

● 徐迪生

# 育统计讲义

学校体育编辑部

# 体育统计讲义

(普及本)

徐迪生

学校体育编辑部

## 前 言

随着学校体育工作的发展，学校体育的科学研究工作空前地活跃起来了。广大学校体育工作者在科学研究过程中，都经常遇到实验设计和数据的统计处理与分析等问题。如何科学地进行实验设计，如何统计处理和分析实验所获得的各种数据（如身体形态、机体能力、心理状态、运动能力、疾病等测验数据），不仅直接影响着科学研究的水平，而且还可能决定着科学研究的成败。因为实验设计错误，就势必把整个科学研究工作引入歧途；而实验所要获得的各项指标，是通过大量的测验数据反映出来的，大量的测验原始数据，如不经过正确处理和分析，就无法显示出有价值的信息，因而也就难以有力地论证问题。体育统计学，作为一门工具课，就是为广大读者提供一种从事体育科学研究的方法。

《体育统计讲义》是在《学校体育》杂志刊载的“体育统计学讲座”的基础上，增改部分章节汇编而成的。本讲义已定为北京体育学院大专班和干部、教练员进修班教材，讲授时间三十二学时。本讲义也可作为三周全脱产的短训班或中专教材。

作 者

1985年7月

A276/20

# 目 录

|      |                 |        |
|------|-----------------|--------|
| 第一讲  | 几个基本概念          | ( 1 )  |
| 第二讲  | 平均数与标准差         | ( 5 )  |
| 第三讲  | 正态分布            | ( 13 ) |
| 第四讲  | 均数的抽样误差及 $t$ 分布 | ( 19 ) |
| 第五讲  | 评分的方法           | ( 30 ) |
| 第六讲  | $U$ 检验与 $t$ 检验  | ( 43 ) |
| 第七讲  | “率”             | ( 48 ) |
| 第八讲  | 直线回归与相关         | ( 55 ) |
| 第九讲  | 多元回归分析          | ( 68 ) |
| 第十讲  | 判别分析            | ( 77 ) |
| 练习解答 |                 | (102 ) |

## 第一讲 几个基本概念

什么叫**体育统计**？体育统计是数理统计在体育科学中的应用。数理统计的基础理论是概率论。概率论是研究偶然事件（又叫随机事件）发生的规律的一门学科。所谓偶然事件是指在相同条件下，可能发生也可能不发生的事件。例如投掷一枚钱币，落地后恰巧正面向上就是偶然事件，又如跳高中试跳某一高度，可能成功也可能失败，所以也属于偶然事件。偶然事件的发生与否是受许多客观因素影响的，这些因素多而复杂，相互联系，相互影响，而又找不出哪一个因素是决定性的因素。因此偶然事件的发生与否事先难以确定，凡是事先没有百分之百的把握必然要发生的事件，都称为偶然事件。如果某个事件，事前就完全可以肯定它一定会发生，那么这种事件就称为必然事件了。

为什么体育科学可以应用概率统计的方法加以研究？原因正是由于体育活动中需要研究的现象，绝大多数都是偶然事件。例如身体素质的提高，运动成绩的出现，以及比赛的胜负等，由于这些事件的发生是受许多客观因素影响的，情况极为复杂，尽管经过严格的科学训练，在计划时间内达到预定指标的可能性极大，但仍存在极小可能性实现不了预定指标，也仍然是属于偶然事件。因此，研究偶然事件的概率统计方法，就可以在体育科学研究中应用。

偶然事件的发生，并非是不可捉摸的，虽然我们不能准

确地预见某一次偶然事件发生与否，但人们在长期的实践中发现，在偶然事件的大量重复中，往往出现几乎必然的规律，这就是大数定律。例如，抛掷钱币，就某一次来说出现正面还是反面，虽然是偶然的，但在大量重复抛掷时，出现正面和反面的次数与抛掷总次数之比，都必然接近于确定的数值，即各占  $1/2$ 。这  $1/2$  就是抛掷钱币时出现正面或反面的概率。那么从一副扑克牌中抽取一张牌，抽得红桃的概率为  $13/52$ ，即  $1/4$ ；抽得老K的概率为  $4/52$ ，即  $1/13$ ；抽得红桃A的概率只有  $1/52$ 。因此每次抽得哪一张牌，虽然无法事前确定，但反复多次抽取后，将会确定某一张牌出现的概率。在体育活动中某一事件的概率并不象上述那样简单得到，需要通过大量的实验，这叫实验概率。如果某一学生试跳某一高度，在100次试跳中有80次跳过去了，我们讲他有80%的把握能成功，这就是成功的概率。习惯上的“十之八九能成功”，“八成是成功了”……这都是根据经验指出某一事件成功的概率。

以上说明偶然事件中蕴藏着必然的规律，我们应用研究偶然事件的规律的数理统计方法来研究人体的各种指标出现的规律是完全可行的。

**总体和样本：**根据研究目的确定的研究对象的全体称为总体；从总体中抽取的有代表性的一部分称为样本。例如要调查全国学生各项运动水平如何，以便制订适合于全国大、中、小学不同年龄组的测验标准，那么这一研究的对象和内容就是总体。这么大的总体（有的总体更大或是抽象的）是很难取到的，因此，只有从总体中抽出对总体有代表性的若干个省市的部分学生进行测验，这叫“抽样”。抽出

的这一部分人测验得到的数据叫“样本”。样本中的数据叫“变量”，用 $X$ 表示。样本所包含的变量个数叫“样本含量”以“ $n$ ”表示。

如果研究的对象是十个学生的跳远水平，即使每人跳一次，得到的成绩仍然不是总体。因为再跳一次得到的成绩与第一次成绩不同。究竟这十个学生客观存在的水平（即总体的“真值”）有多高，是取不到的，每人多次试跳所得到的测验成绩仍然是样本。

我们研究的对象是总体，但我们能得到的只是样本。因此，我们只能通过样本对总体情况做出估计。这是统计学中一个主要内容。

样本既然是估计总体的根据，就要求样本对总体有较高的代表性，样本情况愈接近总体情况，它的代表性就愈高。要想提高样本的代表性，就需要有科学的抽样方法：一是抽样的原则要随机化，即不加选择的抽取。例如，要了解某校初三男生百米水平，准备抽50名进行测验，这50名学生怎么确定呢？可以采取抽签、抓阄的方法；也可以按学号逢五便取，或是在这一年级的几个班中各随机抽取几名都是符合随机原则的。如果有意挑50名好的或差的，那就不能作为这一研究的样本。二是增加样本含量。即从总体中抽得的个体愈多，它对总体情况的代表性也就愈高。一般讲总体愈大样本含量也应愈大；总体中个体之间差别愈大，所采取的样本也应愈大。具体取多少合适，要根据实验要求来确定。

**抽样误差：**由于总体中的个体之间存在着差异，又由于样本仅是总体中的一部分，因此样本的统计量（如样本中各变量的平均数）与总体客观存在的“真值”（如总体中各变

量的平均数)也就存在差异。即使从一个总体中随机抽取相同含量的两个样本,也不会相同。这种由于抽样引起的差别称为“抽样误差”。抽样误差是无法避免的,但这种误差有它一定的规律,研究这个规律,应用统计学方法可以计算出误差的范围和它的概率,从而对总体的“真值”做出估计。其它如测验方法存在缺陷、秒表不准、计算错误、登记错误,都属于工作过失,在实验和统计中应注意防止。没有可靠的数据,无论怎样统计处理也得不到正确的结果和结论,这一点是至关重要的。

## 练 习

1. 以下事件为什么是偶然事件?

(1) 一次图钉落在地上尖向上。

(2) 某次运动会跳高创造1.80米纪录。

(3) 近代世界人口总是在增加。

(4) 我国在第36届乒赛中取得七项冠军。

2. 以下概念错在哪里?

(1) 由于严格遵守随机抽样的原则,因此才能使样本计算得到的某个数值与总体的“真值”相等。

(2) 去年某县中学生运动会100米跑第一名成绩为12"4,今年为11"9,说明全县中学生100米跑水平有大幅度提高。

(3) 我们可以将一所学校篮球代表队看做是从该校抽得的样本。

(4) 从城区、郊区各随机抽得1000名10岁女童体重的

平均数：城区为30.4公斤，郊区为29.6公斤。说明城区10岁女童体重大于郊区。

## 第二讲 平均数与标准差

### 一、怎样收集和整理材料

当我们准备研究某一个问题之前，首先要阅读有关的资料或书刊，了解他人对此问题是否有所研究，研究到什么程度，有哪些论点和争论，已有什么结论，然后才能设计研究方案。研究问题要通过自己的实践或实验来论证才有意义，才会得到有价值、有创造性的结论。研究要有对象，体育教师主要是进行学校体育工作研究，研究对象是有组织的学生，这是从事研究的非常有利的条件，教师可以针对某一个课题（如体育课、课外体育活动、达标活动等等），设计实验方案，制定测验方法，有目的地收集材料（数据）；也可以经常、不断地积累资料（如身体检查结果、考试或达标测验的成绩等等）。

由于研究问题是多种多样的，因此不可能提出一个统一的实验设计方案。但实验设计应遵循基本的原则和方法。

实验设计的原则是要求以尽量少的实验次数，尽量用简易而准确的实验方法来获得足够、有效的资料，从而得到较可靠的结论。实验设计要从这个要求出发来制定最优的实验方案。因此实验不必过于追求复杂的仪器和繁杂的方法。在

体育测验中有不少公认的方法，如田径中的比赛规则、视力测定表等。这些都可以作为实验的方法。

**比较研究是研究中不可少的方法。**有比较才能有鉴别。把实验对象分为实验组和对照组是比较研究中常用的基本方法。在实验过程中对两个组用不同的处理方法来影响它们，一般是对实验组施以所要研究的处理方法，而对照组则有几种：“空白对照”，即不加任何处理，听其自然发展，例如要了解开展锻炼标准活动对于学生健康的影响，空白的对照组可以不开展这项活动，按原有的活动自然发展；“标准对照”是指运用已知的正常值，如正常男子的心率为72次/分，研究长跑锻炼的学生心率是否低于72次/分；“实验对照”是对实验组和对照组施以不同的处理，以比较两种处理结果的差异，例如施以两种不同的教法。

**在实验中要控制非实验因素的干扰。**即除实验因素不同外，其它因素要一致，否则无法判断两组的差别是实验因素不同引起的还是其它因素引起的。例如研究两种跳高教学的效果，不但两个组原来跳高的水平应要求一致，身体素质基本相同，并且教学时间、条件也要一样，这样，经过一段时间后产生不同的效果，才利于确认效果的差异是由于不同的教学方法引起的。

**实验的数据要符合统计要求才能应用统计方法处理。**前边讲过取得数据的要求是随机化，并且要考虑有足够的数量。实验数据愈多说明实验次数也愈多，得到的实验结果可靠性就高。如某个学生的跳远测验以试跳十次的平均成绩比较只跳一次的成绩更能代表他的跳远水平，从这个角度考虑数据自然是愈多愈好，但测验工作量和统计工作量都加大，

因此取得实验数据要根据实际可能，结合专业知识和统计学的要求来确定。

实验数据取到后，需要整理成有条理的、有系统的资料。“频数分布”是最实用和简便的方法。频数分布是将全部数据根据大小分为若干组，其步骤如下：

1. 求两极差：例如100名初三男生仰卧起坐成绩中最少的做1次，最多的做19次，两极差为 $19 - 1 = 18$ 。

2. 分组和定组距：分组的多少是根据变量的个数。一般100人左右可分为7~9组，上例可以分6组左右，那么每组包括的范围应为 $18 \div 6 = 3$ 。

3. 排列组限：组限即一个组的上、下限，在下表中第一组为1——3次，也可写做1——。第二组为4——6次。习惯上数值由小到大从上到下排列。

4. 填数：和选举唱票一样，将每个变量按数值大小填在所属的组内，写成“正”字，便于计数。变量填完后将每组频数和累积频数以及总频数 $n$ 逐栏填写（见表2—1）。

表2—1 某校初三男生仰卧起坐成绩频数分布表

| 组距    | 记 数           | 组 频 数 | 累 积 频 数 |
|-------|---------------|-------|---------|
| 1—3   | —             | 1     | 1       |
| 4—6   | 正 正 丁         | 12    | 13      |
| 7—9   | 正 正 正 正 正 丁   | 27    | 40      |
| 10—12 | 正 正 正 正 正 正 一 | 31    | 71      |
| 13—15 | 正 正 正 正 下     | 23    | 94      |
| 16—18 | 正             | 4     | 98      |
| 19—21 | 丁             | 2     | 100     |

$n = 100$

从频数分布表中可以掌握样本的大致情况：最低成绩不低于1次；最高成绩不超过21次；从累积频数栏中可以知道达到某一水平的频数，例如9次和9次以下者有40人；10—12次者（即11次左右）的频数最多，我们叫11次为“众数”，是平均数的一种。

从频数分布表中我们发现频数的分布情况是中间多两头少，这是偶然事件分布的普遍规律，是一个很重要的性质，在下一讲中再详述。

## 二、平均数和标准差

平均数是分析数据的基本指标，它反映了各变量平均的水平，对各变量的总情况具有代表性。平均数有几种，最常用的是“算术平均数”。如将某班25名男生百米跑的成绩加起来被25除，得到的就是25人百米跑成绩的“算术平均数”，这是大家早已熟习的方法。在统计学中算术平均数又简称为“均数”，代表符号为 $\bar{X}$ （读作“杠x”），计算公式为：

$$\bar{X} = \frac{\sum X}{n} \quad (1)$$

(1)式中 $\Sigma$ 是总加符号， $\sum X$ 是将每个变量 $X$ 加起来。例如测得5个人50米跑的成绩为8.8"、8.6"、8.2"、9.4"、9.5"，则：

$$\begin{aligned} \bar{X} &= \frac{\sum X}{n} = \frac{8.8 + 8.6 + 8.2 + 9.4 + 9.5}{5} \\ &= 8.9 \text{ (秒)} \end{aligned}$$

但是均数还不能全面反映一组变量的情况,例如:8.8"、8.6"、8.2"、9.4"、9.5"和7.8"、7.6"、8.2"、10.4"、10.5"两组均数都是8.9"。显然两组的均数虽然相同,但情况却不同。后者各成绩好坏较前一组悬殊,即后一组各变量值差别大。统计学中对各变量值的变化叫“变异”。变异在一切体育活动中都存在。例如同年龄、同性别学生的身高、体重、跑跳能力等,都不尽相同。为了反映某一群变量之间参差不齐程度,有一个常用统计量叫“标准差”。计算标准差的公式:

$$S = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}} \quad (2)$$

(2)式中S是样本标准差的符号;  $n - 1$  叫自由度;  $X - \bar{X}$  称“离均差”,这是每个变量和它们的均数的差数;  $\sum(X - \bar{X})^2$  表示这些差数的平方总和。下面计算上例的标准差:

$$S = \sqrt{\frac{(8.8-8.9)^2 + (8.6-8.9)^2 + (8.2-8.9)^2 + (9.4-8.9)^2 + (9.5-8.9)^2}{5-1}} = 0.548 \text{ (秒)}$$

由于上述计算公式中

$$\sum(X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n} \quad \text{此公式也可写作:}$$

$$S = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}} \quad (3)$$

(3)式中  $\sum X^2$  表示各变量平方后相加,而  $(\sum X)^2$  表示各变

量相加后再平方,要注意区别。用公式(2)计算上例:

| X          | X <sup>2</sup> |
|------------|----------------|
| 8.8        | 77.44          |
| 8.6        | 73.96          |
| 8.2        | 67.24          |
| 9.4        | 88.36          |
| 9.5        | 90.25          |
| 44.5       | 397.25         |
| $\Sigma X$ | $\Sigma X^2$   |

$$\bar{X} = \frac{44.5}{5} = 8.9 \text{ (秒)}$$

$$S = \sqrt{\frac{397.25 - \frac{(44.5)^2}{5}}{5-1}} = 0.548 \text{ (秒)}$$

也可先根据变量值的大致情况,任意确定一个假设的平均数,然后再加校正值。如果假设上例均数为  $A = 8.5$ , 其计算公式和过程如下:

| X   | d<br>(X - A)       | d <sup>2</sup><br>(X - A) <sup>2</sup> |
|-----|--------------------|----------------------------------------|
| 8.8 | 0.3                | 0.09                                   |
| 8.6 | 0.1                | 0.01                                   |
| 8.2 | -0.3               | 0.09                                   |
| 9.4 | 0.9                | 0.81                                   |
| 9.5 | 1.0                | 1.00                                   |
|     | 2.00<br>$\Sigma d$ | 2.00<br>$\Sigma d^2$                   |

$$\bar{X} = A + \frac{\Sigma d}{n}$$

$$= 8.5 + \frac{2}{5} = 8.9 \text{ (秒)}$$

$$S = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{2 - \frac{2^2}{5}}{5-1}} = 0.548 \text{ (秒)}$$

在样本含量较大时,用频数分布表进行计算,现以表2-2所列仰卧起坐样本为列(见表2-2)。表2-2中第一

表2-2

初三男生仰卧起坐平均数和标准差计算表

| ① | 组 限                                    | 1—3                              | 4—6   | 7—9   | 10—12 | 13—15 | 16—18 | 19—21 | 总 计    |
|---|----------------------------------------|----------------------------------|-------|-------|-------|-------|-------|-------|--------|
| ② | ①行上、下组限平均值                             | 2                                | 5     | 8     | 11    | 14    | 17    | 20    |        |
| ③ | 组 频 数                                  | 1                                | 12    | 27    | 31    | 23    | 4     | 2     | 100    |
| ④ | ② × ③                                  | 2                                | 60    | 216   | 341   | 322   | 68    | 40    | 1049   |
| ⑤ | $\bar{X} = \text{①总计} \div \text{③总计}$ | $\bar{X} = 1049 \div 100 = 10.5$ |       |       |       |       |       |       |        |
| ⑥ | ② - $\bar{X}$                          | -8.5                             | -5.5  | -2.5  | 0.5   | 3.5   | 6.5   | 9.5   |        |
| ⑦ | ⑥ × ⑥                                  | 72.25                            | 30.25 | 6.25  | 0.25  | 12.25 | 42.25 | 90.25 |        |
| ⑧ | ③ × ⑦                                  | 72.3                             | 363.0 | 168.8 | 7.8   | 281.8 | 169.0 | 180.5 | 1243.2 |
| ⑨ | ⑧总计 ÷ ③总计                              | $1243.2 \div 100 = 12.43$        |       |       |       |       |       |       |        |
| ⑩ | $S = \sqrt{\text{⑨}}$                  | $S = \sqrt{12.43} = 3.53$        |       |       |       |       |       |       |        |

列为各行的编号，第二列为计算说明。如第④行为②×③，表示第②行的数乘第③行的数，然后填在第④行中，其它相同，当表按说明填完后，即能得到样本的均数和标准差。此法为计算器程序算法。

平均数和标准差的计算方法有多种，计算结果都相同，可参看徐英超教授和作者所著《体育统计方法》一书。（人民体育出版社，1981年出版）

### 三、变异系数

$$\text{计算变异系数的公式：} C = \frac{S}{\bar{X}} \times 100\% \quad (4)$$

变异系数也是反映变量变异大小的一个统计量，是标准差与均数的比值。变异系数没有单位，反映相对变异，便于不同事物或不同单位的统计量进行比较。例如某样本中10岁女孩身高均数为138厘米，标准差5.7厘米，变异系数  $C = 5.7/138 = 4.1\%$ ；14岁女孩均数157厘米，标准差5.6厘米。变异系数  $C = 5.6/157 = 3.6\%$ 。虽然从此例两个标准差看几乎是相等的但10岁女孩的变异系数大，因此相对讲10岁女孩身高发育参差不齐的程度就比14岁女孩要大。

各种事物都有各自特征的变异系数，因此可以将其估计在某一界值之内，如果实验数据超出这一界值，便可以怀疑实验或数据的可靠性，也可以从中发现新的问题。

## 练 习

1. 以假设平均数方法计算下列10人晨脉的平均数、标准差和变异系数。(单位: 次/分)

79, 72, 72, 73, 70, 69, 71, 68, 75, 73。

2. 在大样本中  $n-1$  的值与  $n$  值近似, 试以  $n$  替代  $n-1$ , 将计算标准差的

$$\text{公式 } S = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}} \quad \text{简化成}$$

$$S = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} \quad (\text{此式即较大样本计算}$$

标准差的公式)。

## 第三讲 正态分布

如果将上一讲表 2—1 中的“正”字组成的图形逆时针方向转  $90^\circ$  就成了一个钟形图, 靠近平均数的频数多, 远离平均数的频数逐渐减少, 两侧趋向于对称。如果在一个班级中做个实验, 请每个学生都将一枚硬币抛投 10 次, 然后记下每人出现正面的次数。结果出现 10 次的一定很少, 甚至没有, 而出现 2 次或 8 次者将会增加, 在 5 次左右的可能人数最多,