水产试验统计学

谢 庄 张元明 主编

中国农业科技出版社

前　言

《水产试验统计学》是水产类科技与教学工作者在科研、教学、生产、推广中一本极有应用价值的参考书，也是水产类本、专科学生必须掌握的一门专业基础课。

随着水产事业的不断发展，水产科研、教学的进展也十分迅猛。如何使更多的水产工作者掌握试验设计与数据处理方法，从而使得水产科研的设计更趋合理、科学，并取得更高层次的结论，是广大水产科研工作者的迫切希望和要求，也是高等农、水院校在培养水产类专业学生的需要。而目前我国针对水产科研方面的这一类参考书和系统教材尚不多见。因此，我们在多年教学、科研的基础上共同编写了这本书，希望得到全国同行的认可。

本书在编写过程中，力求在坚持科学性、系统性的基础上突出应用性、实践性。

本书共分 14 章，主要内容包括大样本资料的收集与整理、数据基本特征值的计算与估计、理论分布、显著性检验、效应值的估计、相关与回归分析、抽样原理、渔获量的估计、试验设计原理与方法及结果的统计分析等内容。本书在强调统计方法基本概念的基础上，强调各种方法与公式的应用及生物学意义。每章后还附有相当数量的复习题以供练习。

本书的编写分工为：谢庄第一章、徐夕水第二章、姚式群第三章、吕小欢第四章、马国平第五章、王惠珍第六章、章元明第七章、刘娟第八章、庄尔铮第九章、管荣展第十章、陈晓生第十一章、朱锦兰第十二章、陈建酬第十三章、倪海儿第十四章。最后全书由谢庄、章元明统稿。

此外，杨坤、谢映海、刘丽均等同志也为本书做了大量的文字录入、校对、制图等工作，在此一并致以衷心感谢。

由于编者水平有限，错误与不足在所难免，恳请读者与同行不吝赐正。

谢　庄

1998 年 2 月

《水产试验统计学》编委会

目　　录

第一章 绪 论

第一节 试验统计学的概念和特点

生物体是自然界中最复杂多变的物体，它所表现出来的性状，特别是带有随机性的数量性状是千变万化、参差不齐的。这就是统计数据的变异性。如何来处理、分析这些看似杂乱无章的数据资料，发现其潜在的规律性，一般的数学显得无能为力。将概率论和数理统计学原理应用于生物界带有随机性的数量变化的研究，就形成了生物统计学。水产试验统计学是应用概率论和数理统计学原理研究如何以有效的方式搜集、整理、分析带有随机性的渔业数据，对所研究的问题作出统计推断，直到对可能作出的决策提供依据或建议的一门学科。

半个多世纪以来，生物统计学已广泛应用于生物学科的各个方面，水产学科也不例外。越来越多的水产科研工作者使用水产试验统计学的知识来设计科学试验，处理试验数据和调查结果，从而得出合理、客观、正确的结论，对水产学科的科学研究和生产的发展起着重要的促进作用。

水产试验统计学的推理思想与其它自然学科不同。例如，“血是红的，天鹅是白的，乌鸦是黑的”是通过简单枚举产生的结论。然而，在南极洲却发现了白色血液的鱼，在日本发现了白乌鸦，在澳洲发现了黑天鹅。如果根据其它自然学科的推理方法，出现反例就否定原假设，这就要否定“血是红的，天鹅是白的，乌鸦是黑的”的结论。但是，如果我们要概括上述结果，就可以这样说，“至少 99%的动物其血是红的，至少 99%的天鹅是白的，至少 99%的乌鸦是黑的”。这种归纳推理称为概率归纳推理。它的特点就是所作结论并不是 100%的正确，而在一定的概率保证下是正确的。水产试验统计学中的“部分推断全部”的统计结论也是在一定概率保证下是正确的，即按概率归纳推理进行的推断。因此，水产试验统计学既不同于描述性学科，也不同于经典数学，它是按概率归纳原理进行的推理，概率性是它的第一个特点。

水产试验统计学有它自身的理论体系，这是任何一门独立学科必不可少的。但同时必须面对大量来源于实践的数据资料，否则就失去了它存在和发展的价值。处理和分析渔业数据资料是水产试验统计学的首要任务。因此，水产试验统计学不是一门纯理论的学科，而是理论和实践并重、理论和实践密切联系的学科。这是水产试验统计学的第二个特点。

理论上，我们总希望获得并处理同一性质的所有资料，然而实践中只能获得其中很少的一部分，并对这很少的一部分资料进行分析和处理，从而得出一个统计结论。因此，水产试验统计学的第三个特点，就是对部分资料（称为样本）进行分析，在一定的概率保证下推断全体数据资料（称为总体）的带有普遍意义的结论和规律，即从特殊推断一般。用统计学自身的语言来说，就是用样本推断总体，或者用样本的数量特征值（平均数、均方等）、数量关系（相关系数）和数量变化（回归方程）来推断总体的数量规律性。其前提条件是样本必须为随机样本。

一般说来，水产试验统计学所要处理的数据资料来源于两个方面：调查和试验。这就涉及到抽样和试验设计。正确地确定抽样方案，进行试验设计是统计分析工作的基础。所谓试验设计，是指在试验工作进行之前，应用水产试验统计学原理，制订合理的试验方案，如样本的最佳配置、正确选择试验动物、正确拟定试验进程等，使我们可以利用尽可能少的人力、物力、财力和时间，在试验结束后能获得尽可能多的、可靠的资料和信息，进行统计分析，以得出可信的科学结论。

第二节　试验统计学的发展简史

水产试验统计学是在生物统计学的基础上发展起来的，而生物统计学又是统计学发展到一定阶段形成的。因此，我们通过介绍统计学的发展简史来阐明水产试验统计学的发展简史。

统计学是随着社会生产发展和适应国家管理的需要而发展起来的。统计学的发展史可追溯到远古时期。例如，管仲说："不明于计数，而欲举大事，犹如无舟楫而欲经于水也。""举事必成，不知计数不可"。但是，它作为一门独立的学科——统计学，距今只不过才 300 多年的历史。

从统计学的产生和发展过程来看，它可划分为古典统计学、近代统计学和现代统计学三个时期，形成国势学派、政治算术学派、社会统计学派和数理统计学派四大主要学派。

古典统计学时期是指 17 世纪中末叶至 18 世纪中末叶的统计学萌芽时期，主要有国势学派和政治算术学派。前者的主要贡献是提出了统计学(G. Achenwall 1749)以及统计数字资料、数字对比等统计术语。后者主要是用计量方法研究社会经济问题，仅仅是用数字表示社会经济规律，主要代表人物是 J. Graunt 和 W. Petty。

近代统计学时期是指 18 世纪末到 19 世纪末，主要有数理统计学派和社会统计学派。比利时的 A. Quetelet 把概率论和数学方法引入统计学以解决生物学、医学和社会学中的问题，形成数理统计学派。他运用大数定律论证社会生活现象并非偶然，有其发展规律性；运用概率论提出"平均人"概念，第一个认识到大量的变异数据中孕育着规律性。这一时期应用生物统计方法的还有 C. R. Darwin、G. Mendel、F. Galton 和 K. Pearson 等。Darwin 进化论的创立离不开生物统计学；Mendel 分析豌豆杂交试验的结果运用了生物统计方法；Galton 把生物统计学引入遗传学研究，通过对生物变异的分析，提出了相关和回归的概念和计算方法。此外，还设计了百分位数法和百分位秩。K. Pearson 把生物统计学应用于生物学研究，发现了 χ^2 分布，提出了度量实际观测值与理论值差异程度的 χ^2 检验以及计算复相关、偏相关和标准差的方法；与 Weldon 一起于 1901～1902 年创办了《生物统计学报》；并与 Galton 一起首次提出生物统计学 Biometry 一词。社会统计学派是研究社会现象变动原因和规律性的实质性科学，用大量观察法研究社会总体。

现代统计学时期是指 20 世纪初到现在的数理统计学时期。在随机抽样基础上建立了推断统计学，形成了贝叶斯学派和非贝叶斯学派。K. Pearson 的学生 W. Gosset(1908)以"Student"为笔名将 t 检验发表于《生物统计学报》，此外，还阐述了样本标准差、样本平均数与标准差之比以及相关系数的抽样分布，由此奠定了小样本理论基础。R. A. Fisher 提出了很多

统计方法，对推动农业科学、生物学和遗传学的研究和发展起到了很大作用，因此被公认为是生物统计学的奠基人。如为了比较不同因素在试验中所起的作用他提出了方差分析法(1923年)；同时为了保证误差估计值的有效性引入了随机性原则。在1915年和1921年完善了样本相关系数的抽样分布并提出小样本相关系数的Fisher氏转换，在1925年提出了随机区组设计和拉丁方设计，编制了t分布概率表和F分布显著水平临界值表。此外，他还认为t分布不仅适用于小样本，还适用于大样本。P. O. Johnson认为，从1920年至今可称为Fisher氏时代。20世纪30年代，J. Neyman和E. S. Pearson共同对假设检验的理论进行了系统的研究。1940～1950年期间，Neyman提出了区间估计理论。A. Wald把现代数理统计学原有的估计理论和假设检验理论结合起来，形成决策理论，于1946年出版了《统计的决策函数》；他又于1947年出版了《序贯分析》一书，从而开拓了一个崭新的研究领域。在W. Gosset和R. A. Fisher之后，S. S. Wilks、J. Wishart、H. Hottening、R. C. Bose和T. W. Anderson等在样本分布理论方面也作出了贡献。W. G. Cochran和G. M. Cox于1957年出版了《试验设计》一书。

我国早在30年代，生物统计学已成为农学专业的必修课，最早出版的专著有王绶编著的《实用生物统计法》。生物统计学在水产研究和生产实际中的应用比较迟，很长一段时间没有水产方面的生物统计学教材和课程，只是在近年，各高校的水产养殖专业才络续开设了本课程。

第三节　试验统计学的几个常用术语

一、总体和样本

总体是指具有相同性质的观测值所组成的集合。由于生物体具有许许多多性状，因此相似的生物体所组成的集合，如同一类群、同一物种，就不是水产试验统计学意义上的总体，只有相似生物体所具有的某一共同性状所表现出来的值的集合才称为总体。例如，一龄草鱼的体重、二龄鲤鱼的体长等。总体可以是无限的，也可以是有限的。所谓无限，既有时间上的含义，又有空间(或地域)上的含义。当限定某一时间和某一地域时，总体就成了有限。如亲鳖的年产卵量，就是一个无限总体，但如果把亲鳖的产卵量限定在某一区域范围内(如湖北省)，再限定在某一年份(如1998年)，这就是一个有限总体。由于总体往往是无限的，即使是有限的，其量也很大，因此不可能在实际工作中对总体中的所有观测值进行一一考察，只能对其中具有代表性的一小部分进行研究。为了能对总体有一个很好的了解和认识，被研究的这一部分观测值必须来自于该总体，并能很好地代表总体。这样的一批观测值的集合就称为样本。在总体中抽取样本的这一过程称为抽样。为了使所得到的样本能无偏地估计总体，必须使总体中的每一个观测值都有同等的机会进入样本，这种抽样方法称为随机抽样法。随机抽样所取得的样本就称为随机样本。没有特别的说明时，随机样本简称为样本。

一个样本内的变量(观测值)个数，称为样本容量。样本容量较大时，称为大样本；样本容量较小时，称为小样本。但大样本和小样本并没有严格的界线。习惯上，可以把30作为大小样本的分界线；但在很多情况下，这一分界线是不够合理的。

二、变数、变量、参数和统计量

在实践中，无论是总体还是样本，无论是调查还是试验，得到的数值都是有差别的，这种差别在水产试验统计学中称为统计数据的变异。例如，同一规格的同一类鱼，其体重不会完全相同。这种具有某一性质或特征的变异在生物统计学中称为变数，变数在某一个体具体表现出来的数值称为变量或观测值。

由总体各观测值计算所得到的用来描述总体特征的数值称为参数。例如，由总体算得的平均数反映了总体观测值的集中程度和一般水平，因此总体平均值是参数。参数往往用希腊字母来表示。相应的，由样本观测值计算得到的描述样本特征的数值称为统计量或统计数。例如，样本平均数反映了样本变量的集中程度和一般水平，因此样本平均数是统计量。统计量常用拉丁字母来表示。

由于总体很大，参数往往不容易获得，常通过样本统计量来估计相应总体的参数。通过统计量无偏地估计参数是生物统计学中一个很重要的内容。

三、误　差

在科学试验中，除了对希望所要研究或讨论的某一个或几个试验条件（生物统计学中称为试验因素）人为地加以区别（生物统计学中称为设置水平）外，其余外部及内部的条件都应当保持一致，以使试验所得到的结果符合真值。然而，在生物学科（也包括水产学科）中，人们几乎无法把非试验条件绝对地控制在同一水平上，同时所试验的对象也是千变万化的生物体，因此很难使所得到的试验结果完全符合真值。试验结果与真值之间的这种偏离，就称为误差。误差按其来源和性质可分为系统误差和随机误差两种。

系统误差是指由于某些特定的非试验条件所造成的使试验结果朝一个方向发生有规律的偏离。造成系统误差的原因有以下几种：(1)度量工具的不正确且未经校正，如称量鱼体重用的秤发生了偏差，结果所有的鱼体重都减少了或增加了一定的重量；(2)活的水生生物体称重时都带有一些水或其它杂质，且这些水或杂质会逐渐积存于称量的容器中而使得称重发生累积性的变化；(3)实验仪器及其读数器未经校正，而使试验所得读数发生偏差，如分光光度计等；(4)外界试验环境发生了较大的变化，如灌排水的水温、溶氧量、肥瘦程度、pH 值等都会使试验用的水生生物发生生长发育方面的变化，传染性疾病的流行使得实验动物生长发育受阻；(5)观测时间及顺序的影响；(6)试验人员操作及观测时个人的偏爱及习惯；(7)试验用的水生动物分组时发生的偏差等。这些因素都会使试验所得结果有规律地偏离真值。有些偏差可能呈线性变化，这一类偏差比较好校正；然而有些偏差可能呈非线性变化，且其规律不易被发现以致几乎无法进行校正。因此，系统误差应当在试验前就加以预防和克服，使得系统误差发生的可能性降为零。系统误差一般来说也是能被消除的。

随机误差是指由于各种偶然因素引起的、无法加以预测和控制的无规律的偏差。因此，随机误差又称为偶然误差。随机误差的大小、方向都无法确定，完全取决于该观测值在观测和试验过程中的机会。例如，在试验和观测过程中，不管实验仪器多么精密、观测手段多么完善、操作过程多么精细，其观测结果总会发生大小不等、方向不定的偏差。排除系统误差以后，试验中主要的或者是唯一的误差来源就是随机误差。因此，随机误差也就是试验误差，或

者说，试验误差的绝大部分来源于随机因素。在不发生歧义的情况下，随机误差简称为误差。可以发现，如果观测次数足够多的话，对同一事物进行重复观测所得到的随机误差有统计学上的意义，即每一次观测时所产生的随机误差都是独立发生的，且误差的分布服从于后面所要讨论到的正态分布。可以通过各种手段把随机误差降到最低的程度，但是无法消灭它。实际上，随机误差是进行统计假设检验的基础。降低随机误差，可以提高试验的精确性，可以更好地区别误差效应和处理效应，使得试验结果更正确，对试验处理间的差异所作出的评定更准确，更可靠。

除以上两类误差外，还有一类误差是由于工作人员的粗心大意或不负责任所造成的，这类误差称为错误。如仪器使用不当、错读数据、记录不准、不完善、任意涂改、凭空捏造等造成数据的不真实。在试验和调查过程中，错误是应当也是完全可以避免的。

误差的来源和控制途径如下：

(1)试验材料的不同质性。生物试验一般所用的试验材料是生物有机体，由于遗传物质的多样性及遗传物质分配的随机性，而使得生物体在遗传上表现出复杂的变化，特别是数量性状方面的变化。这种遗传基础上的差异必然会带来试验结果的不一致。例如，饲养在同一条件下的鱼，吃的饵料相同，饲养管理条件也相同，其增重效果却不一样。这在很大程度上是由于决定鱼生长发育的遗传物质的不同而造成的。即使是同一尾亲鱼的后代，其表现也会有差别。这种由于遗传因素引起的差异至少在目前还无法加以控制。因此，在作试验时，应尽可能找遗传基础基本一致的试验材料，并尽可能地随机化。例如，在做饲养试验时，尽量使用同一尾亲鱼的后代进行分组，同时使样本大一些；进入试验时，各个被试个体应在各方面都保持一致，如做不到这一点，或初始时体重、规格有差异的话，则在试验结束后应使用协方差分析法进行校正。

(2)外部条件的不一致性。即使是在完全由人工控制的情况和环境下进行实验，还是会发现仍有一些不能或无法控制的因素在干扰着试验，使得试验结果产生偏差；在自然环境和条件下进行试验，这种外部干扰就更大，更难控制。例如，鱼池的地理位置不同致使光照、进水的理化性质、风力的大小、疾病及寄生虫的危害程度、水的肥瘦等等都不一样，而这些差别又都使得不同的处理组合带有一定程度的差异；室内试验环境也会受到加热、保温、通风、电力强弱等因素的影响。控制这一类误差，一是要有良好、严格、科学的试验设计；二是要在作统计分析时运用有关知识把这些差异从总变异中剖分出来。这两者在很多情况下是有密切联系的。对于前者，除考虑设置区组并适当增加重复数外，还需考虑试验鱼池的排列、进水与排水的顺序问题及鱼池区组的配置等。一般来说，试验用的鱼池在换水时，每一鱼池都应是独立的，不宜将上风向鱼池的水流向下风向鱼池，使鱼池失去独立性。类似的情况如施肥、用药等都应注意，否则将使试验失去较大的精确性和准确性。当试验是属于推广性或应用型时，还应注意鱼池的生态条件和环境的代表性，否则试验将失去价值。

(3)操作管理的不一致性。在不止一个工作人员参与试验的情况下，由于各人的实践经验、对试验的领会程度、责任性等不同，会使各人所实施试验的部分产生人员方面的差异，即使是同一个工作人员负责管理，也会由于操作时间的先后、疲劳程度、工作情绪的变化、外界的影响而发生一些差异，这种差异对试验的影响有时是很大的。因此，应针对所出现的情况采取对策，使这种人为的误差尽可能减少。一方面要加强对工作人员的培训、教育，特别是责

任性方面的教育；另一方面，应注意鱼池区组和工作人员的调配，不能某一个人集中于某一个特定的处理水平，而使工作人员的影响与处理水平产生混杂而不能区分。在可能的情况下应尽量采用机械化操作，以减少工作人员的疲劳程度和人为的误差。

四、准确度与精确度

准确度与精确度是和两类误差紧密联系的两个术语。

准确度是指观测值与真值接近的程度。当发生系统误差时，观测值都有规律地向某一方向偏离真值，因而降低了试验的准确度。精确度是指在同一处理条件下，同一批观测值间相互接近的程度。当随机误差较大时，数据较离散，精确度较小。

相比之下，准确度是比精确度更重要的一个概念。由于准确度与精确度往往不可兼得，因此在做试验时，应当很好地加以权衡。原则上，可适当放弃一些精确度以取得较高的准确度，即宁可适当地扩大一些随机误差也要把系统误差降至最小的程度直至为零，或把系统误差化为随机误差。

水产试验统计学是水产科研和生产中的一个重要工具。它能帮助水产科技工作者发现隐藏在纷繁表观现象下面的客观规律。在水产科研和生产中用好用活水产试验统计学，除了学好本门学科，掌握其原理、计算公式、数学概念和含义、具有基本的电脑知识和操作技能外，还必须对水产业及与水产业相关的学科有充分的了解和认识。用水产试验统计学处理和分析每一批资料、每一批数据，都必须有充分的生物学意义、水产学科的含义及渔业经济管理学意义，所作的试验应当具有水产学科的理论意义和实践意义；否则，即使公式用得再熟练，计算结果再正确、再精确，也是毫无实际意义的。因此，水产试验统计学的学习和使用不能孤立地、单独地进行，必须结合水产实践，以取得具有指导意义的结果。

第四节　试验统计学的主要内容

水产试验统计学是生物统计学在水产生产和科学研究中的应用。由于水产学科与其它学科间有着较大的差距，因此水产试验统计学除了一般生物统计学所具有的共性外，还有其特殊性。

生物统计学主要分为描述统计学和推断统计学。前者包括数据资料的整理和一般描述，如资料集中性（算术平均数、几何平均数、调和平均数、中数和众数）、离散性或变异性（极差、标准差、方差、变异系数和标准误）和偏斜性（峰度和偏度）的描述；后者包括参数估计和统计推断。所谓统计推断就是用样本的数量特征值（平均数、均方等）、数量关系（相关系数）和数量变化（回归方程）对总体数量规律性做出推断。

水产试验统计学包括以下几部分内容。

一、大样本资料的统计分析

水产试验统计学所面对的资料可分为两大类：数量性状资料和质量性状资料。前者包括连续性资料和间断性资料。连续性资料要通过度量衡等计量工具才能得到，具有单位，并且呈连续变化，如一龄青鱼的体重、亲鱼的怀卵量、每一公顷水面的鱼产量、二龄草鱼的体长

等；间断性资料是通过计数获得的，并且呈非连续性变化。质量性状资料不是通过直接度量而是通过感观获得的资料，如亲鱼的性别、体色等。从渔业生产及调查得来的连续性资料往往是很多的，粗略一看，这些资料很零乱，不易找到其中的规律，因此需要进行整理。资料整理的内容是检查数据的正确性、完整性，找出其分布规律，计算其最基本的统计量，并用它们来估计其相应的参数，此外还有对大样本资料的分布进行适合性检验等。

二、假设检验

从试验中得到的数据仅是一个样本，从该样本中所得到的统计量是否可靠，能否用来估计相应的参数，不同的处理条件下所得到的样本间的差异是否是真实性差异，有多大的可信程度，这些都需要进行检验，这种检验在生物统计学中称为假设检验或显著性检验。

假设检验常用的有以下几种。

(一)平均数(或百分数)的假设检验

样本所属总体与假定总体以及两样本所属总体的平均数(或百分数)间的假设检验。

(二)χ^2 检验

计数资料除可以化成百分数进行 t 检验外，还可以用 χ^2 检验来检验某批计数资料是否服从某一理论值，或某两个性状间是否相互独立。

(三)方差分析和协方差分析

t 检验不能胜任两个以上平均数间的相互比较，因为这会增大犯第一类错误的概率和降低误差估计的精度等。这时可以用方差分析法。方差分析法尤其适用于多因素试验资料的分析，它可以使我们在一次试验中获得更多、更全面的信息，得到因素间的互作效应和主效应。此外，还可以根据不同的数学模型估计出不同的期望均方，进而估计出各种遗传参数。

当所得到的试验资料由于试验动物的初始状况不同使试验的正确性受到影响时，仅用方差分析是不够的。如果用回归分析法进行校正，则可使这些试验资料都能处在同一初始水平上，从而更合理地比较处理间的差异。这就是把回归分析与方差分析相结合对数据资料进行分析的协方差分析法。

三、相关与回归

生物体的性状间总能或多或少地表现出某种关系，这种关系的密切程度，称为相关；当一个性状的量发生变化，另一个量也发生相应的变化时，这种性状间的依赖关系称为回归。回归关系可以是线性的，也可以是非线性的；可以是一个性状对另一个性状的影响，也可以是多个性状对一个或多个性状的影响。因此，回归关系就有直线回归与曲线回归(包括多项式回归)、一元回归与多元回归(又称复回归)之分。例如，亲鱼的怀卵量随着亲鱼体重的增加而增加，鱼的体长、体厚又共同决定了鱼的体重，等等。

四、渔获量的估计

水产业不同于其它行业，因为绝大部分水产品生活在水域中，很难直接用肉眼观测到，因此如何较正确地估计水产品的产量，以做到心中有数，是渔业生产中一个很重要的问题。严格说来，水产品的产量估测也是一个抽样问题，但它又和农学、畜牧学中常用的一般性抽

样技术有着较大的差别。因此，水产资源的调查方法和渔获量的估计就成了水产试验统计学中一个很重要的内容。

五、 试验设计

试验设计包括抽样原理与技术、样本容量的最佳配置、试验设计的基本原理与要求、试验方案的拟定。常用的水产试验设计方法包括完全随机试验设计、随机区组设计、交叉设计、拉丁方设计、析因设计、正交试验设计等。

思考与练习

1. 统计数据的本质特征是什么？

2. 试验统计学的概念和特点是什么？并说明试验统计学的推理思想与其它自然学科推理思想的区别。

3. 试验统计学的主要内容有哪些？

4. 试举例说明样本、总体、参数和统计数的概念及其相互关系。

5. 什么叫试验误差？试验误差与试验的准确度、精确度有什么关系？如何控制试验误差？试验误差有哪些来源？

第二章　资料的整理

第一节　数据的分类

从水产试验和生产中可以取得大量的原始资料。除文字说明以外，这些资料一般以数字的形式出现，这是对试验对象进行观测的结果，通常称之为观测值。大批观测值的集合称为数据。数据往往是试验材料某一特定性状的表现。

当在试验或生产中取得的观测值较多，即所得样本较大时，由于观测值之间表现出纷繁的变化，不易立即找出其潜在的规律、联系和分布状况，必须对其进行整理、分析。在对资料进行整理时，必须注意资料的完整性、真实性和正确性。资料整理中的这三个原则贯穿于所有试验和生产中。

一般来说，我们所能得到的资料大致可分为两大类：数量性状资料和质量性状资料。

一、数量性状资料

生物体中，有些性状必须经过度量才能获得数据，其观测值随着计量工具的不断完善和精密可有不断增多的有效位数并呈连续性，这样的资料又称为连续性资料，如体重、体长、体液内某些生化物质的含量等。有些性状的观测值需通过记数的方法获得，其最小单位以个数来表示，呈不连续性，这样的资料又称为不连续性资料或间断性资料，如一个鱼塘内某类鱼的尾数、一次捕获量等。这样的资料往往有一个分布的范围，在这个范围内，观测值的分布并不是均匀的。

二、质量性状资料

有些性状不能测量，只能描述，如鱼体颜色、性别、死亡等，这样的性状称为质量性状。对质量性状的资料进行统计分析时，必须将其数量化，数量化的方法常用的有以下两种。

（一）统计次数法

将水产品进行归类和计数，同时计算每一类别的百分比。例如，根据归类和统计，可以知道一批鲤鱼中健康鱼占 84%，病鱼占 16%。该法又称为百分数法。

（二）评分法

对某一质量性状的不同表现进行打分，并统计每一分值下的个数。例如，对鱼体的肥瘦程度订立 5 级分制，最好的为 5 分，最差的为 1 分，每一分级下有多少尾鱼，同时也可计算每一分级下鱼所占的百分比，以此来评定各个渔场的好坏。

对于质量性状的资料，得到统计结果后，还可用图表将其表示出来。

第二节　资料的校核

根据性状的不同，将资料进行分类整理，不同性状的资料不能相互混淆。分类以后，应对各批资料，特别是数量性状资料进行必要的校核。

首先，应检查所用单位是否正确和规范，所有的市制一律要换算成公制：g（克）、kg（千克）、cm（厘米）、km（千米）、J（焦耳）、mJ（兆焦）、hm^2（公顷）等。

其次，检查记录是否完整，是否有遗漏、多余或重复；记载是否错误，是否有极大、极小等异常数据或可疑数据出现。一旦出现异常数据，应反复核实，包括计量工具、计量技术、记录等都在核查范围以内。

再次，检查数据记录是否掺有工作人员的主观因素，取样是否具有代表性、是否随机。

最后，检查资料的合并是否合理，同时还应考虑合并的场别、年份、鱼类及鱼的规格等。如果场际之间、年度之间、规格之间差别过大，就应考虑进行适当的分类。

总之，资料的检查是一项十分重要的准备工作，只有经过检查认为合理、可靠的数据才能进行统计分析；否则，不但不能得到真实可靠的统计结果，而且还有可能产生错误的结论和导向，给生产和科研带来不应有的损失。

第三节　资料的整理

间断性资料的整理与分组一般是采用单项式分组法。其整理方法是统计样本中各观测值出现的次数。即每组均用一个观测值来表示，它的整理实质上就是将相同的观测值作为一组，然后统计各组观测值的次数并制成次数分布表。但是，当样本观测值个数较多，变异幅度较大时，统计每个观测值出现的次数并不能反映资料潜在的规律性，这时可将几个变数合为一组，然后统计各组观测值出现的次数并制成次数分布表。

对于连续性资料，当样本较小时，不必进行分组；样本较大时，由于不能看出其分布规律必须进行整理和分组。

表 2-1　120 尾大银鱼体长资料　(cm)

12.5	12.8	12.7	12.4	13.2	13.3	11.2	14.9	14.3	13.2	10.3	14.8
14.0	13.6	13.7	14.6	13.4	10.3	11.5	14.0	11.4	12.5	13.4	12.0
12.3	13.5	13.2	15.1	16.0	11.6	13.2	12.6	12.8	12.0	11.8	11.9
13.4	13.9	12.9	11.3	13.2	11.6	10.7	12.6	12.9	14.6	12.8	12.2
11.7	12.1	10.8	12.1	9.8	11.9	11.6	12.8	15.1	10.5	14.8	13.7
15.6	11.5	13.8	13.5	12.0	14.1	14.7	11.9	12.2	10.1	11.7	13.5
14.8	14.3	12.7	12.7	13.5	12.3	12.8	12.0	12.3	13.6	12.7	11.2
13.4	12.8	12.6	11.2	12.8	11.3	11.2	13.3	11.8	12.5	13.9	11.8
11.9	12.7	13.8	14.2	12.9	13.8	12.5	11.6	13.4	12.2	12.3	12.6
12.9	11.0	12.0	13.3	12.5	12.2	13.7	12.3	10.7	11.7	13.0	12.5

常用的整理方法为组距式分组法。它的一般步骤是确定该批资料的全距、组数和组距、

组中值和组限，将各观测值归入相应的组内，统计各组观测值个数，画出次数分布表和次数分布图。

现以大银鱼体长为例，说明连续性资料分组法的一般步骤。表 2-1 是某研究所于某年对所捕捞的大银鱼进行抽样的体长资料，样本容量 n＝120。

首先，求出资料全距。找出最大值和最小值，本例中最大值和最小值分别为 16.0cm 和 9.8cm，全距为两者之差，等于 6.2cm。

其次，确定组数和组距。样本大，组数多；样本小，组数少。并且，组数和组距的关系也是相互制约的。显然，组数过多和组数过少均不能反映资料潜在的规律性。因此，求出一个适宜的组数对揭示资料的分布规律性是有意义的。组数和样本容量的大致关系可参考表 2-2。经研究表明，组数和样本容量的关系可用以下两公式描述：

$k_1 = 1 + 3.322\lg(n)$　　$k_2 = 1.87(n-1)^{0.4}$

较适分组数是在 k_1 和 k_2 之间的整数。对于本例，

$k_1 = 1 + 3.322 \times \lg 120 = 7.91 \approx 8$

$k_2 = 1.87 \times (120-1)^{0.4} = 12.6 \approx 13$

表 2-2　样本容量与组数的关系

样本容量	分组数
50～100	6～10
100～500	8～18
500～1 000	16～25
＞1 000	＞25

由此，可初步确定为 10 组。这里讨论的是等距分组（即各组的组距相等）法，这是资料整理中最常用、也是最简单的一种方法。所谓组距是每一组的间距，即该组最大值与最小值的距离。组距的计算方法是：

组距＝全距/组数

本例的组距＝6.2/10＝0.62≈0.6cm。为计算方便，组距总取一个较为简单的数。

第三，计算组中值和组限。一个组的中点称为该组的组中值。一个组的最大值和最小值分别称为该组的上限和下限。先确定第一组的组中值，由此确定该组的组限，以此确定其余各组的组中值和组限。第一组的组中值一般可取样本最小观测值或接近最小值的一个较简单的数，这样可避免第一组的观测值次数过多，较正确地反映资料的规律性。

第一组的组中值减去一半组距等于该组下限，下限加组距等于上限，该上限同时也是下一组的下限。每一组的上下限以此类推。

下限＝组中值$-\frac{1}{2}$组距，上限＝组中值$+\frac{1}{2}$组距

对于本例，第一组的组中值取资料的最小值 9.8。第一组的下限则为 $9.8-\frac{1}{2}\times 0.6$ ＝9.5，上限即为 $9.8+\frac{1}{2}\times 0.6$＝10.1 或 9.5＋0.6＝10.1。10.1 既是第一组的上限，同时也是第二组的下限。第二组的组中值为 $10.1+\frac{1}{2}\times 0.6$＝10.4，第二组的上限为 10.1＋0.6＝10.7。以此类推，直至某一个组的上限大于资料中的最大值，该组即为最后一个组。可以发现，实际得到的组数可能比原定的组数多。本例组数定为 10 组，实际分为 11 组。最后一组的上限和组中值分别为 16.1 和 15.8。

由于一个组的上限等于下一个组的下限，所以在具体分组中可能会发生某些数据既可归入这个组，也可归入下一个组的情况。为了使资料中的每个数据都只有唯一的归组，可采取组限比实际数据多保留一位小数的做法；也可不写出上限，让每一组的上限无限逼近下一