前一篇分享了统计学需要掌握的知识，在数据分析过程中，广泛用于数据质量处理，分析模型构建以及数据挖掘。今天这篇文章将详细讲统计学中最基础的描述统计。

试想，当你拿到一份数据会怎么做？二话不说做个图？

此前也无数次强调，拿到数据需要观察数据情况和数据质量，对数据进行描述统计分析，以发现其内在的规律，再选择进一步分析的方法。

什么是描述性统计？

描述性统计分析要对调查总体所有变量的有关数据做统计性描述，主要包括数据的频数分析、数据的集中趋势分析、数据离散程度分析、数据的分布、以及一些基本的统计图形。

常用的指标有均值、中位数、众数、方差、标准差等等。数据的集中趋势一般采用平均值、中位数表示。数据的离散程度一般采用方差、标准差表示。数据的分布情况一般采用直方图表示。

具体概念前一篇有做详解，就不赘述了。接下来我将用 Excel 来分别解释每一种统计方法的应用以及这些统计方法在 Excel 中的实现方式。

Excel 数据分析工具库

专业的统计分析工具有 SPSS，R 或 Python，但对于大部分新手一天两天比较难上手。永远不要忘记万能的 Excel，Excel 2016 里自带以了一个统计分析工具——“分析工具库”。实际上就是一个外部宏（程序）模块，专门为用户提供一些高级统计函数和实用的数据分析工具。

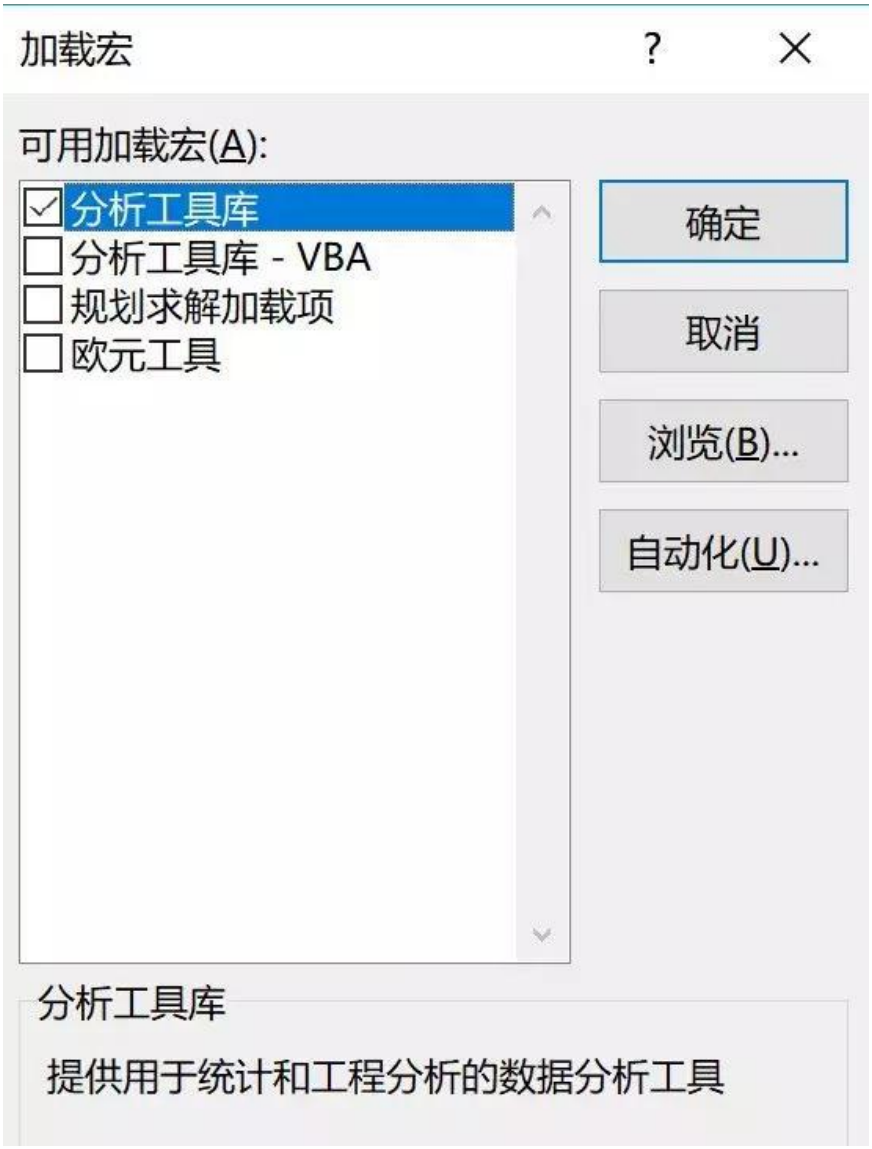
分析工具库内置了 19 个模块，可以分为以下几大类：

分类	工具模块
抽样设计	随机数发生器
	抽样
数据整理	直方图
参数估计	描述统计
	排位与白费排位
假设检验	z-检验：双样本均值差检验
	t-检验：平均值的成对二仰恩分析
	t-检验：双样本等方差假设
	t-检验：双样本异方差假设
	F检验：双样本方差检验
方差分析	方差分析：单因素方差分析
	方差分析：无重复双因素方差分析
	方差分析：可重复双因素方差分析
相关与回归分析	相关系数
	协方差
	回归
时间序列预测	移动平均
	指数平滑
	傅里叶分析

加载 EXCEL 分析工具库

首先你得要有 Excel 2016 。（ 文末有获取方式 ）

安装好 2016 版后，文件—选项—切换到“加载项”选项卡，在“管理”下拉列表中选择“EXCEL 加载项”选项，单击“转到”按钮，跳转到如下“加载宏”对话框，勾选“分析工具库”复选框，再单击“确定”按钮



以上一波操作后，“数据”选项卡中会显示出添加的“数据分析”功能。

案例分析：

现在有一份北京房价数据：

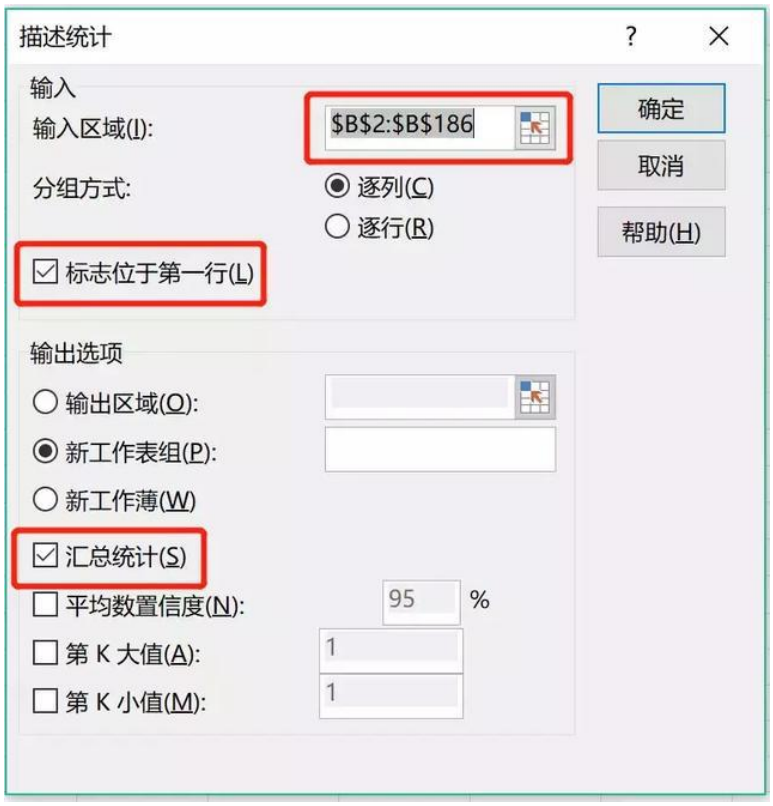
1）北京市政府为调控房地产价格，希望知道北京各小区房屋价格的分布，请分析房地产价格的集中趋势，并选择合适的图形呈现。

2）房地产商想知道北京各个环线房屋装修状况的对比情况，以便进行产品设计和市场拓展，计算指标并设计合适的图形呈现结果，最后给房地产商一些建议。

3）选择合适的图形反映北京各个区住宅区房屋分布情况

操作步骤：

- 基本描述统计打开 excel 数据文件
- 选择描述统计，单击“确定”按钮。



随后，就会生成如下的统计分析结果，就省得一个个函数去计算了。

Avgprice	
平均	3.182772
标准误差	0.095965
中位数	2.9
众数	2.5
标准差	1.30173
方差	1.694502
峰度	3.788973
偏度	1.574287
区域	7.8
最小值	0.8
最大值	8.6
求和	585.63
观测数	184

直方图

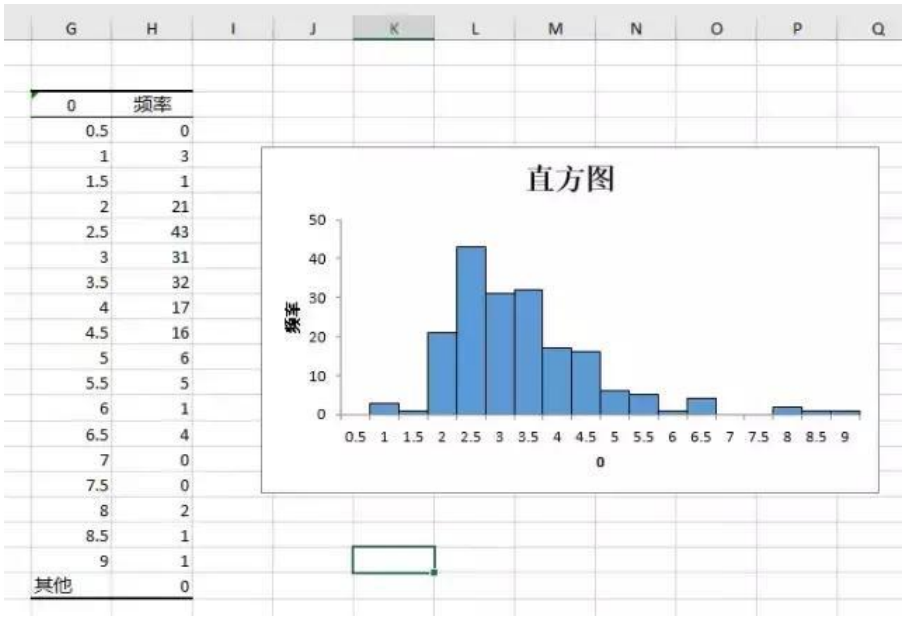
根据描述统计的结果，在空白列构造间隔为 0.5 的等差数列作为接收区域 D1:D19，最大值为 9,最小值为 0。

	A	B	C	D	E
1	Avgprice			0	
2				0.5	
3	平均	3.182772		1	
4	标准误差	0.095965		1.5	
5	中位数	2.9		2	
6	众数	2.5		2.5	
7	标准差	1.30173		3	
8	方差	1.694502		3.5	
9	峰度	3.788973		4	
10	偏度	1.574287		4.5	
11	区域	7.8		5	
12	最小值	0.8		5.5	
13	最大值	8.6		6	
14	求和	585.63		6.5	
15	观测数	184		7	
16				7.5	
17				8	
18				8.5	
19				9	
20					

选择数据，单击“数据”选项卡，选择“数据分析”选项框中的“直方图”选项

输入区域选择房屋价格 avgprice 列$B$2:$B$186 ,接收区域选择第一步构造的接收数据 ,即 D1:D19 数据。

输出区域选择 G3，勾选图表输出，然后单击“确定”按钮。



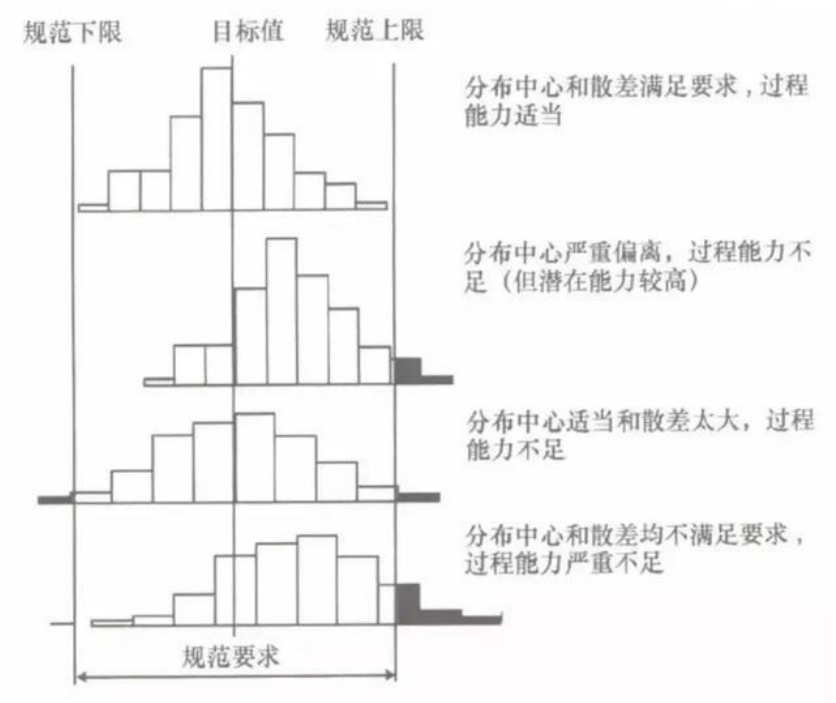
0	频率
0.5	0
1	3
1.5	1
2	21
2.5	43
3	31
3.5	32
4	17
4.5	16
5	6
5.5	5
6	1
6.5	4
7	0
7.5	0
8	2
8.5	1
9	1
其他	0

选中整个直方图，右键单击选择“设置数据系列格式”，单击“系列选项”，分类间距设为 0。

可以看出，北京的房价普遍分布在 2W~4.5W，2.5W 占绝大多数。

关于直方图

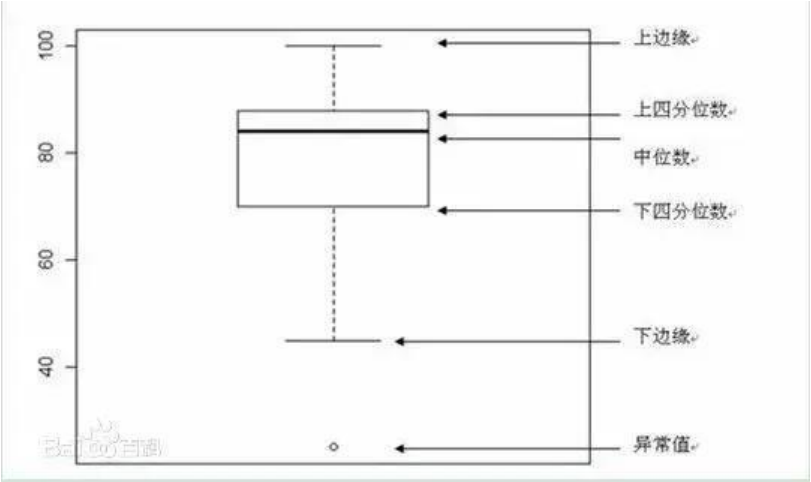
直方图是描述统计中很常见的一个应用，不同直方图代表的业务意义不同。



箱型图

对于数据的离散情况，还有一个更直观的方法，就是箱线图。箱线图利用 6 个指标描述数据的离散情况。这 6 个指标分别是最小值，第一四分位数、中位数、第三四分位数与最大值和异常值。

- 中位数：中位数是一组从小到大排序数据中位置在最中间的一个数据（两个数据取均值）。
- 第 1(下)四分位数：第 1 四分位数与中位数算法类似，是对一组数据中 50%数据再取中位数。一组数据中如果有 25%的数据小于这个数，那么这个数是第 1 四分位数。
- 第 3(上)四分位数：一组数据中如果有 75%的数据小于这个数据，那么这个数是第 3 四分位数。
- 异常值：异常值是指这个数据与四分位数的差达到 5 倍的值。箱线图中异常值的表示方法有两种，1.5 倍-3 倍差之间用空心的点表示。超过 3 倍的异常值，用实心点表示。
- 上限和下限数：除了异常值之外，最靠近上边缘和下边缘的两个数值为上限数和下限数。



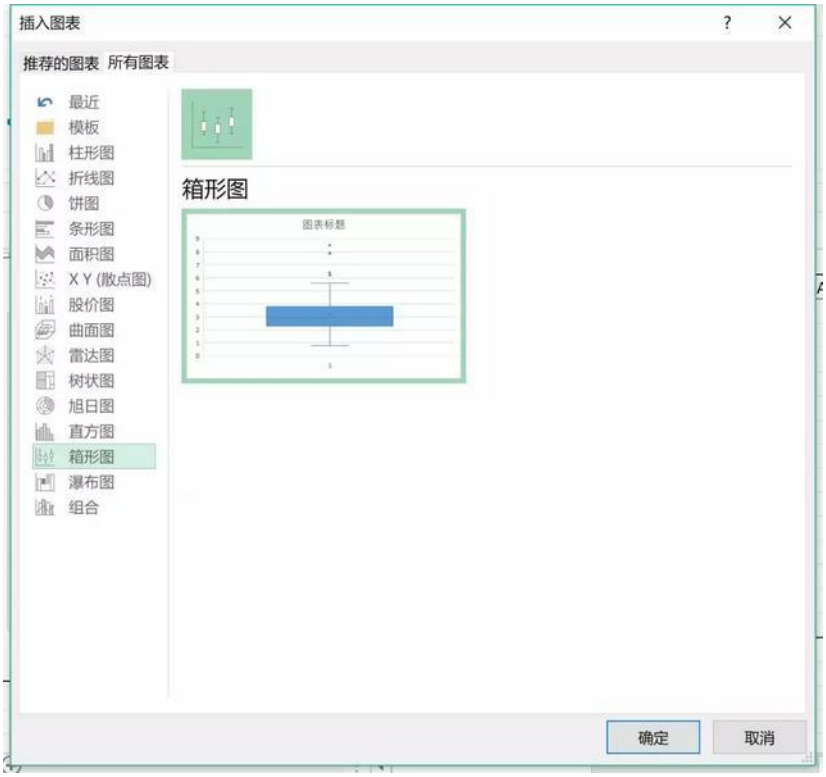
现在来了解北京各区的房价分析，把他加工成箱型图，这也是最常用的描述统计图表。

Excel 2016 可以直接制作箱型图。Excel 的箱型图定位 6 个数据：最大值、最小值、中位数、上四分位数、下四分位数、平均值，还有异常值。

操作步骤：

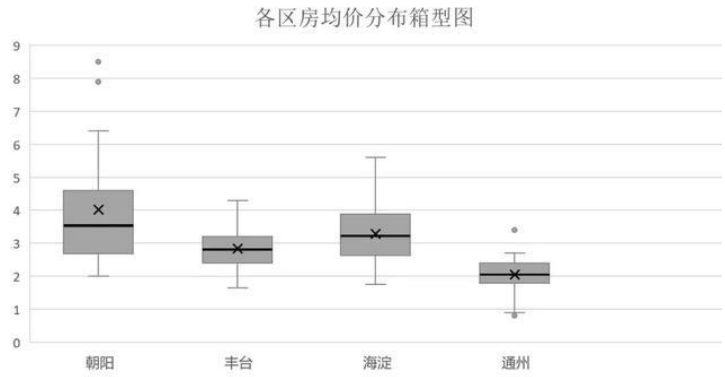
1、选择所要统计的数据，即均价。

2、选择箱型图



3、“选择数据源”中，水平分类轴加上“区域”，如下

调整一下样式得到如下箱型图。



中间黑色出现是各区域中游水平的房价标准（中位数）；x 是全区域的平均房价水平（平均值）；箱型上端代表中上游水平；箱型下端代表中下游水平，以此类推。简而言之，房价分布被四等分了。

我们来解读一下：朝阳区的房价分布范围较广，高低值差异较大，可能和横跨多环有关，整体平均水平位于四区域前列。海淀区平均房价次之，但也不低。丰台区房价分布较为集中且偏态较小，跨度相对较小。通州区很明显整体房价最低。

这张图能一眼看出不少内容，想必大家已经明白箱线图的作用了，它能读出数据的整体分布和倾斜趋势（偏态）。

到这里，描述统计的内容就结束了。描述统计是分析数据的一种技巧，包含数据的集中度量(平均数、中位数、众数)、数据的离散（方差、标准差）、数据的分布（箱线图、条形图、直方图）三块。