

频率分布表和频率分布直方图

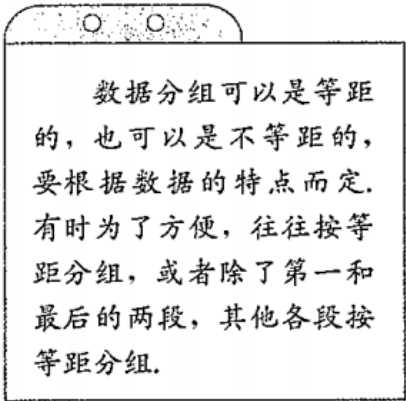
1. 在初中我们学习过频数分布表和频数分布直方图，那么我们为什么还要学习频率分布表和频率分布直方图？

用样本在某区间内的频率可以估计总体在相应区间内的比例，但用样本在某区间内的频数不能直接估计总体在相应区间内的频数，需要将样本频数除以样本容量后再乘总体中的个体数。也就是说，相对而言，频率分布直方图使用更方便。

2. 为什么频率分布直方图的纵坐标用频率/组距，而不是直接用频率？

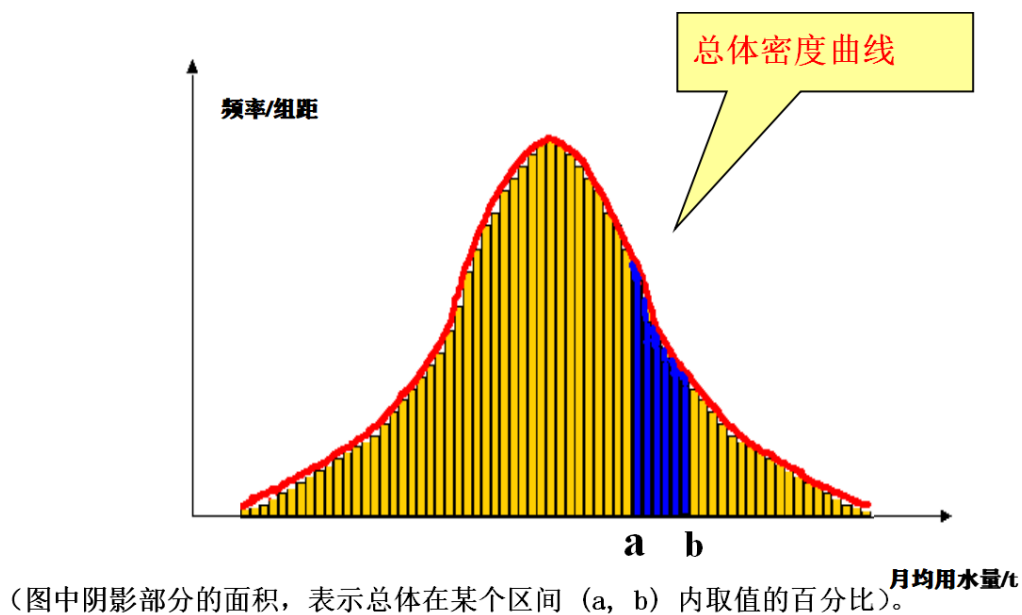
在课本中有这样一个边注：

所以，对于不等距分组的频率分布直方图，如果小长方形的高为频率，那么一般来说，区间越长，小长方形就越高，而两个长度不同的区间的频率比较就没有意义；



数据分组可以是等距的，也可以是不等距的，要根据数据的特点而定。有时为了方便，往往按等距分组，或者除了第一和最后的两段，其他各段按等距分组。

另外，全部小长方形的面积之和为 1，这个与总体中落在全部取值范围内的频率为 100%是一致的，可以作为总体分布的估计。一般来说，总体的极差是有限的，随着样本量增大，我们的组数也随之增加，而组距=极差/组数，所以，组距会越来越小，于是我们的频率分布直方图就近似变成了一条曲线，统计中称这条光滑曲线为**总体密度曲线**。



正确解读统计数据

统计的最终目的还是做出推断，教科书 200 页边空中的问题提到，“你能得出“2016 年的空气质量比 2015 年明显改善了”的结论吗？为什么？”

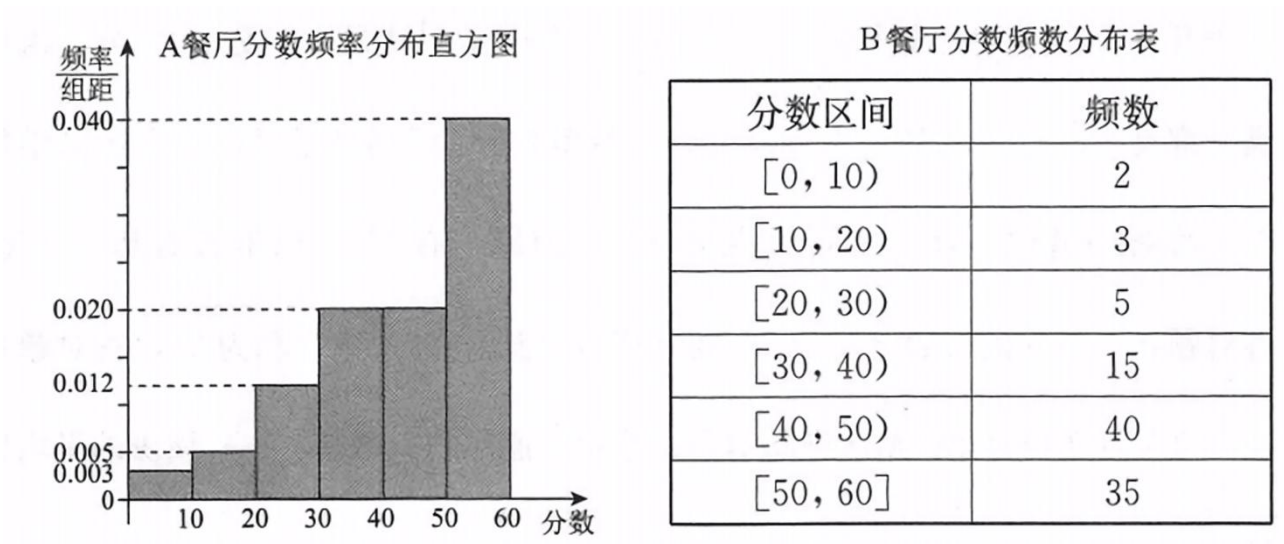
2016 年 6 月的空气质量数据与 2015 年全年的空气质量数据相比，2016 年 6 月中度以上的污染天气频率明显小于 2015 年，只能说从整体上看，2016 年 6 月的空气质量要好于 2015 年全年的空气质量，但不能说明 2016 年的空气质量比 2015 年明显改善。原因是空气质量与季节有关，比如冬天因为北方地区烧煤等因素的影响，往往空气质量会比较差，用 6 月的空气质量数据代表全年的空气质量的情况，代表性差（也就是说，这是一个极端样本）。若要比 2016 年与 2015 年的空气质量，可以把 2016 年 365 天随机的 30 天空气质量数据与 2015 年全年数据进行比较，或者两年都用 6 月的数据进行比较，最好两年都用全年的数据进行比较，来说明 2016 年

的空气质量是否比 2015 年有明显的改善。

统计结论的得出不是简单的数据比较，而是要结合问题的实际背景进行解释。

同一份数据，从不同的角度进行解释也会有不同的结论。比如教材 207 页关于招聘的例子。在实际生活中这样的例子也很多大家可以跟自己的同学讨论一下下面这个例子：

某大学为调研学生在 A，B 两家餐厅用餐的满意度，从在 A，B 两家餐厅都用过餐的学生中随机抽取了 100 人，每人分别对这两家餐厅进行评分，满分均为 60 分. 整理评分数据, 将分数以 10 为组距分成 6 组: $[0,10)$, $[10,20)$, $[20,30)$, $[30,40)$, $[40,50)$, $[50,60)$ ，得到 A 餐厅分数的频率分布直方图，和 B 餐厅分数的频数分布表：



如果从 A，B 两家餐厅中选择一家用餐，你会选择哪一家？说明理由.

用样本估计总体中的误差

教材 201 页边空有一个问题：你认为 14t 这个标准一定能够么？如果

不一定，哪些环节可能会导致结论的差别？

我们得到的 **14t** 这个结论，是来自样本的观测数据。是不一定能保证 **80%**的居民用水不超标的，原因有以下两点：抽样方法未必是合适的；相对于全体居民而言，**100** 的样本量实在是不大，样本代表性可能不够好；

如果我们用频率分布直方图去估计这个数值，得到的结论为 **14.2** 吨，与直接运用原始数据进行运算得到的数据 **13.7** 为什么有差别呢？其实，我们用频率分布直方图计算的样本平均值、样本中位数都是和直接运用原始数据计算得到的结果有差别的，这个差别是如何产生的呢？

运用频率分布直方图进行估计时，我们只能认为数据在每个小组之内是均匀分布的，可事实并非如此，相当于我们人为插入了好多个数值，使得数据达到了均匀分布，结果的差别也就因此而产生了。