## §14.4　用样本估计总体

## 14.4.1　用样本估计总体的集中趋势参数

学习目标　1.了解平均数为什么是“最理想”的近似值.2.会计算一组数据的平均数.3.会根据频率分布表或频率直方图估计平均数．

知识点一　平均数

1．一般地，使(*x*－*a*1)2＋(*x*－*a*2)2＋…＋(*x*－*an*)2＝*nx*2－2(*a*1＋*a*2＋…＋*an*)*x*＋*a*＋*a*＋…＋*a*最小的*x*＝称为这*n*个数据*a*1，*a*2，…，*an*的平均数．

一般记为：＝.

2．一般地，若取值为*x*1，*x*2，…，*xn*的频率分别为*p*1，*p*2，…，*pn*，则其平均数为＝*x*1*p*1＋*x*2*p*2＋…＋*xnpn*.

知识点二　众数与中位数

1．众数：一般地，将一组数据中出现次数最多的那个数据叫作该组数据的众数．

2．中位数：一般地，将一组数据按照从小到大的顺序排成一列，如果数据的个数为奇数，那么排在正中间的数据就是这组数据的中位数；如果数据的个数为偶数，那么排在正中间的两个数据的平均数即为这组数据的中位数．

1．中位数是一组数据中间的数．(　×　)

2．众数是一组数据中出现次数最多的数．(　√　)

3．如果在*n*个数据中，*x*1，*x*2，…，*xn*出现的频率分别为*f*1，*f*2，…，*fn*，则＝.

(　×　)

4．平均数反映了总体的平均水平．(　√　)

一、平均数的计算

例1　(1)某班45名同学的年龄(单位：岁)如下：

14　15　14　16　15　17　16　15　16　16　15　15

17　13　14　15　16　16　15　14　15　15　14　15

16　17　16　15　15　15　16　15　13　16　15　15

17　14　15　16　16　15　14　15　15

求全班的平均年龄．

(2)从高三年级中抽出50名学生参加数学竞赛，由成绩得到如图所示的频率直方图．

试利用频率直方图估计高三年级学生的平均成绩．

解　(1)方法一　利用平均数的公式计算：

＝×(14＋15＋…＋15)＝×684＝15.2.

方法二　利用平均数的简化公式计算．

取*a*＝15，将已知各数减去15，得

－1　0　－1　1　0　2　1　0　1　1　0　0　2　－2　－1　0　1　1　0　－1　0　0　－1　0　1　2　1　0　0　0　1　0　－2　1　0　0　2　－1　0　1　1　0　－1　0　0

′＝×(－1＋0＋…＋0)＝×9＝0.2，

＝′＋*a*＝0.2＋15＝15.2.

方法三　利用加权平均数公式计算．

＝×(13×2＋14×7＋15×20＋16×12＋17×4)＝×684＝15.2.

即全班的平均年龄是15.2岁．

(2)样本平均数是频率直方图的“重心”，即所有数据的平均数，取每个小矩形底边的中点值乘以每个小矩形的面积再求和即可．

故平均成绩为45×(0.004×10)＋55×(0.006×10)＋65×(0.02×10)＋75×(0.03×10)＋85×(0.024×10)＋95×(0.016×10)＝76.2(分)．

反思感悟　(1)给定一组数据，要求其平均数可直接套用公式，若这组数据都在某一数据附近波动，可用平均数的简化公式计算，若这组数据某些数多次出现，可用加权平均数公式计算．

(2)在频率分布表中，平均数可用各组区间的组中值与对应频率之积进行估计．

(3)若一组数据的个数未知，但每一数据所占比例已知，可用频率平均数公式．

跟踪训练1　一批乒乓球，随机抽取100个进行检查，球的直径频率直方图如图．试估计这个样本的平均数．

解　平均数为39.96×0.1＋39.98×0.2＋40×0.5＋40.02×0.2＝39.996.

二、频率分布与数字特征的综合应用

例2　已知一组数据：

125　121　123　125　127　129　125　128　130　129

126　124　125　127　126　122　124　125　126　128

(1)填写下面的频率分布表：

|  |  |  |
| --- | --- | --- |
| 分组 | 频数 | 频率 |
| [121,123) |  |  |
| [123,125) |  |  |
| [125,127) |  |  |
| [127,129) |  |  |
| [129,131] |  |  |
| 合计 |  |  |

(2)作出频率直方图；

(3)根据频率直方图或频率分布表求这组数据的众数、中位数和平均数．

解　(1)频率分布表如下．

|  |  |  |
| --- | --- | --- |
| 分组 | 频数 | 频率 |
| [121,123) | 2 | 0.10 |
| [123,125) | 3 | 0.15 |
| [125,127) | 8 | 0.40 |
| [127,129) | 4 | 0.20 |
| [129,131] | 3 | 0.15 |
| 合计 | 20 | 1.00 |

(2)频率直方图如下．

(3)在[125,127)中的数据最多，取这个区间的中点值作为众数的近似值，得众数126，事实上，众数的精确值为125.图中虚线对应的数据是125＋2×＝126.25，事实上中位数为125.5.使用“组中值”求平均数：＝122×0.1＋124×0.15＋126×0.4＋128×0.2＋130×0.15＝126.3，

平均数的精确值为＝125.75.

反思感悟　众数、中位数、平均数与频率分布表、频率直方图的关系

(1)众数：众数一般用频率表中频率最高的一小组的组中值来显示，即在样本数据的频率直方图中，最高矩形的底边中点的横坐标．

(2)中位数：在频率直方图中，中位数左边和右边的直方图的面积应该相等．

(3)平均数：平均数在频率分布表中等于组中值与对应频率之积的和．

注意　因为频率直方图不能体现样本数据，因此由频率直方图得到的中位数不一定在样本数据中出现．

跟踪训练2　在一次中学生田径运动会上，参加男子跳高的17名运动员的成绩如表所示：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 成绩(单位：m) | 1.50 | 1.60 | 1.65 | 1.70 | 1.75 | 1.80 | 1.85 | 1.90 |
| 人数 | 2 | 3 | 2 | 3 | 4 | 1 | 1 | 1 |

分别求这些运动员成绩的众数、中位数与平均数．

解　在17个数据中，1.75出现了4次，出现的次数最多，即这组数据的众数是1.75.题中表里的17个数据可看成是按从小到大的顺序排列的，其中第9个数据1.70是最中间的一个数据，即这组数据的中位数是1.70；这组数据的平均数是＝(1.50×2＋1.60×3＋…＋1.90×1)＝≈1.69.

三、数字特征的应用

例3　某公司的33名职工的月工资(单位：元)如下表：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 职位 | 董事长 | 副董事长 | 董事 | 总经理 | 经理 | 管理员 | 职员 |
| 人数 | 1 | 1 | 2 | 1 | 5 | 3 | 20 |
| 工资 | 5 500 | 5 000 | 3 500 | 3 000 | 2 500 | 2 000 | 1 500 |

(1)求该公司职工月工资的平均数、中位数、众数；

(2)若董事长、副董事长的工资分别从5 500元、5 000元提升到30 000元、20 000元，那么公司职工的月工资的新的平均数、中位数和众数又是什么？

(3)你认为哪个统计量更能反映这个公司职工的工资水平？

解　(1)公司职工月工资的平均数为

＝＝≈2 091.

若把所有数据从小到大排序，则得到中位数是1 500，众数是1 500.

(2)当董事长、副董事长的工资提升后，职工月工资的平均数为

＝＝≈3 288.

中位数是1 500，众数是1 500.

(3)在这个问题中，中位数和众数都能反映出这个公司职工的工资水平，因为公司少数人的工资额与大多数人的工资额差别较大，这样导致平均数与中位数偏差较大，所以平均数不能反映这个公司职工的工资水平．

反思感悟　三种数字特征的比较

|  |  |  |
| --- | --- | --- |
| 名称 | 优点 | 缺点 |
| 众数 | ①体现了样本数据的最大集中点；②容易计算 | ①它只能表达样本数据中很少的一部分信息；②无法客观地反映总体的特征 |
| 中位数 | ①不受少数几个极端数据(即排序靠前或靠后的数据)的影响；②容易计算，便于利用中间数据的信息 | 对极端值不敏感 |
| 平均数 | 代表性较好，是反映数据集中趋势的量．一般情况下，可以反映出更多的关于样本数据全体的信息 | 任何一个数据的改变都会引起平均数的改变．数据越“离群”，对平均数的影响越大 |

跟踪训练3　某地区100位居民的人均月用水量(单位：t)的分组及各组的频数如下：

[0,0.5)，4；[0.5,1)，8；[1,1.5)，15；[1.5,2)，22；

[2,2.5)，25；[2.5,3)，14；[3,3.5)，6；[3.5,4)，4；

[4,4.5]，2.

(1)列出样本的频率分布表；

(2)画出频率直方图，并根据直方图估计这组数据的平均数、中位数、众数；

(3)当地政府制定了人均月用水量为3t的标准，若超出标准加倍收费，当地政府说，85%以上的居民不超过这个标准，这个解释对吗？为什么？

解　(1)频率分布表．

|  |  |  |
| --- | --- | --- |
| 分组 | 频数 | 频率 |
| [0,0.5) | 4 | 0.04 |
| [0.5,1) | 8 | 0.08 |
| [1,1.5) | 15 | 0.15 |
| [1.5,2) | 22 | 0.22 |
| [2,2.5) | 25 | 0.25 |
| [2.5,3) | 14 | 0.14 |
| [3,3.5) | 6 | 0.06 |
| [3.5,4) | 4 | 0.04 |
| [4,4.5] | 2 | 0.02 |
| 合计 | 100 | 1 |

(2)频率直方图如图．

众数：2.25；中位数：2.02；平均数：2.02.

(3)人均月用水量在3t以上的居民所占的比例为6%＋4%＋2%＝12%，即大约有12%的居民月用水量在3t以上，88%的居民月用水量在3t以下，因此政府的解释是正确的．

1．(多选)下列说法正确的是(　　)

A．在统计里，把所需考察对象的全体叫作总体

B．一组数据的平均数一定大于这组数据中的每个数据

C．平均数、众数与中位数从不同的角度描述了一组数据的集中趋势

D．众数是一组数据中出现次数最多的数

答案　ACD

解析　平均数不大于最大值，不小于最小值．故B错．

2．对于数据3,3,2,3,6,3,10,3,6,3,2，下列结论正确的是(　　)

A．这组数据的众数是3

B．这组数据的众数与中位数的数值不相等

C．这组数据的中位数与平均数的数值相等

D．这组数据的平均数与众数的数值相等

答案　A

解析　在这11个数中，数3出现了6次，出现的次数最多，故众数是3；将这11个数按从小到大的顺序排列得2，2,3,3,3,3,3,3,6,6,10，中间数据是3，故中位数是3；而平均数＝＝4.故只有A正确．

3．下列关于平均数、中位数、众数的说法中正确的是(　　)

A．中位数可以准确地反映出总体的情况

B．平均数可以准确地反映出总体的情况

C．众数可以准确地反映出总体的情况

D．平均数、中位数、众数都有局限性，都不能准确地反映出总体的情况

答案　D

4．某高校有甲、乙两个数学建模兴趣班，其中甲班40人，乙班50人．现分析两个班的一次考试成绩，算得甲班的平均成绩是90分，乙班的平均成绩是81分，则该校数学建模兴趣班的平均成绩是\_\_\_\_\_\_\_\_分．

答案　85

解析　平均成绩为＝85(分)．

5．样本容量为100的频率直方图如图所示，根据样本频率直方图，得平均数为\_\_\_\_\_\_\_\_．

答案　14.84

解析　平均数＝10×0.06＋12×0.1＋14×0.4＋16×0.24＋18×0.2＝14.84.

1．知识清单：

(1)平均数、平均数的估计．

(2)众数、中位数．

2．方法归纳：数学运算、数据分析．

3．常见误区：数据较多时，计算平均数易漏加或重复．

1．10名工人某天生产同一零件，生产的件数是15,17,14,10,15,17,17,16,14,12，设其平均数为*a*，中位数为*b*，众数为*c*，则有(　　)

A．*a*＞*b*＞*c* B．*b*＞*c*＞*a* C．*c*＞*a*＞*b* D．*c*＞*b*＞*a*

答案　D

解析　由已知得*a*＝×(15＋17＋14＋10＋15＋17＋17＋16＋14＋12)＝14.7，*b*＝×(15＋15)＝15，*c*＝17，

∴*c*＞*b*＞*a*.

2．某台机床加工的五批同数量的产品中次品数的频率分布如表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 次品数 | 0 | 1 | 2 | 3 | 4 |
| 频率 | 0.5 | 0.2 | 0.05 | 0.2 | 0.05 |

则次品数的平均数为(　　)

A．1.1 B．3

C．1.5 D．2

答案　A

解析　设数据*xi*出现的频率为*pi*(*i*＝1,2，…，*n*)，则*x*1，*x*2，…，*xn*的平均数为*x*1*p*1＋*x*2*p*2＋…＋*xnpn*＝0×0.5＋1×0.2＋2×0.05＋3×0.2＋4×0.05＝1.1.

3．某学习小组在一次数学测验中，得100分的有1人，95分的有1人，90分的有2人，85分的有4人，80分和75分的各有1人，则该小组成绩的平均数为(　　)

A．83 B．85

C．87 D．89

答案　C

解析　平均数是×(100＋95＋2×90＋4×85＋80＋75)＝87.

∴平均数是87.

4．某班全体学生参加物理测试成绩的频率直方图如图所示，则估计该班物理测试的平均成绩是(　　)

A．70分 B．75分

C．68分 D．66分

答案　C

解析　平均成绩就是频率直方图中每个小矩形的面积乘以小矩形底边中点的横坐标再求和，即0.005×20×30＋0.01×20×50＋0.02×20×70＋0.015×20×90＝68(分)．

5．某同学使用计算器求30个数据的平均数时，错将其中一个数据105输入为15，那么由此求出的平均数与实际平均数的差为(　　)

A．－1 B．1

C．－3 D．3

答案　C

解析　少输入90，＝3，平均数少3，求出的平均数减去实际的平均数等于－3.

6．已知一组数据4,6,5,8,7,6，那么这组数据的平均数为\_\_\_\_\_\_\_\_．

答案　6

解析　由已知得，所求平均数为＝6.

7．已知样本数据*x*1，*x*2，…，*x*10，其中*x*1，*x*2，*x*3的平均数为*a*，*x*4，*x*5，*x*6，…，*x*10的平均数为*b*，则样本数据的平均数为\_\_\_\_\_\_\_\_．

答案

解析　前3个数据的和为3*a*，后7个数据的和为7*b*，样本平均数为10个数据的和除以10，即.

8．已知一组数据按从小到大排列为－1,0,4，*x*,6,15，且这组数据的中位数是5，那么数据的众数是\_\_\_\_\_\_\_\_，平均数是\_\_\_\_\_\_\_\_．

答案　6　5

解析　∵中位数为5，∴＝5，即*x*＝6.

∴该组数据的众数为6，平均数为＝5.

9．某地区全体九年级的3 000名学生参加了一次科学测试，为了估计学生的成绩，从不同学校的不同程度的学生中抽取了100名学生的成绩如下：

100分12人，90分30人，80分18人，70分24人，60分12人，50分4人．

请根据以上数据估计该地区3 000名学生的平均分、合格率(60或60分以上均属合格)．

解　平均分为＝79.4(分)，

(12＋30＋18＋24＋12)÷100×100%＝96%，

所以样本的平均分是79.4分，合格率是96%，由此来估计总体3 000名学生的平均分是79.4分，合格率是96%.

10．某小区广场上有甲、乙两群市民正在进行晨练，两群市民的年龄如下(单位：岁):

甲群：13,13,14,15,15,15,15,16,17,17；

乙群：54,3,4,4,6,6,6,6,6,55.

(1)甲群市民年龄的平均数、中位数和众数各是多少？其中哪个统计量能较好地反映甲群市民的年龄特征？

(2)乙群市民年龄的平均数、中位数和众数各是多少？其中哪个统计量能较好地反映乙群市民的年龄特征？

解　(1)甲群市民年龄的平均数为＝15，中位数为15，众数为15.

平均数、中位数和众数相等，因此它们都能较好地反映甲群市民的年龄特征．

(2)乙群市民年龄的平均数为＝15，中位数为6，众数为6.

由于乙群市民大多数是儿童，所以中位数和众数能较好地反映乙群市民的年龄特征，而平均数的可靠性较差．

11．有容量为100的样本，数据分组及各组的频数、频率如下：

[12.5,14.5)，6,0.06；[14.5,16.5)，16,0.16；[16.5，18.5)，18,0.18；[18.5,20.5)，22,0.22；[20.5,22.5)，20,0.20；[22.5,24.5)，10,0.10；[24.5,26.5]，8,0.08.则估计总体的平均数为(　　)

A．19.40 B．19.42 C．19.44 D．19.46

答案　B

解析　由于每组数据是一个范围，所以可以用组中值近似地表示平均数．

方法一　总体的平均数约为(13.5×6＋15.5×16＋17.5×18＋19.5×22＋21.5×20＋23.5×10＋25.5×8)＝19.42.故总体的平均数约为19.42.

方法二　组中值与对应频率积的和为13.5×0.06＋15.5×0.16＋17.5×0.18＋19.5×0.22＋21.5×0.20＋23.5×0.10＋25.5×0.08＝19.42.故总体的平均数约为19.42.

12．某中学举行电脑知识竞赛，现将高一参赛学生的成绩进行整理后分成五组，绘制成如图所示的频率直方图，已知图中从左到右的第一、二、三、四、五小组的频率分别是0.30，0.40,0.15,0.10,0.05.

则估计高一参赛学生的成绩的众数、中位数分别为\_\_\_\_\_\_，\_\_\_\_\_\_.

答案　65　65

解析　由题图可知众数为65，

又∵第一个小矩形的面积为0.3，

∴设中位数为60＋*x*，则0.3＋*x*×0.04＝0.5，得*x*＝5，

∴中位数为60＋5＝65.

13．某企业有3个分厂生产同一种电子产品，第一、二、三分厂的产量之比为1∶2∶1，用分层抽样方法(每个分厂的产品为一层)从3个分厂生产的电子产品中共抽取100件做使用寿命的测试，由所得的测试结果算得从第一、二、三分厂取出的产品的使用寿命的平均数分别为980 h，1 020 h，1 032 h，则抽取的100件产品的使用寿命的平均数为\_\_\_\_\_\_\_\_ h.

答案　1 013

解析　依题意可知平均数＝＝1 013.

14．如图是某路段的一个检测点对200辆汽车的车速进行检测所得结果的频率直方图，利用组中值估计，则下列说法正确的是\_\_\_\_\_\_\_\_(填序号)．

①平均数为62.5；②中位数为62.5；③众数为65.

答案　②③

解析　由频率直方图可知，平均数为0.01×10×45＋0.03×10×55＋0.04×10×65＋0.02×10×75＝62.

设中位数为*a*，由题意得0.01×10＋0.03×10＋0.04×(*a*－60)＝0.5，解得*a*＝62.5.

众数为频率直方图中最高矩形底边中点值，故其众数为65.

15．为了普及环保知识，增强环保意识，某学校随机抽取30名学生参加环保知识测试，得分(十分制)如图所示，假设得分的中位数为*m*，众数为*n*，平均数为，则*m*，*n*，的大小关系为\_\_\_\_\_．

答案　*n*＜*m*＜

解析　由题目所给的统计图可知，30个数据按从小到大的顺序排列好后，中间两个数为5,6，故中位数为*m*＝＝5.5.

又众数为*n*＝5，平均数＝＝.所以*n*＜*m*＜.

16．我国是世界上严重缺水的国家，某市为了制定合理的节水方案，对居民用水情况进行了调查，通过抽样，获得了某年100位居民每人的月均用水量(单位：吨)，将数据按照[0,0.5)，[0.5,1)，…，[4,4.5]分成9组，制成了如图所示的频率直方图．

(1)求直方图中*a*的值；

(2)设该市有30万居民，估计全市居民中月均用水量不低于3吨的人数；

(3)估计居民月均用水量的中位数．

解　(1)由频率直方图可知，月均用水量在[0,0.5)的频率为0.08×0.5＝0.04.

同理，在[0.5,1)，[1.5,2)，[2,2.5)，[3,3.5)，[3.5,4)，[4,4.5]的频率分别为0.08,0.21,0.25,0.06,0.04,0.02.

由1－(0.04＋0.08＋0.21＋0.25＋0.06＋0.04＋0.02)＝0.5×*a*＋0.5×*a*，

解得*a*＝0.30.

(2)由(1)知，100位居民月均用水量不低于3吨的频率为0.06＋0.04＋0.02＝0.12.由以上样本的频率分布，可以估计30万居民中月均用水量不低于3吨的人数为300 000×0.12＝36 000.

(3)设中位数为*x*.

因为前5组的频率之和为0.04＋0.08＋0.15＋0.21＋0.25＝0.73>0.5.

而前4组的频率之和为0.04＋0.08＋0.15＋0.21＝0.48<0.5.

所以2≤*x*<2.5.

由0.50×(*x*－2)＝0.5－0.48，解得*x*＝2.04.

故可估计居民月均用水量的中位数为2.04.