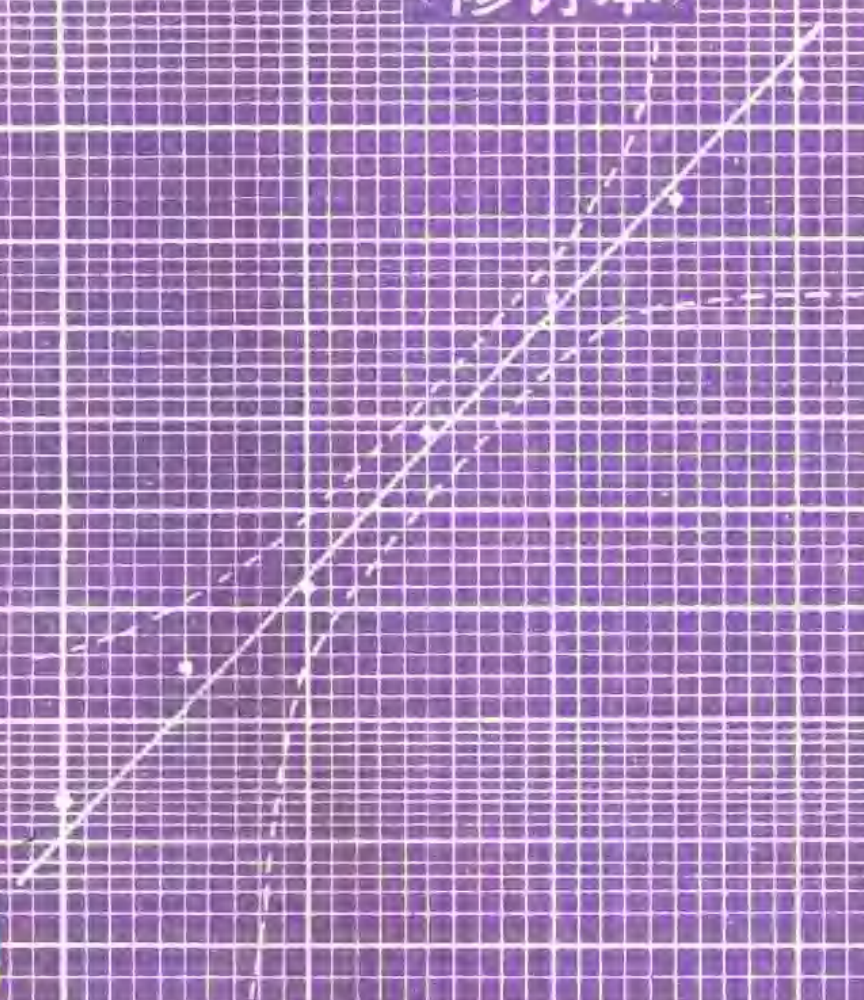


体育统计方法

TIYU TONGJIFANGFA

(修订本)



湖北省体育科学研究所

前 言

体育统计方法是正确地进行体育科研、教学、训练中数据处理的重要工具。

随着体育事业的不断发展，广大体育教师、教练员和科研人员越来越感到掌握体育统计这一工具的重要性，迫切希望有一本通俗易懂而又比较结合体育实际的体育统计方法参考书，供平时学习和进行统计计算时参阅。

编写这本“体育统计方法”的指导思想，就是以结合实例介绍方法为主，辅以用通俗的语言说明基本原理。因而，它不是按一般的数理统计书籍的体系由抽象的数学概念开始，而是从资料的收集、整理入手，一步步由浅入深地用体育教学、训练、科研中的实例来介绍体育统计的各种方法。希望学过这本书的读者能进一步感觉到体育统计不是深奥莫测的数学理论，而是从事体育教学、训练和研究工作时有用的工具。从而能联系自己的工作实际用学到的方法去研究问题、分析问题。

一九八三年四月在湖北省体育科研所的领导和同志们的关心和支持下，在体育统计研究会刘厚生、施丽影、常庆、彭寅安等同志的热情协助下，出版了这本体育统计方法。

两年多来，不少体育学院、体育系以它作为体育统计的教材或参考书，许多体育工作者、卫生工作者把它作为进行统计工作的参考手册。

这次趁再版之机，对内容作了较多修改和补充，增加了一些章节，力图在保持通俗易懂的特点的同时，增加更多的体育教学、训练和科研的实例。但是，限于自己的水平，缺点错误还是难免的。敬请批评指正。

王 路 德

一九八五年九月

目 录

绪论	(1)
第一节 数理统计在体育教学、运动训练和体育科研中的作用	(1)
第二节 正确运用体育统计方法	(2)
第一章 统计资料的收集与整理	(3)
第一节 统计资料的来源	(3)
第二节 统计资料的收集与整理	(6)
第三节 统计资料的分组	(8)
习题	(13)
第二章 几个常用的统计量	(14)
第一节 算术平均数、标准差、变异系数	(14)
第二节 算术平均数、标准差在体育中的应用实例	(19)
第三节 标准误	(25)
第四节 率及其标准误	(26)
第五节 百分位数	(27)
习题	(29)
第三章 概率知识	(30)
第一节 随机事件和概率	(30)
第二节 随机变量与概率分布	(31)
第三节 正态分布	(32)
第四章 假设检验	(34)
第一节 假设检验的原理简介	(34)
第二节 样本平均数差异显著性的 t 检验	(35)
第三节 样本率差异显著性的 t 检验	(39)
第四节 t 检验在体育中的应用实例	(40)
第五节 χ^2 检验	(43)
第六节 正态检验	(46)
第七节 秩和检验	(49)
第八节 进行显著性检验时应注意事项	(53)
习题	(54)

第五章 动态分析	(55)
第一节 动态分析的计算方法.....	(55)
第二节 动态分析的统计图、表.....	(58)
第三节 动态分析在体育中的应用实例.....	(60)
习题.....	(61)
第六章 方差分析	(62)
第一节 方差分析原理简介.....	(62)
第二节 单因素方差分析.....	(62)
第三节 各组平均数之间的相互比较.....	(64)
第四节 双因素方差分析.....	(67)
第五节 正交设计简介.....	(71)
习题.....	(75)
第七章 相关与一元回归	(76)
第一节 相关与回归.....	(76)
第二节 散点图与回归直线.....	(77)
第三节 相关系数及其检验.....	(78)
第四节 表格式计算相关系数与一元回归方程.....	(80)
第五节 检验一元回归方程的效果.....	(81)
第六节 两个一元回归方程的比较.....	(84)
第七节 用电子计算器计算一元回归方程.....	(85)
第八节 一元回归在体育中的应用实例.....	(86)
习题.....	(88)
第八章 多元线性回归	(89)
第一节 二元线性回归方程.....	(89)
第二节 表格式计算多元线性回归方程.....	(91)
第三节 多元线性回归的方差分析.....	(96)
第四节 复相关系数与偏相关系数.....	(98)
第五节 利用标准回归系数或偏回归平方和进行因素分析.....	(100)
第六节 多元回归分析的实例.....	(102)
第七节 两个多元回归方程的比较.....	(107)
第八节 合理地挑选指标建立回归方程.....	(109)
习题.....	(110)
第九章 逐步回归	(112)
第一节 逐步回归分析方法的计算过程.....	(112)
第二节 逐步回归在体育研究中的应用实例.....	(119)
第十章 曲线回归	(121)

第一节 配曲线的方法.....	(122)
第二节 曲线拟合优度的比较.....	(125)
第三节 曲线回归在体育中应用的实例.....	(126)
习题.....	(126)
第十一章 聚类分析	(127)
第一节 R 型聚类.....	(127)
第二节 Q 型聚类.....	(131)
第三节 模糊聚类.....	(134)
习题.....	(136)
第十二章 判别分析	(137)
习题.....	(141)
第十三章 主成分分析	(142)
习题答案	(147)
附 录	
1 正态概率纸.....	(149)
2 随机数表.....	(150)
3 正态分布表.....	(152)
4 t 值 表.....	(153)
5 相关系数可靠度表.....	(153)
6 等级相关系数可靠度表.....	(153)
7 F 值 表.....	(154)
8 多重比较的Q 值表.....	(156)
9 χ^2 检验临界值表.....	(158)
10 正态性D 检验临界值表.....	(159)
11 秩和检验临界值表.....	(160)
12 确定最大(小)秩和数的临界系数 C 表.....	(162)

绪 论

第一节 数理统计在体育教学、运动训练和体育科研中的作用

体育运动的目的任务是要增强人民体质，提高运动技术水平。为了更好地达到这一目的，就要求广大体育教师、教练员和科研人员不断地总结经验，开展科学研究，探索体育教学和运动训练的客观规律，以便提高教学和训练工作的水平。

用什么方法去探索客观规律呢？

数理统计方法是研究客观事物规律的一种方法。在客观世界中许多现象，人们可以事先预言它们在一定条件下是否会出现。例如：在标准大气压下把水加热到 100°C 就一定会沸腾；把一个球用力向天上抛去，球最后一定会落到地上。这些无论试验多少次都无一例外必然出现的现象称为必然事件。又如：对空射击，子弹是不可能飞到地球以外去的，这一类在一定条件下必然不可能出现的现象称为不可能事件。但是，除了可以事先预言的必然事件和不可能事件之外，在客观世界中还有着更大量的现象，却是事先无法预言其结果的，这在体育运动中也是极多极多的。例如：球赛之前裁判员召集双方抽签挑选场地时，使用圆形塑料挑边器，它一面是红色，另一面是绿色，裁判员将它向上抛起后，落下来将会是红面向上还是绿面向上？这是事先无法预言的；又如：一个人投篮时，在球出手前无法准确地预言球投中篮还是不中篮？一个人跳远，事先也无法准确地预言他跳出的成绩是多少？这一类事先无法预言其结果的带有偶然性的现象称为随机现象。随机现象的每一种表现就称为随机事件。从表面上看随机现象完全是带有偶然性的，好象并无规律可言，但是观测大量次数以后就可以发现随机现象也是具有一定的内在规律性的。例如，一个质地均匀的塑料挑边器，只掷一次或几次，红面向上几次是没什么规律的，但是如果掷几百、几千次并记下每次的结果，最后就会发现红面向上和绿面向上的比例数几乎是相同的。又如：让某运动员投篮，对于他的每一次投篮来说，事先无法准确地预言投中还是不中，但是，让他投许多许多次，就可以统计出他的投篮命中率。假设该运动员投了100次篮，投中60次，我们就可算出他的投篮命中率为60%，根据这一命中率，当然我们也就可以估计说他每一次投篮，投中的可能性是60%。假设又对另一名运动员统计了许多次投篮算出他的投篮命中率为20%，那么，当他某一次投篮前我们就可以预言他这一次投中的可能性是20%，数学上把事件发生的可能性大小称为概率，习惯用符号 P 表示，对于不可能事件 $P=0$ ，对于必然事件 $P=1$ ，而随机事件的概率就介于0与1之间。

专门研究随机现象的数量规律性的学科叫做概率论，它是数学的一个分支。而数理统计就是通过对有计划地收集到的大量数据，按照概率论的法则进行统计分析从而寻找出客观事物规律性的一种数学方法。

把数理统计方法应用到工农业方面，就产生了工业统计、农业统计，应用到医学卫生领

域就产生了卫生统计。运用数理统计方法来研究体育领域中的大量随机现象的规律性,就是体育统计的任务。

体育教师、教练员在日常教学和训练工作中经常用秒表、皮尺和其他器材测得学生的数据,特别是1981年7月教育部发出了“关于试行《高等学校学生体质健康卡片》和《中学、小学学生体质健康卡片》的通知”。规定每年都要对学生的身高、体重、胸围、肺活量、50米跑、立定跳远等等形态、机能、素质和健康状况作详细的测定,从中可以获得大量反映学生体质状况的宝贵数据。对于体育科研人员更需要测得各种数据来进行研究分析。这些数据如:身高170厘米,183厘米;肺活量3500毫升,4850毫升;50米跑7.3秒,8.1秒等都是用数值记录下来的随机现象的结果,数学的术语称为随机变量(简称变量),如果能够正确地运用数理统计的原理和方法进行分析研究,就可以从一大堆看来杂乱无章的变量中,找出事物的内在规律,发现事物相互之间的关系,并为预测今后可能的发展趋势提供丰富的材料和科学的依据。这对改进体育教学和训练工作是会有很大用处的。

因此,我们可以说,数理统计方法是体育工作者为探索体育运动的规律所必须学习和掌握的一项十分重要的工具。

第二节 正确运用体育统计方法

数理统计是数学的一个分支,有其严格的理论根据,各种统计计算公式都是经过推导和证明的。体育工作者学习数理统计应该要懂得基本的数理统计知识,学会在体育教学、训练和科学研究中常用的统计分析方法,以便能熟练而准确地分析研究体育领域中的各种随机现象。

本书着重介绍数理统计在体育领域中的应用方法。各种统计方法和公式都尽量通过实例来介绍,关于公式的来源和推导过程,有兴趣的同志可进一步参阅有关数理统计的专著。

但是,要做到正确地运用体育统计方法,必须注意以下几点:

1. 必须把体育专业知识和数理统计的原理紧密结合。也就是说一定要根据体育专业的特点结合数理统计的原理才能经过统计分析得出正确的结论。例如,工业上由于实验时的温度、时间、压力、配料比例等等都比较容易准确地控制,各种偶然因素对实验结果的干扰较小,所以,工业统计中有时允许以较少的观察例数就得出统计分析结论。但是,体育统计研究的对象是人,人的个体差异是很大的,同性别,同年龄的人身高、体重、运动成绩的差异都较大,由于存在个体差异,所以,相同的训练方法,相同的运动量对每一个人的训练效果就不会完全一样。因此,在研究教学、训练的效果时,观察的人数太少了就不合适。又如,统计分析时必须注意“同质”,即被研究的对象应是性质相同的,否则统计分析的结果就不能说明问题。例如,研究某种发展弹跳力的训练方法,虽然研究的对象是同性别、同年龄的人,但是还必须考虑他们的运动水平是否相同,因此常常还要选择运动等级相同的人才行。假如把一般学生和优秀运动员作比较就不对了。

2. 要有严肃的科学态度。对待统计的数据和计算结果必须实事求是,一丝不苟,来不得半点虚假。有的人在取数据时,为了希望说明自己的预定结论而弄虚作假,有的人在测数据和计算时粗枝大叶,这样算出的数据不真实、不可靠,根据它作出的结论是会误人的。

第一章 统计资料的收集与整理

收集资料、整理资料和分析资料是统计工作的三个基本步骤。

收集资料之前必须根据研究目的制订研究设计，然后再按设计的要求去收集完整、准确的原始资料。这是进行统计工作的基础。

整理资料是对收集到的原始资料进行审核、分组、归纳汇总，以便于统计计算。这是保证统计结果的准确性不可缺少的一步。

分析资料是对经过整理的原始资料计算相应的指标和绘制必要的统计图表，结合专业知识进行分析、比较、找出其中规律性。

以上三个基本步骤是互相有联系的，任何一步有了缺陷或差错都会影响研究结论的正确性。

第一节 统计资料的来源

体育统计资料的来源主要有三方面：

一、日常工作中积累的资料。体育教师和教练员在平时上课、训练、开展群众性体育活动和组织运动竞赛中都会积累许多宝贵的数据。如训练时测验的运动成绩，体育课考核的各项数据，运动会的比赛成绩等等。

用这一类数据进行统计前，要特别注意数据的准确性如何？有时上课记录的数据笔迹模糊，有时不同班级测试时条件不相同，如一个班跑100米时正巧刮大风，而另一个班跑100米时天气很好，这些数据究竟能否用于计算就要注意审核。

二、全面普查。如一年一度对学生的“体质健康卡片”普查。这种全面调查，工作量比较大，而学生又往往由于某种原因，容易漏测某些项目，甚至有时还有卡片上忘了填写姓名等等情况，因此，事先要有周密的安排，做到测试时不忙乱，不出差错，并且在测完最后一项，回收卡片时要有专人负责逐项审查，发现漏填或漏测的项目就要督促学生及时补上，以免造成整理卡片的困难。

三、专题研究。这包括专题的调查研究 and 专题的实验研究。像1979年全国青少年儿童体质研究就是我国首次进行的大规模的专题调查研究。其目的是为了摸清我国青少年儿童的体质现状、特点及生长发育规律，有时为了研究某种新的训练方法，就要组织实验组和对照组，通过一定时间实验，取得数据，对比新训练方法的效果。

在进行专题研究前应该根据研究工作的目的、要求，结合统计原理的要求对研究工作的全过程事先制订一个周密的工作计划，通常把制订这样的计划称为研究设计（或课题设计）。如果研究设计考虑得比较周密、完善就能以较少的人力、物力、时间获得可靠的资料，使误差减小到最低限度，提高研究工作的质量和效率，并有助于得到科学的可靠的结论。反之，事先不作研究设计，匆匆忙忙收集来一大堆数据就进行统计计算，往往是不可能得到可靠的结

论的。

研究设计的基本内容大致应包括以下几方面:

1、研究目的。围绕着增强体质和提高运动技术水平,有许多大大小小的研究课题需要我们去研究。一般讲一项研究工作不宜涉及过多的问题,最好只要求解决一、二个问题,否则容易影响工作的质量。

2、研究对象和人数。包括抽样的方法及分组的办法等。如果本研究课题是为了比较不同的教学、训练方法的效果,就应该设立实验组和对照组,这两组的人数和实验前的水平应该大体相仿。

3、研究时间和方法。包括:研究工作开始和预计结束的时间,在整个研究时间内是否划分成几个阶段,研究工作开始时搜集哪几项统计指标,研究过程中每隔多少时间再次搜集哪几项统计指标,对这几项指标用什么方法、什么仪器设备进行测试也应同时订出详细的测试细则。应该特别重视指标的设计。要使每一项指标所测得的结果,都能比较客观地、真实地反映出测试对象某一方面的情况。

4、各种原始记录表格的设计、制订。

5、初步规定数据的整理、统计分析方法。

下面,讲几个与制订研究设计有关的概念。

(1) 总体、个体与样本

总体:是指某一项研究的全体对象。如研究湖北省7~25岁青少年儿童的体质情况时,总体就是湖北省7~25岁的全部青少年儿童,研究某校男生100米跑情况时,总体就是该校的全体男生100米跑的数据。

个体:是组成总体的最小研究单位。在体育研究中,通常就是指人。

样本:有时总体很大,要对其中每一个个体进行观察是不可能的,有时由于人力、物力、时间的限制不允许对总体内的每一个体进行观察,因此,要用适当的方法从总体中抽出一部分个体来进行观察、研究。被抽出的这一部分就叫做样本。样本一定要能很好地代表总体的特征,形象地说,样本应是总体的“缩影”才有研究价值。要做到这一点,就必须严格按照“随机抽样”的原则——即总体里各个个体都有同等的被抽取的机会,抽样时绝对不能事先抱有丝毫主观偏见。

(2) 样本的含量

样本内所包括的个体数称为样本的含量,通常用 n 表示。习惯上认为 $n < 30$ 为小样本, $n > 30$ 为大样本。体育运动中所测得的数据由于人的个体差异较大,所以在条件许可时,样本最好尽量大一些。但是其前提是要保证所测数据的准确性,如果单纯追求样本大一些,结果测试的人数增加了,却会使得一些数据测得不够准确,那么这样的大样本统计出来的结论也是没有什么意义的。倒不如样本略为小一点,把每一个数据都测准确更好。当然样本太小了也不行的。由于样本小了抽样误差就会变大,因此,例如,用甲方法训练是比用乙方法训练成绩提高得很多,但是如果在进行实验时,所取的样本太小,计算出来的结论就可能认为两种方法的差别不显著。所以,在制订研究设计时,要根据本课题的人力、物力、时间等多方面的情况确定样本的含量。样本含量 n 的大小还应该根据研究的课题所允许的误差大小而决定,其计算方法可见第53页。

(3) 随机抽样

为了获得对总体有较好代表性的样本,并排除选定样本时的主观偏见,一定要按“随机化”原则抽样,其方法有:

①抽签法:事先把总体内各个体按顺序编号,然后抽签决定被抽出的个体。如全校男生共有800人要抽出80人,则先把全校男生编成为001~800号,然后抽签,把抽中的80个号码所代表的人组成样本。

用“随机数表”或fx-180p计算器的随机数发生器作抽签法抽样比较方便。(其使用方法见后)

②机械抽样法(亦称系统抽样):事先也要把总体内各个体按顺序编号,然后按固定的间隔抽样。如在800人中要抽出80人,即抽出 $1/10$,因此,可以抽签决定一个数,假设是“3”,则3、13、23、33……直至793号都抽出来组成样本。这种方法比较简便。

③分层抽样法:即先把总体按某种特征分成若干部分(称为层),然后在每一层内随机抽样。如把全校800人分成一、二、三、四年级,再在各年级内随机抽样 $1/10$,这种抽样方法抽样误差比较小。

④整群抽样法:即先把总体分为若干个集体(称为群)随机抽出几个群组成样本。如:全校800人共为25个教学班,抽样时规定共抽3个整班组成样本。这种方法在进行大规模调研时,便于组织可以节省人力、物力,但抽样误差较大。

在制定研究设计时,可以根据具体情况决定采用哪一种抽样方法,有时也可以把几种方法结合起来。例如要从某校随机抽样100人可以规定先用整群抽样法在全校抽出10个班,然后再用“随机数表”按抽签法从每班抽出10个人组成样本。

随机数表使用方法:

随机数表见150页是由0到9的数字随机排列而成的,每页有 $50 \times 50 = 2500$ 个数字,排成50行、50列,可以根据需要把它当成任何位数来使用。随机数表的用处很多,这里只介绍用它进行随机抽样和随机分组的方法。

无论抽样或分组,首先要决定随机数表的起始点和前进方向。

1、起始点由闭眼指表三次来决定页、行、列。

①决定页:闭眼指表,指到的数字是奇数则用第一页表,指到偶数和0则用第二页表。

②决定行:闭眼指表,按指到的二位数决定。如:指到38,则由上向下数到第38横行即为这次抽签的起点行。假如指到的数大于50,则应减去50,如指到75,则 $75 - 50 = 25$,应把第25行定为起点行,假如指到00,则把第50行定为起点行。

③决定列:闭眼指表,按指到二位数决定。如:指到23,则应由左向右数到第23列定为起点列(指到的数大于50时,也应减去50)。

假设:闭眼指到6,决定用第二页表;

再闭眼指到47,决定用第47行;

再闭眼指到57,决定用第7列。

那么,第二页表第47行第7列的数“7”就是这一次抽样或分组的起始点。

2、取数的前进方向按下列原则决定:被抽的总数是一位数或二位数时,由起始点向右取数,到达最右端后下移一行继续由左向右取数,被抽的总数是三位或三位数以上数时,由起始点向下取数,到达下端后移至下一列的上端继续往下取。

现在用上面决定的起始点(2页47行7列)举例说明前进方向和取数分组的方法。

(例一)有80人已按顺序编成01~80号,需要随机地从80人中抽出20人。

因为被抽的最大数为80是二位数,所以,从起始点二位、二位向右取数,把大于80的数不要,重复的数也不要,则抽中的20个号码是:79、80、24、36、59、38、07、53、35、23、18、05、46、40、62、54、20、56、15、74。

(例二)有200人已按顺序编成001~200号,要从中随机抽出20人。

因为被抽的最大数200是三位数,所以,从起始点向下取数,从表上可见起始点的7和其右边两列的数组成798,向下为988、034,……,去掉抽出的大于200的数就可得到034、055、045、024、166……,直至抽满20人止。由于三位数中大于200的数很多,所以也可事先规定,由起始点开始抽取的数如果大于200均被200除,余数即为被抽中的号码,这样去掉重复的数后就应抽出:198,188,034,055,161,099,006,165,001,171,048,101,151,057,137,060,045,051,119,013。

(例三)有40人已按顺序编成01~40号,要随机地将他们分至甲、乙两组。

由起始点向右取数。现规定第一个数代表1号抽到的数,第二个数代表2号抽到的数……,如抽到的数为奇数该号就编入甲组,该数为偶数即编入乙组。按此规定这40人就可以分成:

编 号	1	2	3	4	5	6	7	……
组 别	甲	甲	乙	乙	乙	乙	甲	……

40人全部分完后,如两组人数不等
可再随机地抽出多余的人至另一组补齐。

Casio fx-180p电子计算器上随机数字发生器使用方法:

在计算器的小数点键下面印了红字符RAN#,这是英文RANDOM(随机)的缩写。所以,每一次按INV键后再按小数点键,计算器上就可以发出一个0.000至0.999之内的随机数字,如果不看小数点,可以把发出的随机数字看成是0至999的抽签号,则每一次按出的数,就是随机抽签的号数了。如果利用fx-180p计算器上编程序的功能还可以更方便地用于随机抽签。(详见fx-180p计算器统计运算实用手册)

第二节 统计资料的收集与整理

按照研究设计的要求去收集资料时,最重要的是要保证测试数据的准确性,例如给学生进行“体质健康卡片”的测试时,是由许多人分工进行测试的,有的人负责测身高,有的人负责测50米跑等等,或者同一个测试指标是由几个人负责测的,甲测这一班学生的立定跳远,乙测另一班学生的立定跳远。为了使测得的数据尽可能地减小误差,首先必须按拟定的“测试细则”统一测试时的操作方法、记录方法,测试前还应严格校正测试器材(如身高计、秒表、皮尺等)这样取得的数据才比较准确。假使对测试的要求不统一,有的人称体重时严格规定受测者只能穿短裤,有的人又允许受测者不脱外衣称体重;或者对同一批运动员测100米跑成绩,开始时测验用的秒表和练习了三个月短跑后再测验时用的秒表快慢不同未经校对,用这样的数据计算出来的结果能说明什么问题呢?

即使是某一位教师自己作教学试验,所有的数据都是他一个人收集的,也同样应该在制订研究设计时,拟定出测试细则以保证数据的准确可靠。

特别应该指出,体育统计中所要测的运动成绩,和受测者当时的心理状态有很大的关系。测学生的身高、体重等形态指标时,如果受测学生积极性不高不愿测试,只要测试人员严格按照规定操作,还是可以测到准确的身高、体重数据。但是,如果测试100米跑,受测者当时积极性不高,他不用全力跑,尽管发令、计时都严格按规格操作,但测得的成绩就可能和该学生的真实成绩差很多,因此,收集身体素质、运动成绩的数据时,一定要设法调动受测者的积极性,要求他们和测试人员紧密合作以便测到反映真实情况的数据。这除了要做好宣传动员工作外,还应考虑一些具体措施,如:把测试和体育课考核、运动竞赛结合,这样测得的成绩就比较准确,有一些测试的规则也应根据实际情况制定,例如:跳远按田径规则预赛跳三次,某运动员前二次踏板时犯规,不计成绩,他第三次为了怕再次踏出线犯规,就在离起跳板较远的地方起跳,这时测出的成绩看起来是严格按田径规则,其实却没能反应出该运动员的真实水平,假使把测试规则改为犯规不计成绩但可重跳一次,以二次中最好的一次成绩为准就会更准确一些。另外测试人员在测试时发现某人的成绩明显地不可靠时,要及时让他重测,如当时无法重测,就应该在数据卡片上加以注明。

收集受测者的机能数据,如心率、血压等,也要十分注意受测者当时的情况,因为有些指标,受情绪的影响很大,如情绪紧张时心率加快,血压也会升高。这些指标还受到当时身体状况的影响,如刚进行过运动,疲劳没有充分恢复,心率就较快,这些也是必须十分注意的。

在收集各种统计数据时,尽管测试人员主观上是严格按照研究设计规定的要求认真测试的,但收集到的原始数据与其真实情况之间仍然会存在或多或少的差别,称为误差。

误差可分以下几种:

1、随机误差:它是由一系列偶然因素引起的一类不易控制的测量误差,在测试过程中是难免的,其误差值可正可负,可大可小,具有统计的规律性,随着样本的增大,随机误差的算术平均值可愈来愈小而渐近于零。

2、系统误差:它是服从确定性规律的误差,多数情况下系统误差是一个常量,它不可能通过扩大样本而消除。例如跳远时丈量用的皮尺,因为使用时间过久,已经拉长了,和标准的钢尺校对每米差2厘米,那么凡是用这一皮尺丈量出的跳远成绩,就存在着系统误差(每米2厘米)。系统误差要在测试前通过校正仪器设备,以及在测试过程中通过一定方法识别后加以清除。

3、过失误差:这是在测试过程中的过失造成的对数据的歪曲,如记录成绩时由于笔误把13.7“误写成17.3”,或由于读数的错误把跳远5.50米误读成4.50米等。过失误差造成的误差数有时会很大,它对计算结果的准确性的影响比其他几类误差要大得多。因此,测试人员工作时一定要高度认真负责,还应遵守一定的操作规定,加记录员记录时应复诵记录的成绩,记录时必须字迹端正等。

4、抽样误差:这是抽取的样本与总体之间的差别。即使采用随机抽样的方法仍不能避免的,但当样本含量加大后,可使抽样误差端小。

由于收集到的统计资料中存在着以上误差,所以,必须经过审核、整理,把其中错误的、可疑的数据剔除或更正后才能进行统计计算。制定研究设计时,应该规定出对原始数据的审核、整理办法。大规模的专题研究、“体质健康卡片”测试时都应有严格的数据审核、验收制度以保证数据的可靠性。

审核原始数据大体可分以下步骤:

(1) 仔细地检查全部原始记录表格或卡片。首先要核对姓名、性别、出生年月等项是否填写清楚。然后要逐栏检查测试数据是否有“缺、疑、误”,缺是指缺项未填,疑是指记录的数字由于书写不端正而不能明确辨认或对记录的数据真实性有怀疑,误是指明显存在的差错。

发现“缺、疑、误”应该及时处理。缺项应补填或补测,如无法补就应考虑此人其他数据是否参加统计。字迹不清楚的数据要经几人共同辨别确认后填写清楚,如无法确认时就不能使用,在改正“疑、误”时,一定要严肃认真,不能主观估计、臆造,随便填写没有事实根据的假数字。

检查“缺、疑、误”应该反复进行几遍,特别在原始资料很多时,只进行一、二遍检查往往还会有遗漏的。

(2) 逻辑检查。对原始资料进行初步的“缺、疑、误”检查后还应进行逻辑检查,以便进一步找出存在的误差。利用已知的测试指标之间的比例关系可进行检查。如:某学生身高为165厘米,坐高为65厘米,初看时数字记录得很清楚好象没有差错,但因为已知人的身高和坐高的比例约 $1/2$,而该生相差很大,就可肯定必然有错。又如一般学生的100米跑成绩中发现跑10秒的人,也必然可以怀疑数据错了。

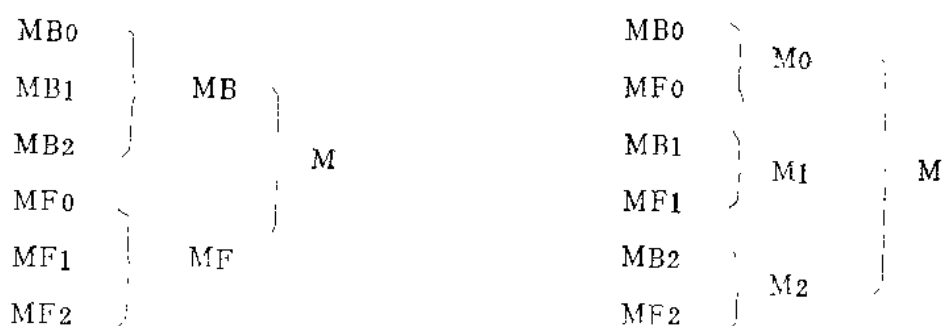
大规模的专题研究为了便于进行逻辑检查,可参考过去的有关统计资料制订一个检查标准,如:最大值至最小值的范围,当发现某一数据超出此范围时就作为可疑,对其进行仔细分析,必要时可作复测。也可以参考本人过去的统计数据进行检查,如某学生一年前进行测验100米跑成绩为14",现在又进行测验成绩为15"3,退步一秒多,就应考虑为什么退步?是一年来未锻炼所以成绩退步了,还是测试时未认真跑?

(3) 抽查复核。在经过几遍全面检查,逻辑检查之后,最好再抽查复核。可规定抽查 $1/10$, $1/50$ 等,对抽出的卡片再仔细的检查,确实无误了才可验收,如发现误差则应对全部原始资料重新检查一遍。

统计资料的审核、整理是一项要求十分细致的工作,必须以高度的责任心认真地进行,千万不可粗心大意。

第三节 统计资料的分组

经过审核整理可以进行统计计算的原始资料,还要根据研究设计的规定进行分组整理。首先应该将同质资料归纳在一起,例如按性别、年龄分组,有时还需要按运动员等级分组或按运动项目分组等等。总之,要根据研究设计中统计的要求,先将原始数据合理地分成若干组使得需要分类汇总统计时,可以很方便地进行各种组合,既方便又不重复劳动。在分组较多时,最好对各组的资料进行统一编号。例如,测了一批射击运动员的资料,其中按运动等级分有健将、一级、二级;按项目分有飞碟、步枪;按性别分有男、女。可以按性别、项目、等级的不同分成若干组。然后再按项目、等级进行合并统计,为此,统一编号:第一个字符为性别,第二个字符为项目,第三个字符为等级。如:MB0表示:男步枪健将,WF1表示:女飞碟一级等等。下图表示男子分成6个组后,可以用不同的办法组合,汇总统计。



在质量分组的基础上还要进一步按数量大小分组，才能从原始数据中看出某些内在规律。通常是用“频数分布表”进行分组整理，下面举例说明分组方法：

(例四) 有一组18岁女生身高的数据 ($n=120$)，数据如下，用频数分布表进行分组整理。

147.5	153.0	149.0	150.2	151.7	152.9	154.0	155.3	155.5	156.3
156.7	149.9	157.1	154.1	155.1	159.7	160.0	153.5	160.1	160.3
159.1	150.8	161.0	155.0	156.8	153.7	152.7	160.4	154.6	160.7
156.9	160.3	160.5	151.0	157.3	161.2	157.3	160.9	161.3	161.2
161.7	155.9	161.4	153.2	152.1	162.0	154.9	162.8	163.0	163.0
162.0	155.9	151.1	154.7	158.5	156.2	158.1	159.9	163.7	165.1
152.5	164.9	153.4	156.1	159.7	164.0	157.9	165.5	158.0	165.8
156.9	165.1	158.5	157.1	152.0	164.1	164.2	162.7	166.6	166.7
159.5	157.0	168.2	154.3	158.4	168.0	167.2	167.9	160.8	169.4
168.5	159.0	157.0	154.9	159.0	156.9	158.0	162.1	172.1	161.0
158.5	165.0	159.3	161.0	159.1	166.7	165.9	162.9	165.0	168.2
164.7	158.6	159.3	159.0	163.1	163.9	162.8	169.7	170.0	169.1

分组步骤：

①找最大值，最小值。从上述数据中可找出最大值为172.1，最小值为147.5。

②计算极差。

极差 = 最大值 - 最小值 = $172.1 - 147.5 = 24.6$

③确定分组数和每组的组距。确定分组数时根据 n 的大小可参考表一确定分组数。

表一 分 组 数 参 考 表

n	50	80	100	150	200	300	500	1000	2000	5000
分组数	9	11	12	14	16	18	22	30	39	56

本例 $n=120$ ，现确定分12组，计算组距： $\text{组距} = \frac{\text{极差}}{\text{分组数}} = \frac{24.6}{12} = 2.05$ （厘米）

根据上述计算结果将组距凑成整数，所以，定为2厘米（这样分组数就会增加一组）。

关于确定组距的方法，有时并不采用上述计算方法，而是事先早作规定，如1979年中国青少年儿童体质研究统计计算时规定各年龄组频数分布表一律统一组距为：身高2厘米，体重1公斤，肩宽0.5厘米，胸围1厘米，仰卧起坐2次/分，60米跑0.3秒，立定跳远5厘米，屈臂悬垂4秒，400米跑4秒等。

④在频数分布表（表二）上填写组限栏。本例因为已找出最小值为147.5，所以第一组可以从147开始，写成：147~，又因为组距已定为2厘米，所以第二组为149~，如此共填写13个组，到171~时由于本例最大值172.1已可分在这一组内，所以下面不要再设组了。

⑤把原始数据用划“正”字的方法，将120人分别划记到各组去。如第一人身高147.5应划记在147~的一组，第二人身高153.0应划记在153~的一组内，全部120人都一一划记完后，数一下每一组内划记的数目，填在频数（f）栏内。

⑥把各组的频数（f）加在一起应该等于样本的总数，即 $\sum f = n$ ，如不等就应该重新复核。

通过频数分布表对原始数据进行了数量分组整理后，就可以从中看出一些规律，如从表二我们可以看出18岁女生120人中最高的、最矮的在哪一组？在哪一组中人数最多等等。

按照频数分布表上各组的频数，可以画出一张直方图（图一）。在图的横轴上标出各组的组限，在纵轴上标出频数，根据各组的f数可在图上画出代表各组频数的直方条。本例经计算平均数 $=159.7$ （计算方法详后）在图上以虚线代表。

从直方图上我们可以看到：

①在平均数附近的组分布的人数多，越离开平均数远的组，组内人数越少。

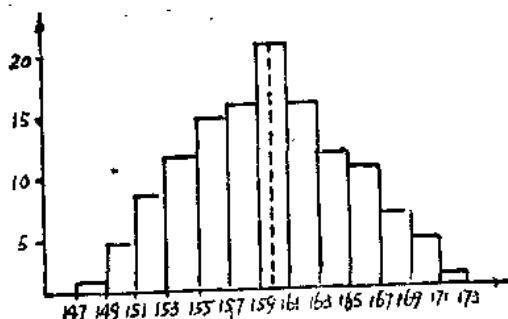
②平均数两边的各组人数分布是接近于对称的。

类似这种“中间多、两头少”的分布情况，在统计学上称为：正态分布。如果把频数分布的直方图上各直条顶端的中点连起来，就可以成为一条钟形曲线，它与数学上的正态曲线形状相似（图二）。

在体育统计中，过去人们一般都认为取得的数据只要样本含量不是太小，又是随机抽样的，则数据的分布都是接近于正态分布的。但是，近年来的一些研究表明体育领域中的各项

表二 频数分布表

组 限	划 记	f
147~	—	1
149~	正	4
151~	正下	8
153~	正正一	11
155~	正正正	14
157~	正正正	15
159~	正正正正	20
161~	正正正	15
163~	正正一	11
165~	正正	10
167~	正一	6
169~	正	4
171~	—	1
	Σ	120



图一

指标并不都是正态分布。用频数分布表作直方图后可以从图上直观地看出数据的分布是否接近正态分布。但是，有时由于分组的方法不同，用直方图来判断是否正态分布也会出现错觉的。例如表三、表四是同一个60米跑的样本，在作频数分布表时，组距也都是0.2秒，只是表三的第一组下限定为7.8"，而表四的第一组下限定为7.9"，这样各组的频数就不同了。从表四上看频数分布是接近正态分布的。但从表三上看8.6~的一组 $f=18$ ，明显地少于其相邻的两组，好象不成正态分布了。



图二

这说明单凭直方图有时会有错觉。数理统计有许多计算方法可以检验数据是否趋近于正态分布，有一种用“正态概率纸”作图检验的方法，比较简便，现介绍如下。

正态概率纸的横坐标是均匀的刻度，纵坐标则是按正态分布的规律刻划的。作图时，先要把频数分布表上各组频数进一步计算出累计频率，然后按累计频率在正态概率纸上打点，如果频数分布是接近正态分布的，则打出的点就几乎在一条直线上，如果有的点子偏离太多了，就认为这批数据不成正态分布了。

下面以表二之数据说明作图方法。

第一步：把频数分布表上的频数填在累计频率表上（表五）。

第二步：计算累计频数。这是到某组为止的频数之和，如：到151~组的累计频数为 $1+4+8=13$ 。因此，最后一组的累计频数一定应该等于样本含量 n （本例为120）。

表三 60米跑频数分布表

组 限	划 记	f
7.8~	—	1
8.0~	正正丁	12
8.2~	正正正	15
8.4~	正正正正正丁	27
8.6~	正正正下	18
8.8~	正正正正正下	28
9.0~	正正正正	20
9.2~	正正正丁	17
9.4~	正正	10
9.6~	正丁	7
9.8~	丁	2
10.0~	—	1
10.2~	—	1
Σ		159

表四 60米跑频数分布表

组 限	划 记	f
7.9~	下	3
8.1~	正正正丁	17
8.3~	正正正下	18
8.5~	正正正正下	24
8.7~	正正正正正下	28
8.9~	正正正正下	23
9.1~	正正正正	19
9.3~	正正	10
9.5~	正正	9
9.7~	正	5
9.9~	—	1
10.1~	丁	2
Σ		159

表五 累 计 频 率 表

组 限	频 数	累 计 频 数	累 计 频 率 [%]
147~	1	1	0.83
149~	4	5	4.17
151~	8	13	10.83
153~	11	24	20.00
155~	14	38	31.67
157~	15	53	44.17
159~	20	73	60.83
161~	16	88	73.33
163~	11	99	82.50
165~	10	109	90.83
167~	6	115	95.83
169~	4	119	99.17
171~	1	120	100.00
Σ	120	—	—

第三步：计算累计频率。其公式如下：

$$\text{累计频率} = \frac{\text{累计频数}}{n} \times 100\%$$

$$\text{如：} \frac{1}{120} \times 100\% = 0.83\%$$

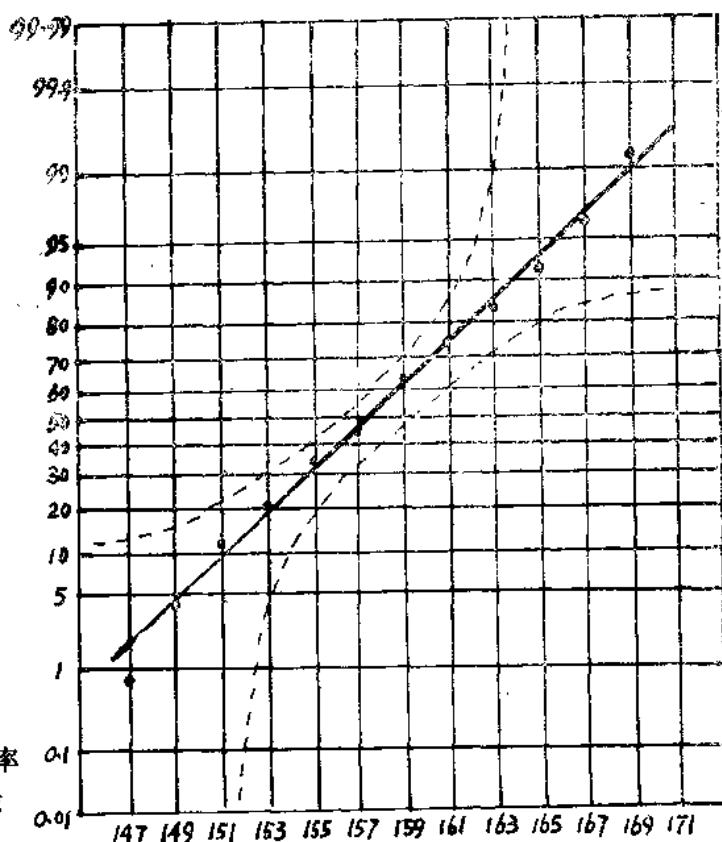
$$\frac{5}{120} \times 100\% = 4.17\%$$

第四步：在正态概率纸上以组限为横坐标，累计频率为纵坐标依次作点（见图三）由于在正态概率纸上不能作概率为0或1的点，所以“171~”这一组的点就不作了。

第五步：在纸上用直尺画一条线，画线的原则是要使这条线离图上各点的距离尽可能相近，且使直线两侧的点数尽可能相同。

由于抽样的随机性，用样本数据作出的点不可能和理论上的正态分布完全一致，所以这些点就不可能恰好成一条直线，只要点子偏离直线不远就行了。那么允许偏多大的距离呢？

可以根据柯尔莫哥洛夫（Колмогоров）检验法在图上的直线上下画出两条弧线，只要有一个点子落在这两条弧线形成的区间之外，就认为这个样本不服从正态分布。



图三 用正态概率
纸作检验