总体集中趋势估计---学习指南

一、学习目标：

1．结合实例，能用样本估计总体的集中趋势参数（平均数、中位数、众数）.

2．会从频率分布直方图中提取基本的数字特征（众数、中位数、平均数），并做出统计学意义的解释,

形成对数据处理过程进行初步评价的意识.

3．理解集中趋势参数的统计含义.

二、学法指导：

数据集中趋势的刻画包括平均数、中位数、众数等数字特征，这些概念学生初中已经有了解，且在抽样调查中已经学习了总体平均数的估计，因此对于总体集中趋势的估计，教科书主要是结合案例，通过平均数、中位数、众数关系的讨论，以及如何从图表中估计它们，让学生进一步理解它们的统计含义．频率分布直方图是数据描述部分的重点，各种数字特征的统计含义是数据分析部分的重点；而能根据实际问题的特点，灵活应用所学统计知识是难点.对于本节课利用频率分布直方图求样本的众数、中位数和平均数是本节课的重点，而其中利用频率分布直方图求样本的中位数和平均数是本节课的难点问题.

三、学习过程：

为了了解总体的情况，我们前两节课学习了总体取值规律的估计，总体百分位数的估计，用样本的频率分布估计总体分布，体会了用样本估计的总体的这种方法。但有时候，我们可能不太关心总体的分布规律，而更关心总体取值在某一方面的特征。例如，对于某县今年小麦的收成情况，我们可能会更关注该县今年小麦的总产量或平均每公顷的产量，而不是产量的分布；对于一个国家国民的身高情况，我们可能更关注身高的平均数或中位数，而不是身高的分布；等等.

初中的学习中我们了解到了平均数、中位数和众数等都是刻画“中心位置”的量，它们从不同角度刻画了一组数据的集中趋势.下面我们就通过具体实例进一步了解这些量的意义，探究它们之间的联系与区别，并根据样本的集中趋势估计总体的集中趋势.

探究一：众数、中位数和平均数在具体数据中的应用

问题 1：利用 100 个居民用户月均用水量的样本数据，求样本平均数和中位数，从而估计总体平均数和中位数。

9.0	13.6	14.9	5.9	4.0	7.1	6.4	5.4	19.4	2.0
2.2	8.6	13.8	5.4	10.2	4.9	6.8	14.0	2.0	10.5
2.1	5.7	5.1	16.8	6.0	11.1	1.3	11.2	7.7	4.9
2.3	10.0	16.7	12.0	12.4	7.8	5.2	13.6	2.6	22.4
3.6	7.1	8.8	25.6	3.2	18.3	5.1	2.0	3.0	12.0
22.2	10.8	5.5	2.0	24.3	9.9	3.6	5.6	4.4	7.9
5.1	24.5	6.4	7.5	4.7	20.5	5.5	15.7	2.6	5.7
5.5	6.0	16.0	2.4	9.5	3.7	17.0	3.8	4.1	2.3
5.3	7.8	8.1	4.3	13.3	6.8	1.3	7.0	4.9	1.8
7.1	28.0	10.2	13.8	17.9	10.1	5.5	4.6	3.2	21.6

利用 Excle 计算这 100 个数据的平均数，中位数和众数

样本平均数：$\bar{y}=\frac{y_1+y_2+\cdots+y_{100}}{100}=8.79$　　样本中位数：$\frac{6.8+6.8}{2}=6.8$

样本的众数：2.0 与 5.5

估计：因为数据是抽自全市居民的简单随机样本，所以我们可以据此估计全市居民用户的月均用水量约为 8.79t，其中位数约为 6.8t.

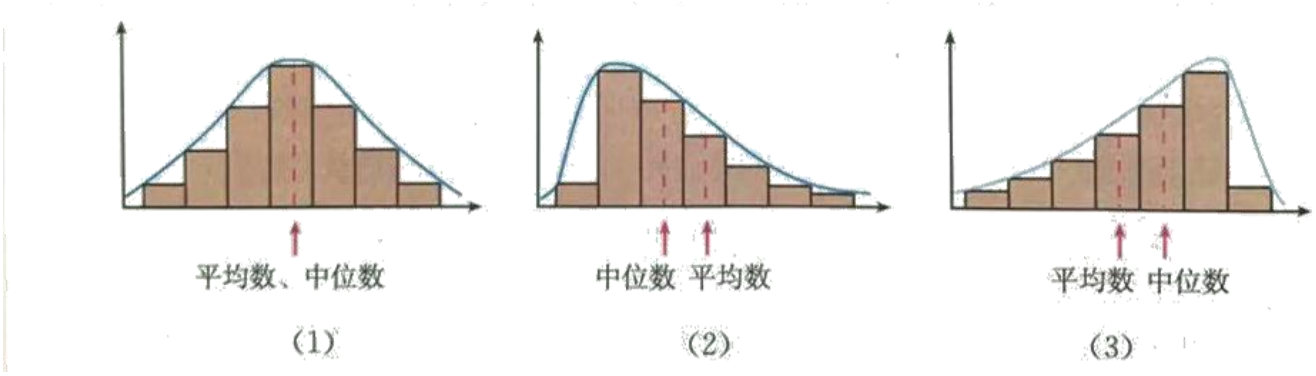
问题 2：如果录入数据时，将 7.7 输入为 77，则平均数和中位数有什么变化？说明什么问题？

通过简单计算可以发现，平均数由原来的 8.79 t 变为 9.483 t，中位数没有变化，还是 6.8 t. 这是因为样本平均数与每一个样本数据有关，样本中的任何一个数据的改变都会引起平均数的改变；但中位数只利用了样本数据中间位置的一个或两个值，并未利用其他数据，所以不是任何一个样本数据的改变都会引起中位数的改变. 因此，与中位数比较，平均数反映出样本数据中的更多信息，对样本中的极端值更加敏感.

追问：若一组数据确定了，中位数是唯一的吗？任何一个样本数据的改变都会或者都不会影响中位数吗？

样本数据确定了，中位数是唯一确定的，但个别样本数据的变化不一定影响中位数.

问题 3：平均数和中位数都描述了数据的集中趋势，它们的大小关系和数据分布的形态有关在下图的三种分布形态中，平均数和中位数的大小存在什么关系？



图（1）

一般来说，

对于一个单峰的频率分布直方图来说，如果直方图的形状是对称的，则平均数和中位数大体上差不多；

如果直方图在右边“拖尾”，则平均数大于中位数；

如果直方图在左边“拖尾”，则平均数小于中位数；

即平均数总在“长尾巴”那边。

问题 4：某学校要定制高一年级的校服，学生根据厂家提供的参考身高选择校服规格，据统计，高一年级女生需要不同规格的校服频数如下表：

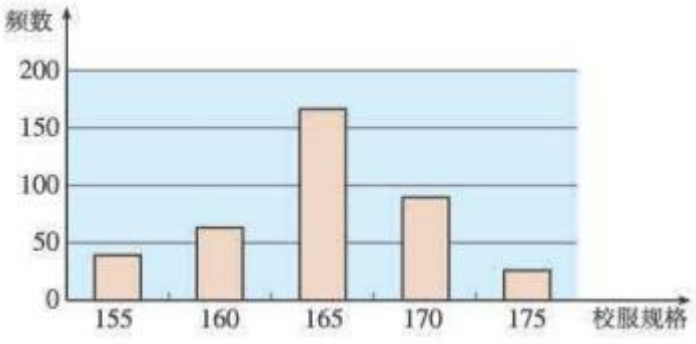
校服规格	155	160	165	170	175	合计
频数	39	64	167	90	26	386

如果要用一个量来代表该校高一女生所需校服的规格，那么在中位数、平均数和众数中，哪个量比较合适？用该表中的数据估计全国高一年级女生校服规格是否合理？

分析：虽然校服规格是用数字表示的，但它们事实上是几种不同的类别. 对于这样的分类数据，用众数作为这组数据的代表比较合适.

解：为了更直观地观察数据的特征，我们用条形图(图 2）来表示表中的数据可以发现，选择校服规格“165”的女生的频数最高，所以用众数 165 作为该校高一年级女生校服的规格比较合适.

由于全国各地的高一年级女生的身高存在一定的差异，所以用一个学校的数据估计全国高一年级女生校服规格不合理.



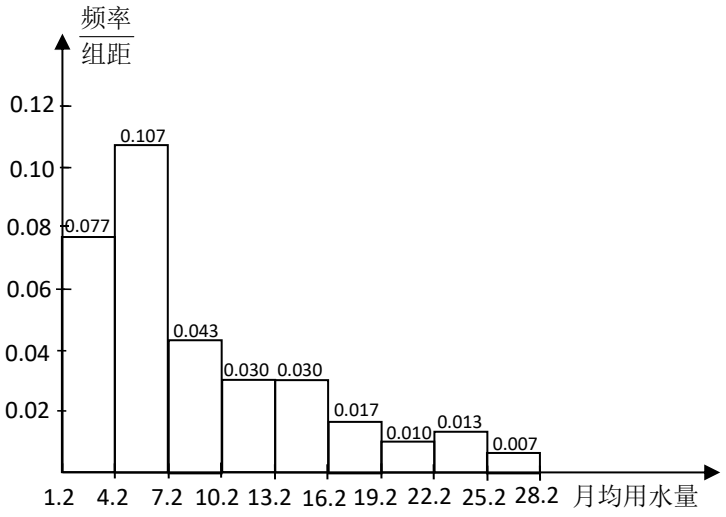
图二

众数只利用了出现次数最多的那个值的信息. 众数只能告诉我们它比其他值出现的次数多，但并未告诉我们它比别的值多的程度. 因此，众数只能传递数据中的信息的很少一部分，对极端值也不敏感.

探究二：在频率分布直方图中求众数、中位数和平均数

问题 5：假如我们没有数据，只有频率分布直方图。那我们该如何利用样本的频率分布直方图来估计样本的的众数、中位数、平均数等，从而估计总体的众数、中位数、平均数。

前面一节在调查 100 位居民的月均用水量的问题中，给出了这些样本数据的频率分布直方图如图:



图三

（1） 从频率分布直方图估计众数

如何从频率分布直方图中估计众数？众数能够说明什么问题？

在频率分布直方图中，月均用水量在区间[4.2，7.2)内的居民最多，可以讲这个区间的中点 5.7 作为众数的估计值，众数常常用在描述分类型数据中，在这个实际问题中，众数“5.7”让我们知道月均用水量在区间[4.2，7.2)内的居民用户最多.这个信息具有实际意义.

从抽样的 100 个数据看众数是 5. 5, 和估计出这个数值不同（相同或不同），原因是频率分布直方图隐藏了样本的一部分数据.

（2） 从频率分布直方图估计中位数

如何从频率分布直方图估计中位数？

根据中位数的意义，在样本中，有 50%的个体小于或等于中位数，也有 50%的个体大于或等于中位数.因此，在频率分布直方图中，中位数左边和右边的直方图的面积应该相等.由于

$$0.077\times3=0.231，（0.077+0.107）\times3=0.551$$

因此中位数落在区间[4.2，7.2)内.设中位数为 x，由

$$0.077\times3+0.107\times（x-4.2）=0.5 \text{ 得到 } x\approx6.71.$$

因此，中位数约为 6.71，

追问：根据问题 3 的计算，中位数值是 6.8，和估计出这个数值不同（相同或不同）原因是频率分布直方图隐藏了样本的一部分数据.

追问：中位数不受少数几个极端值的影响，这在某些情况下是一个优点，但是它对极端值的不敏感有时也会成为缺点，你能举例说明吗？

比如书中的案例居民用水问题，居民用水中位数 6.8，而居民月均用水最大量为 28.0，相差的比较多，中位数对极端值的反应不敏感.

（3） 从频率分布直方图估计平均数

如何从频率分布直方图中估计平均数？

追问:初中我们是如何求平均数的？

例：数据 7，8，6，8，6，5，8，10，7，4 中的众数，中位数，平均数，分别是多少?

4，5，6，6，7，7，8，8，8，10

平均数：$\dfrac{4+5+6\times2+7\times2+8\times3+10}{10}=6.9$

$$\frac{4+5+6\times2+7\times2+8\times3+10}{10}$$

$$=4\times\frac{1}{10}+5\times\frac{1}{10}+6\times\frac{2}{10}+7\times\frac{2}{10}+8\times\frac{3}{10}+10\times\frac{1}{10}=6.9$$

因为样本平均数可以表示为数据与它的频率的乘积之和，所以在频率分布直方图中，样本平均数可以用每个小矩形底边中点的横坐标与小矩形的面积之和近似代替.如图三所示，可以知道每个小矩形的高度，于是平均数的近似值为

$$0.077\times3\times\left(\frac{1.2+4.2}{2}\right)+0.107\times3\times\left(\frac{4.2+7.2}{2}\right)+\cdots+0.007\times3\times\left(\frac{25.2+28.2}{2}\right)=8.96,$$

根据问题 3 的计算，平均数是 8.79,和估计出这个数值不同（相同或不同），原因是频率分布直方图隐藏了样本的一部分数据.

问题 6：你能总结分析一下平均数，中位数和众数与频率分布表，频率分布直方图的关系吗？

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

(2)中位数：在频率分布表中，中位数是累计频率(样本数据小于某一数值的频率叫做该数值点的累计频率)为 0.5 时所对应的样本数据的值，而在样本中有 50%的个体小于或等于中位数，也有 50%的个体大于或等于中位数．因此，在频率分布直方图中，中位数左边和右边的直方图的面积应该相等.

(3)平均数：平均数在频率分布直方图中等于每个小矩形底边中点的横坐标与小矩形的面积的乘积之和.

(4)利用直方图求众数、中位数、平均数均为近似值，往往与实际数据得出的不一致，但它们能粗略估计其众数、中位数和平均数.

问题 7：学习了总体集中趋势的估计，认识了平均数，中位数和众数，假设你现在去人力市场去找工作，有一个企业老板告诉你，“我们企业员工的年平均收入是 20 万元”，你如何理解这句话？

这句话是真实的，但它可能描述的是差异巨大的实际情况.例如，可能这个企业的共资水平普遍偏高，也就是员工年收入的中位数、众数与平均数差不多；也可能是绝大多数员工的年收入较低（如大多数是 5 万元左右），而少数员工的年收入很高，甚至达到 100 万元，在这种情况下年的后入的平均数就比中位数大得多.尽管在后一种情况下，用中位数或众数比用平均数更合理些，但这个老板为了招揽员工，却用了平均数.

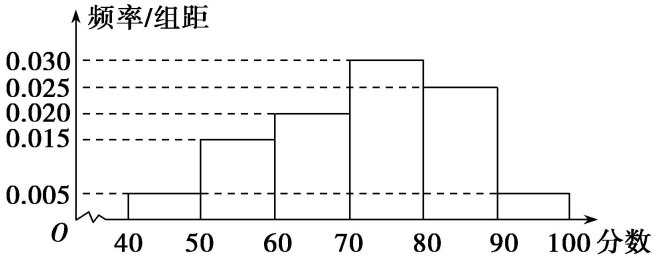
所以，我们强调“用数据说话”，但同时又要防止被数据误导.

问题 8：你能总结分析一下平均数，中位数和众数的优缺点吗？

名称	优点	缺点
众数	①体现了样本数据的最大集中点;②容易得到	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几个数据的影响;②容易得到,便于利用中间数据的信息	对极端值不敏感
平均数	能反映出更多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

例题分析：

某校从参加高二年级学业水平测试的学生中抽出 80 名学生，其数学成绩(均为整数)的频率分布直方图如图所示.



(1)求这次测试数学成绩的众数；

(2)求这次测试数学成绩的中位数；

(3)求这次测试数学成绩的平均分.

[解析] (1)由图知众数为$\frac{70+80}{2}=75$.

(2)由图知，设中位数为x，由于前三个矩形面积之和为0.4，第四个矩形面积为0.3,0.3+0.4>0.5，因此中位数位于第四个矩形内，得0.1=0.03(x−70)，所以$x\approx73.3$.

(3)由图知这次数学成绩的平均分为：

$\frac{40+50}{2}\times0.005\times10+\frac{50+60}{2}\times0.015\times10+\frac{60+70}{2}\times0.02\times10+\frac{70+80}{2}\times0.03\times10+\frac{80+90}{2}\times0.025\times10+\frac{90+100}{2}\times0.005\times10=72$.

变式练习：

本例条件不变，试估计80分以上的学生人数.

解析：[80,90)分的频率为：0.025×10=0.25，

频数为：0.25×80=20.

[90,100)分的频率为：0.005×10=0.05，

频数为：0.05×80=4.

所以估计80分以上的学生人数为20+4=24.

问题9：小结本节课的主要内容：

知识点一　平均数

(1)定义：一组数据的和与这组数据的个数的商．数据x_1，x_2，…，x_n的平均数为$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$．在频率分布直方图中，平均数$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}f_ix_i$，其中$f_i$为第$i$个小矩形对应的频率，$x_i$为第$i$个小矩形底边中点的横坐标.

(2)特征：样本平均数与<u>每一个</u>样本数据有关，样本中的任何一个数据的改变都会引起<u>平均数</u>的改变，这是中位数不具有的性质．所以与中位数比较，平均数反映出样本数据中的更多<u>信息</u>，但平均数受样本中的<u>极端值</u>的影响较大，使平均数在估计总体时可靠性降低.

知识点二　中位数

(1)定义：一组数据按从小到大的顺序排成一列，处于<u>中间</u>位置的数称为这组数据的中位数．在频率分布直方图中，中位数左边和右边的直方图的面积<u>相等</u>.

(2)特征：一组数据中的中位数是<u>唯一</u>的，中位数只利用了样本数据中间位置的<u>一个</u>或<u>两个</u>值，并未利用其他数据，所以不是任何一个样本数据的改变都会引起<u>中位数</u>的改变.

知识点三　众数

(1)定义：一组数据中出现次数<u>最多</u>的数称为这组数据的众数．在频率分布直方图中，众数是最高矩形的底边的<u>中点</u>.

(2)特征：一组数据中的众数可能<u>不止</u>一个．对分类型数据(如校服规格、性别、产品质量等级等)集中趋势的描述可以用众数，但众数只能告诉我们它比其他值出现的次数多，但并未告诉我们它比别的数值多的程度．因此，众数只能传递数据中的信息的<u>很少一部分</u>，对极端值也不敏