.6	70.5	97.8	58.2	78.0	64.6	84.3	50.5	68.7
40.5	56.8	58.0	41.5	85.8	58.8	51.8	60.5	91.8	63.7
78.2	29.3	23.5	67.8	62.5	96.5	84.0	26.0	74.5	79.6
58.2	41.5	63.8	67.3	80.7	11.6	64.0	34.4	28.0	65.0
75.9	59.8	79.5	57.2	76.5	34.5	49.6	67.5	18.4	53.0
61.2	63.3	71.7	65.8	103.5	12.0	58.5	69.8	40.2	55.0
85.0	66.5	39.8	75.5	55.5	104.7	85.0	55.2	104.2	75.3
66.0	96.7	20.5	15.7	20.0	26.3	79.3	77.4	102.4	95.1
83.0	56.0	70.8	76.8	74.0	118.5	42.4	26.6	86.8	53.0
43.8	83.0	71.8	36.0	67.8	72.5	90.5	101.0	41.7	76.1
89.0	65.0	100.0	31.4	71.8	94.5	70.1	88.7	56.4	27.9
81.5	36.7	88.5	79.4	69.5	79.5	72.1	65.5	23.2	40.2
49.9	81.0	41.5	76.9	72.0	82.1	64.4	69.8	54.1	68.6
40.5	56.7	59.5	25.0	62.8	70.0	90.6	67.1	56.7	73.8
62.3	72.3	113.0	90.3	57.8	86.1	34.0	39.4	49.7	41.0
50.4	67.5	79.5	68.3	38.3	68.6	34.0	14.6	25.5	54.3
94.3	88.4	85.0	60.8	81.0	52.5	117.0	73.0	23.3	71.7
85.1	78.0	81.0	74.5	72.9	64.0	80.5	108.6	72.5	87.5
56.0	90.8	89.4	70.6	55.5	102.2	93.5	94.4	67.3	47.3

表 2—7 样本大小与组数多少的关系

变数的个数 (n)	分组的组数
30—60	6—8
60—100	7—10
100—200	9—12
200—500	12—17
500 以上	17—30

(四) 求组中值 组中值是两组限 (每组上限与下限) 的中间数值。它是该组的代表值。组中值最好是整数, 当组距为偶数时, 它就可能是整数, 这样可为以后计算避免许多小数的麻烦。由于相邻两组的组中值间的距离等于组距, 所以当第一组组中值确定后加上组距就是第二组的组中值, 第二组的组中值加上组距就是第三组的组中值, 余类推。

$$\text{组中值} = \frac{\text{组下限} + \text{组上限}}{2}$$

$$\text{或} = \text{组下限} + \frac{1}{2} \text{组距}$$