原来的 10 倍。显然，这个样本与其他的样本相比是异常值。

- 测量误差：这是异常值最常见的来源。例如，有 10 台称重机，其中 9 台是正常的，1 台是有故障的，那么这台有故障的机器测量的值就是异常值。
- 实验误差：实验误差也会导致出现异常值。
- 有意造成异常值：这通常发生在一些涉及敏感数据的报告中。例如，当要求青少年报告消费的酒精量时，他们可能会上报比真实数据小的值。
- 数据处理误差：在操作或数据提取的过程中造成的误差。
- 采样误差：例如，我们要测量运动员的身高，而样本中包括几名很高的篮球运动员，这种就可能会导致数据集中出现异常值。

异常值对模型和预测分析的影响主要有增加错误方差，降低模型的拟合能力；异常值的非随机分布会降低正态性；与真实值可能存在偏差；影响回归、方差分析等统计模型的基本假设。

举一个简单的例子说明：

对比表 1-2-5 中的两组数据可以发现，有异常值的数据集具有显著的平均值和标准差。在无异常值的情况下，数据集的平均值是 5.45，标准差是 0.99。在加入一个异常值后，数据集的平均值上升为 30，标准差则升至 81.41，这将彻底改变估计。

表 1-2-5

无异常值	有异常值
4,4,5,5,5,5,6,6,6,7,7	4,4,5,5,5,5,6,6,6,7,7,300
平均值=5.45	平均值=30.00
中位数=5.00	中位数=5.50
众数=5.00	众数=5.00
标准差=0.99	标准差=81.41

2. 异常值的检测

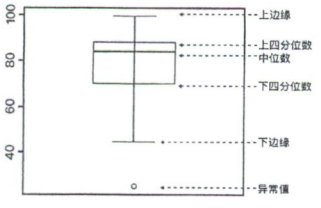

图 1-2-6　箱线图

一般可以采用可视化方法进行异常值的检测，常用工具包括箱线图、直方图、散点图等。

如图 1-2-6 所示，箱线图是一种很好的可视化工具，常用于可视化基本的统计数据，如异常值、最小值、最大值、四分位数等。

利用箱线图检测异常值的原则如下：

不在-1.5*IQR 和 1.5*IQR 之间的样本点认为是异常值，如图 1-2-7 所示。

使用封顶方法可以认为在第 5 和第 95 百分位数范围之外的任何值都是异常值。

距离平均值为三倍标准差或更大的数据点可以被认为是异常值。

下面举例进一步说明异常值的检测。例如，一组客户的年收入是 80 万美元，但是其中两个客户的年收入分别为 400 万美元和 420 万美元，如图 1-2-8 所示。由于这两个客户的年收入