

总体百分位数的估计学习任务单

一、学习目标

1. 理解百分位数的概念，结合实例能用样本百分位数，并能描述百分位数的统计含义；
2. 由一组数据的频数表、频率表、频数分布直方图或频率分布直方图可以近似计算这组数据的百分位数，知道计算百分位数的步骤；
3. 了解百分位数的特点，知道几个重要的百分位数.

二、学法指导

学生在初中已初步了解用样本估计总体的思想，知道可用样本的平均数和方差估计总体的平均数和方差. 上一节课我们已经学会了处理数据的一些方法：频率分布表、频率分布直方图，并且明白不同数据图在表示统计数据上有不同特点. 这一节课是学习从另外一种角度看问题，一种新的处理数据的方式.

首先通过引例让学生感受百分位数的实际意义然后再借助在上一节课问题 2 “市政府希望使 80%的居民用户生活用水费支出不受影响”中 80%这个要求出发，分析出问题的关键就是找一个数据 a ，大于 a 的占 20%，这样很自然就可以启发学生对数据进行一个从小到大的重排，引出第 80 百分位数，在引出第 80 百分位数的过程中要突出数据处理的基本过程，引出第 80 百分位数后侧重对问题 2 的解决，让学生感受到百分位数的统计含义.

在问题 2 解决后，回过头重点分析“第 80 百分位数”，给出百分位数的概念和计算步骤，然后通过两个组具体的数据示范按照步骤计算百分位数的一般过程.

引例：老师在对某个学生数学成绩分析时，比如该生考了 72 分，教师一般不单独依据该生考了多少分论其成绩好与坏，通常状况下，会对比其他考生的成绩进行分析，如果该生的成绩位于所有考生成绩从小到大排列的中间，则认为该生成绩中等，位于靠后的位置，则认为该生成绩比较优异，所以在实际生活中我们经常会通过要研究对象所处的“位置”来做决策，本节课我们就来研究一下，数学中的用什么概念来刻画这个内容。

任务一：学习百分数的概念，了解百分位数的统计含义

问题 2 如果该市政府希望使 80%的居民用户生活用水费支出不受影响，根据 9.2.1 节中 100 户居民用户的月均用水量数据，你能给市政府提出确定居民用户月均

用水量标准的建议吗？

假设通过简单随机抽样，获得了 100 户居民用户的月均用水量数据(单位: t):

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

首先要明确一下问题: 根据市政府的要求确定居民用户月均用水量标准, 就是要寻找一个数 a , 使全市居民用户月均用水量中不超过 a 的占 80%, 大于 a 的占 20%. 下面我们通过样本数据对 a 的值进行估计.

把 100 个样本数据按从小到大排序, 得到第 80 个和第 81 个数据分别为 13.6 和 13.8, 可以发现, 区间(13.6, 13.8)内的任意一个数, 都能把样本数据分成符合要求的两部分. 一般地, 我们取这两个数的平均数 $\frac{13.6+13.8}{2} = 13.7$, 并称此数为这组数据的第 80 百分位数或 80%分位数.

根据样本数据的第 80 百分位数, 我们可以估计总体数据的第 80 百分位数为 13.7 左右, 由于样本的取值规律与总体的取值规律之间会存在偏差, 而在决策问题中, 只要临界值近似为第 80 百分位数即可, 因此为了实际中操作的方便, 可以建议市政府把月均用水量标准定为 14 t, 或者把年用水量标准定为 168 t

思考题: 你认为 14t 这个标准一定能够保证 80%的居民用水不超标吗? 如果不一定, 那么哪些环节可能会导致结论的差别?

解答: 不一定, 尽管这组数据中 80%的居民户用水不超过 14 t, 因为它来自样本观测数据, 只是总体 80%分位数的一个估计值, 抽样方法和样本的随机性都可能导致样本百分位数估计总体百分位数的误差.

在问题 2 中, 13.7 是上面那 100 个数据的第 80 百分位数或 80%分位数.

一般地，一组数据的第 p 百分位数是这样—个值，它使得这组数据中至少有 $p\%$ 的数据小于或等于这个值，且至少有 $(100-p)\%$ 的数据大于或等于这值.

可以通过下面的步骤计算—组 n 个数据的第 p 百分位数：

第 1 步，按从小到大排列原始数据.

第 2 步，计算 $i=n \times p \%$.

第 3 步，若 i 不是整数，而大于 i 的比邻整数为 j ，则第 p 百分位数为第 j 项数据；若 i 是整数，则第 p 百分位数为第 i 项与第 $(i+1)$ 项数据的平均数.

百分位数的特点

(1) 第 0 百分位数为数据组中的最小数，第 100 百分位数为数据组中的最大数.

(2) —组数据的百分位数既可能是这组数据中的数，也可能不是这组数据中的数.

(3) —组数据的某些百分位数可能是同一个数.

我们在初中学过的中位数，相当于是第 50 百分位数. 在实际应用中，除了中位数外，常用的分位数还有第 25 百分位数，第 75 百分位数，这三个分位数把—组由小到大排列后的数据分成四等份，因此称为四分位数，其中第 25 百分位数也称为第一四分位数或下四分位数等，第 75 百分位数也称为第三四分位数或上四分位数等. 另外，像第 1 百分位数，第 5 百分位数，第 95 百分位数和第 99 百分位数在统计中也经常被使用.

任务二：计算—组数据的百分位数

例 2 根据 9.1.2 节问题 3 中女生的样本数据，估计树人中学高—年级女生的第 25，50，75 百分位数.

解：把 27 名女生的样本数据按从小到大排序，可得

148.0 149.0 154.0 154.0 155.0 155.0 155.5 157.0 157.0

158.0 158.0 159.0 161.0 161.0 162.0 162.5 162.5 163.0

163.0 164.0 164.0 164.0 165.0 170.0 171.0 172.0 172.0

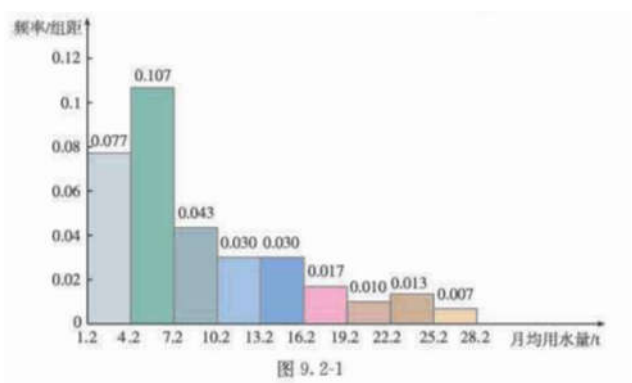
由 $25\% \times 27 = 6.75$ ， $50\% \times 27 = 13.5$ ， $75\% \times 27 = 20.25$ ，可知样本数据的第 25，50，75 百分位数为第 7，14，21 项数据，分别为 155.5，161，164，据此可以估计树人中学高—年级女生的第 25，50，75 百分位数分别约为 155.5，

161 和 164.

例 3 根据表 9. 2-1 或图 9. 2-1，估计月均用水量的样本数据的 80%和 95%分位数.

表 9. 2-1

分组	频数累计	频数	频率
[1. 2, 4. 2)	正正正正正	23	0. 23
[4. 2, 7. 2)	正正正正正正正	32	0. 32
[7. 2, 10. 2)	正正正	13	0. 13
[10. 2, 13. 2)	正正	9	0. 09
[13. 2, 16. 2)	正正	9	0. 09
[16. 2, 19. 2)	正	5	0. 05
[19. 2, 22. 2)	正	3	0. 03
[22. 2, 25. 2)	正	4	0. 04
[25. 2, 28. 2]	正	2	0. 02
合计		100	1. 00



分析: 在某些情况下，我们只能获得整理好的统计表或统计图，与原始数据相比，它们损失了一些信息. 例如由表 9. 2-1，我们知道在[16. 2， 19. 2)内有 5 个数据，但不知道这 5 个数据具体是多少，此时，我们通常把它们看成均匀地分布在此区间上.

解:由表 9. 2-1 可知，月均用水量在 13. 2t 以下的居民用户所占比例为

$$23\%+32\%+13\%+9\%=77\%.$$

在 16. 2t 以下的居民用户所占比例为

$$77\%+9\%=86\%.$$

因此， 80%分位数一定位于[13. 2， 16. 2)内. 由

$$13. 2 + 3 \times \frac{0.80-0.77}{0.86-0.77}=14. 2,$$

可以估计月均用水量的样本数据的 80%分位数约为 14. 2.

类似地，由

$$22. 2 + 3 \times \frac{0.95-0.94}{0.98-0.94}=22. 95,$$

可以估计月均用水量的样本数据的 95%分位数约为 22.95.

思考：在问题 2 中由“100 户居民用户的月均用水量数据”计算出的第 80 百分位数是 13.7，在例 3 中利用这“100 户居民用户的月均用水量数据”整理好的统计表或统计图计算出的第 80 百分位数是 14.2，同一个问题最后计算结果不一样，你知道为什么吗？

解答：问题 2 中计算结果 13.7 是 100 个原始数据的第 80 百分位数，而例 3 中经过整理后原始数据已经损失，14.2 是一个近似数，因此他们的值会有差别.