



描述性统计学



目录

CONTENTS

01

单变量

02

双变量

03

数值方法

04

仪表盘





01

描述统计：单变量



4

数据类型

分类型变量

用标签或名称来识别项目的类型。比如：性别、国家、种类等

数量型数据

表示多少或大小的数值。比如：年龄、身高、体重、重量等

频数分布

频数分布是一种数据的表格汇总方法，表示几个互不重叠组别中，每一组项目的个数（即频数）。

条形图

用来描绘已汇总的分类型数据的频数分布，相对频数分布。

相对频数分布

对每一组的项目所占的比例或百分比更感兴趣。一组的相对频数是属于该组别的项目个数占总数的比例

饼形图

是一种描绘分类型数据的相对频数和百分数频数分布的图形方法。



购买饮料的频数分布

	频数
软饮	
百事可乐	16
芬达	7
果粒橙	13
可口可乐	24
雪碧	12
怡宝	55
总计	127

定义

频数分布式一种数据的表格汇总方法，表示几个互不重叠组别中，每一组项目的个数（即频数）。

观测

从频数分布表中我们可以看出每种饮料的分布详情，可以得到饮料受欢迎程度的信息。

购买饮料的相对频数分布

行标签	频数	相对频数
百事可乐	16	0.125984
芬达	7	0.055118
果粒橙	13	0.102362
可口可乐	24	0.188976
雪碧	12	0.094488
怡宝	55	0.433071
总计	127	1

定义

对每一组的项目所占的比例或百分比更感兴趣。一组的相对频数是属于该组别的项目个数占总数的比例

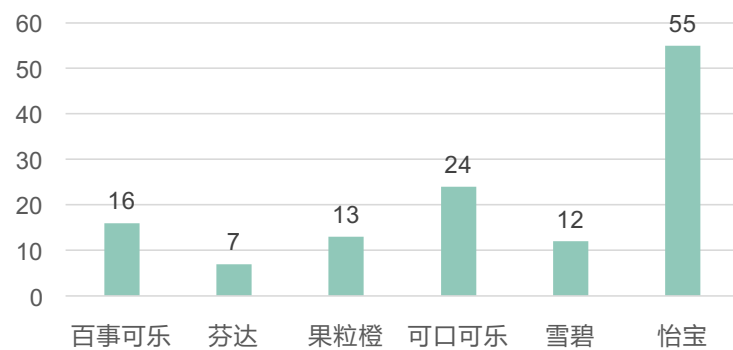
相对频数

$$\frac{\text{组的频数}}{n}$$

8

条形图

频数分布条形图



定义

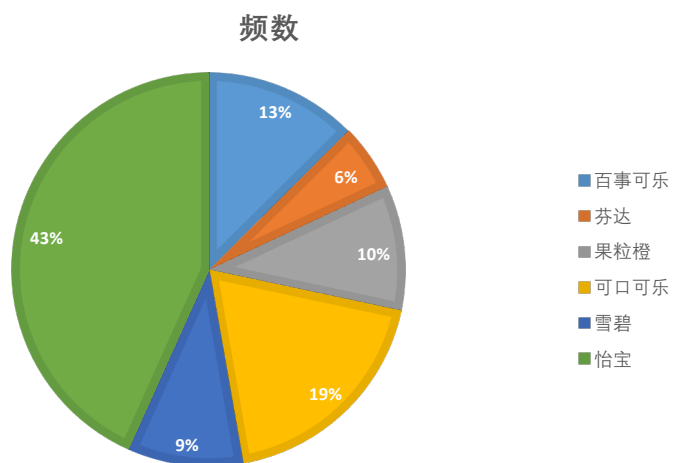
用来描绘已汇总的分类型数据的频数分布，相对频数分布。

画法

在横轴上用来对数据分组做标记，纵轴上标出频数，相对频数的刻度。用一个固定宽度的长条绘制在每一组的标记上，将这个长条的高度延伸，直到达到该组的频数。

解读

怡宝是最受欢迎的饮料，芬达是最不受欢迎的饮料。



定 义

是一种描绘分类型数据的相对频数和百分数频数分布的图形方法。

画 法

使用相对频数对圆进行切分成若干扇形，这些扇形与每一组的相对频数对应。

NBA球员评分

Rank	Player	PPG
1	LeBron James, MIA	27
2	Kevin Durant, OKC	28.8
3	James Harden, HOU	26.4
4	Kobe Bryant, LAL	27.1
5	Russell Westbrook, OKC	22.9
6	Carmelo Anthony, NY	28.4
7	David Lee, GS	19.2
8	Stephen Curry, GS	21
9	LaMarcus Aldridge, POR	20.8
10	Paul George, IND	17.6

确定组数

组是通过对数据规定范围形成的，这个规定的范围作用于对数据进行分组。分组的目的是用足够多的组来显示数据的变异性。一般性原则：5~20组即可

确定组宽

一般性原则：每组的宽度相同。组宽和组数是互相依赖的，较大的数组意味着较小的组宽，反之亦然。

$$\text{近似组宽} = \frac{\text{数据最大值} - \text{数据最小值}}{\text{组数}}$$

确定组限

选组组限必须使每一个数据值属于且仅属于一组。

1

2

3



11

频数分布

球员分数的频数分布

分数分布	频数
(15.975, 18.112]	20
(18.112, 20.25]	8
(13.837, 15.975]	7
(26.662, 28.8]	4
(20.25, 22.388]	4
(11.682, 13.837]	4
(22.388, 24.525]	2
(24.525, 26.662]	1

```
import pandas as pd
df = pd.read_csv("NBAPlayerPts.csv")
df['group'] = pd.cut(df['PPG'], bins = 8, right=True, include_lowest = True)
pd.DataFrame(df['group'].value_counts())
```

确定组数

假设置为8组。

确定组宽

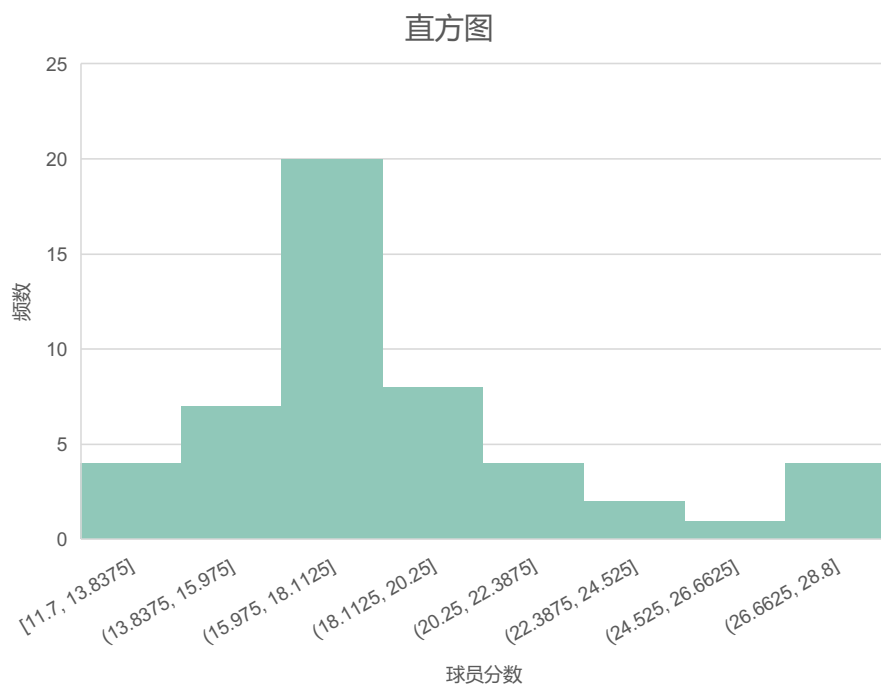
最大值: 28.8, 最小值11.7

组宽: $\frac{28.8-11.7}{8} = 2.1375$

组 限

12

直 方 图



直 方 图

常见的数量型数据的图形描述法。由先前已汇总的频数分布等数据进行绘制。变量标签放置在横轴上，频数放置在纵轴上。

对 比

条形图的类别之间都是分隔的，相互独立。但是直方图类别之间是没有间隔的，是连续的，代表是数值变量。

意 义

直方图更重要的意义是表明一个变量的偏态分布，能够知道数据的分布趋势，我们会在概率章节重点讲解。

球员分数	累积频数	累计相对频数
(11.0, 14.0]	4	0.08
(14.0, 16.0]	11	0.22
(16.0, 18.0]	31	0.62
(18.0, 20.0]	39	0.78
(20.0, 22.0]	43	0.86
(22.0, 25.0]	45	0.90
(25.0, 27.0]	46	0.92
(27.0, 29.0]	50	1.00

累积频数分布

表明的是小于或者等于每一组上组限的数据项个数。

累积相对频数分布

表示数值小于或等于每一组上组限的数据项的比例。
数据表明有92%的球员得分不超过27分

条形图与直方图本质上是同一事物，他们都是频数分布数据的图形表示

开口组是指只有一个下组限或上组限的组




对于数量型数据，适当的组限依赖于数据的精度水平

累积频数分布的最后一个数据项总等于观测值的总数



02

描述统计：双变量



300家洛杉矶饭店质量等级和餐价

(Restaurant.csv)

Restaurant	Quality Rating	Meal Price (\$)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23

交叉分组表

是一种汇总两个变量数据的方法。两个变量可以是分类的或是数量的。但是更常见的是一个为分类一个为数量

数据应用

如数据表中所示，质量是分类变量，价格是数量变量，我们对这两个变量进行交叉表计算

质量等级	餐价				总价	
	(9.0, 20.0]	(20.0, 29.0]	(29.0, 38.0]	(38.0, 48.0]		
Excellent	2	14	28	22	66	
Good	42	40	2	0	84	
Very Good	34	64	46	6	150	
总计	78	118	76	28	300	

交叉分组表

左边栏和顶部边栏的标记确定了两个变量的组别。样本中的每个饭店都会落在交叉表的一个单元格里。

表格解读

从表中可以看出，样本质量等级很好且餐价在20~29的饭店最多。

作用意义

交叉分组表的主要价值在于提供了变量间关系的深刻含义，它揭示了较高餐价与较高质量等级相关，而较低的餐价对应于较低的质量等级。

质量等级	相对频数
Excellent	0.28
Good	0.5
Very Good	0.22
总计	1

质量等级变量的相对频数表

餐价	相对频数
(9.0, 20.0]	0.26
(20.0, 29.0]	0.39
(29.0, 38.0]	0.25
(38.0, 48.0]	0.09
总计	1

餐价变量的相对频数表

散点图和趋势图

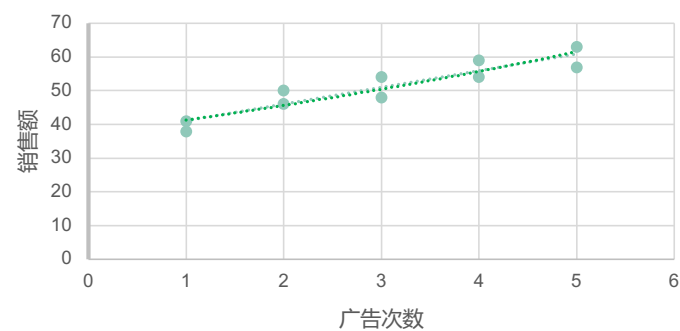
音像设备商店的样本数据		
Week	No. of Commercials	Sales Volume
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

图形定义

散点图是对两个数量变量间关系的图形表述

趋势图是显示相关性程度的一条直线

散点图与趋势图

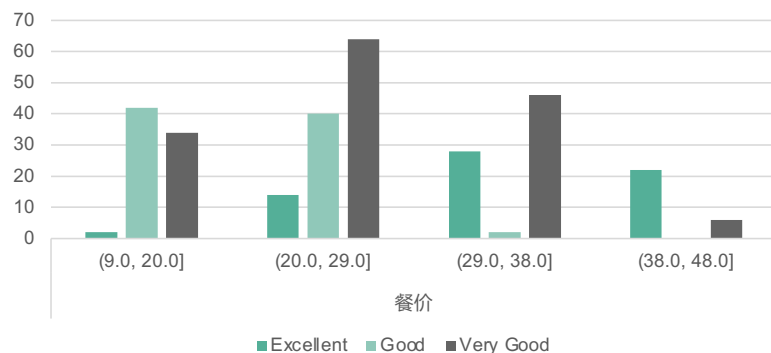


散点图与趋势图

图形表明，广告次数和销售额之间存在正相关关系。较高的销售额与较高的广告次数相联系

质量等级	餐价				总价
	(9.0, 20.0]	(20.0, 29.0]	(29.0, 38.0]	(38.0, 48.0]	
Excellent	2	14	28	22	66
Good	42	40	2	0	84
Very Good	34	64	46	6	150
总计	78	118	76	28	300

质量等级与餐价数据的关系



定义

对于已经汇总的多个条形图同时显示的一种图形显示方法

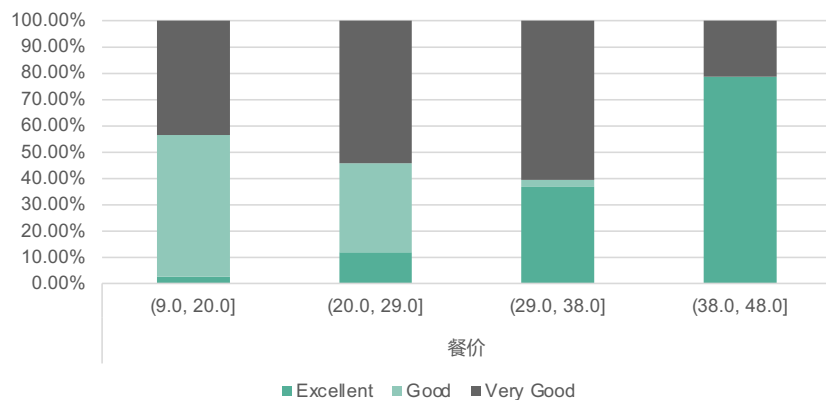
解读

我们可以看到在最低价区间，大部分餐厅都可以得到好与很好的评价，随着价格的升高，餐厅得到优秀的评价变多，尤其是价格最高的区间饭店，都是优秀和很好的评价。

同时我们关注到随着价格的升高，“好”评价在减少，“优秀”评价在增加。表明了随着价格的升高，质量也趋于升高

质量等级	餐价			
	(9.0, 20.0]	(20.0, 29.0]	(29.0, 38.0]	(38.0, 48.0]
Excellent	2.56%	11.86%	36.84%	78.57%
Good	53.85%	33.90%	2.63%	0.00%
Very Good	43.59%	54.24%	60.53%	21.43%
总计	100.00%	100.00%	100.00%	100.00%

质量等级与餐价数据的关系



结构条形图

每一个长条被切割成不同颜色的矩形段，每一个段都表明一个频数。


解 读

从图中我们可以更容易的观察到，随着价格的升高，“好”长条的长度在减少，而“优秀”长条的长度在增加。



03

描述统计-数值方法





位置的度量



五数概括法和箱形图



变异程度的度量



两变量关系的度量

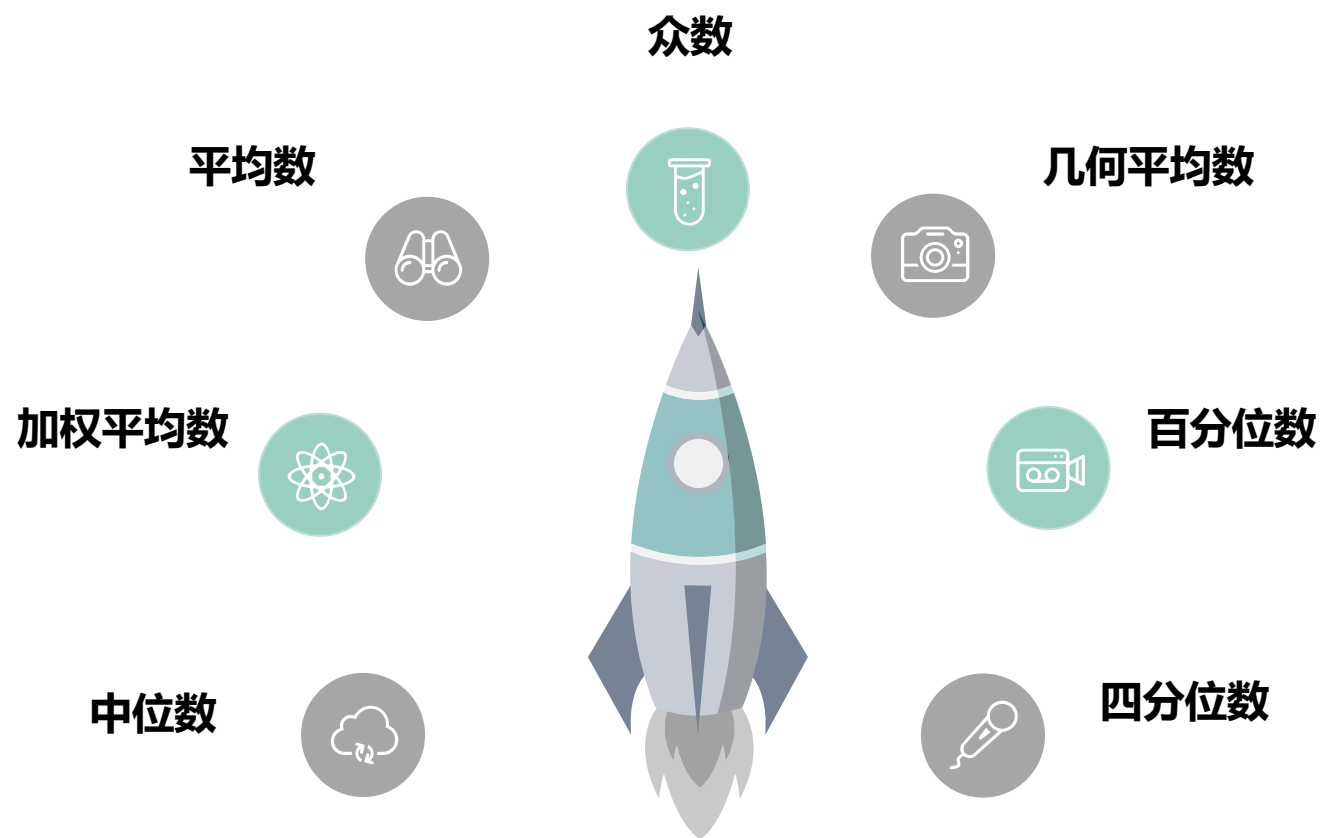


分布形态

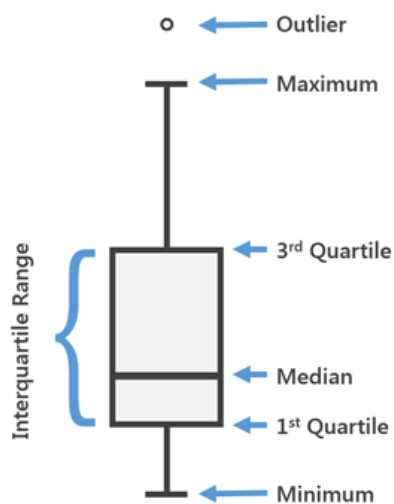


数据仪表盘









定义

百分位数提供了数据如何散布在从最小值到最大值的区间上的信息。第 p 个百分位数是满足下列条件的一个数值：至少有 $p\%$ 的观测值小于或者等于该值，且至少有 $(100-p)\%$ 的观测值大于或者等于该值

计算方式

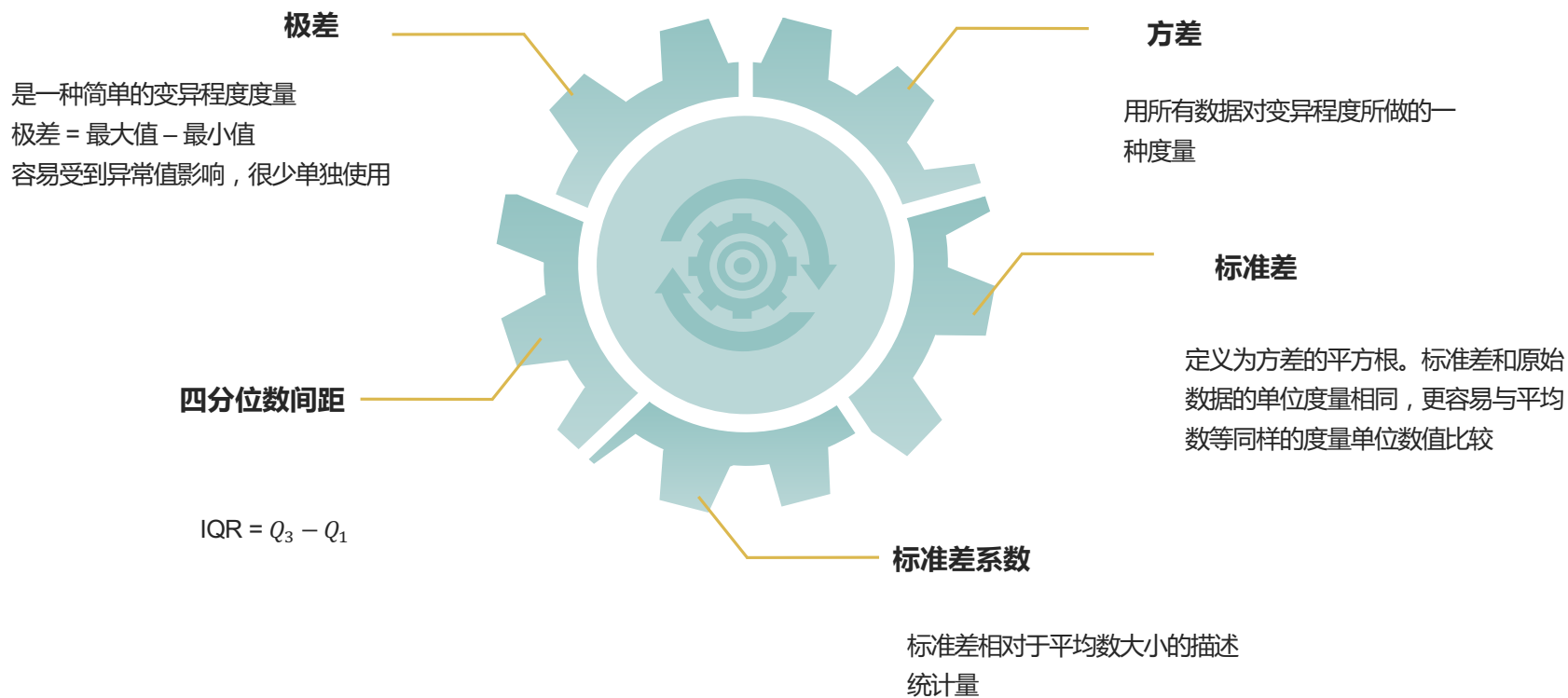
1. 把数据从小到大排序
2. 计算指数 $i = (p/100)n$
3. p 是所求的百分位数， n 是观测值的个数
4. 若 i 不是整数，则向上取整。大于 i 的下一个整数表示第 p 百分位数的位置。若 i 是整数，则第 p 百分位数是第 i 项和第 $i+1$ 项数据的平均数

四分位数

- $Q_1 = 25$ 百分位数
- $Q_2 = 50$ 百分位数
- $Q_3 = 75$ 百分位数

假如你向两个供应商采购货物。两个供应商答应的供货时间都是 10 天，但是它们在按时交货方面是否拥有相同的可信度呢？

我们可以通过两家公司历史交货天数的数据，计算它们各自的变异程度，尽可能选择变异程度少并且控制在10天之内！



方差

用所有数据对变异程度所做的一种度量。方差依赖于每个观察值与平均值之间的差异
拥有较大方差的变量显示其变异程度较大

$$\text{总体方差: } \sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

$$\text{样本方差: } s^2 = \frac{\sum(x_i - \mu)^2}{n-1}$$

标准差

定义为方差的平方根。
标准差和原始数据的单位度量相同，更容易与平均数等同样的度量单位数值比较

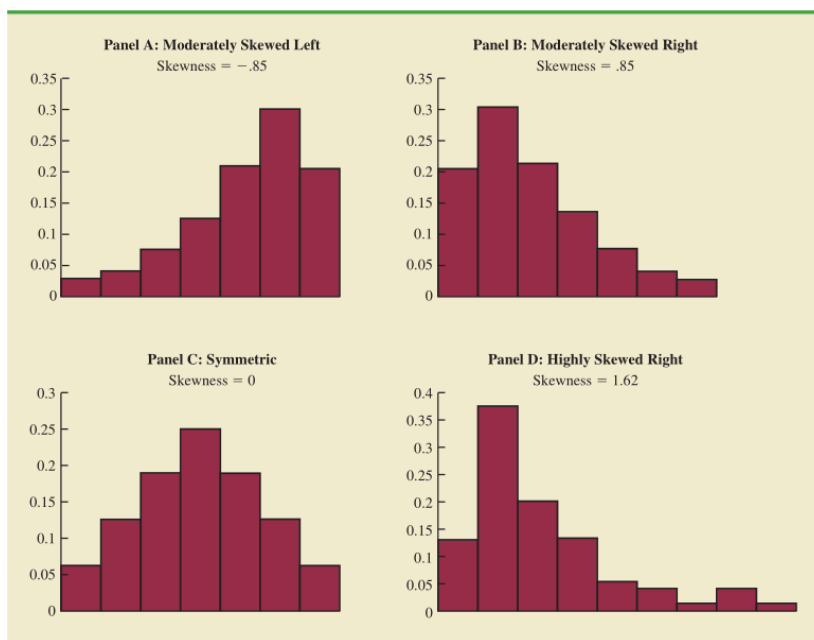
$$\text{样本标准差: } s = \sqrt{s^2}$$

$$\text{总体标准差: } \sigma = \sqrt{\sigma^2}$$

标准差系数

标准差相对于平均数大小的描述
统计量

$$\left(\frac{\text{标准差}}{\text{平均数}} \times 100 \right) \%$$



介绍

前面的直方图对分布形态提供了一种很好的图形描述。分布形态的一种重要的数值度量称为偏度(skewness)

解读

- 当数据偏度是正数时，平均数比中位数要大
- 当数据偏度是负数时，平均数比中位数要小
- 当数据严重偏离时，中位数是位置的首选度量

Restaurant	Quality Rating	x	x - \bar{x}	z-score
1	Good	18	-5.3	-0.61886
2	Very Good	22	-1.3	-0.1518
3	Good	28	4.7	0.5488
4	Excellent	38	14.7	1.71646
5	Very Good	33	9.7	1.13263
6	Good	28	4.7	0.5488
7	Very Good	19	-4.3	-0.50209
8	Very Good	11	-12.3	-1.43622
9	Very Good	23	-0.3	-0.03503
10	Good	13	-10.3	-1.20269

定义

z-分数被认为是对数据集中观测值相对位置的量度，帮助我们确定一个数值距平均数有多远

$$z_i = \frac{x_i - \bar{x}}{s}$$

意义

z-分数常被称为标准化数值，它能被解释为 x_i 与平均数的距离是 z_i 个标准差。比如第二个饭店的标准化分数为-0.15，这表明该样本的值比平均数少0.15个标准差。

两个不同的数据集的观测值具有相同的z分数，就可以说明他们具有相同的相对位置。

标准化还有一层作用：可以消除不同量纲单位带来的影响

切比雪夫的案例：假设某大学有100名学生的考试成绩平均分为70分，标准差为5分。那么有多少学生的考试成绩在60~80分？

意义

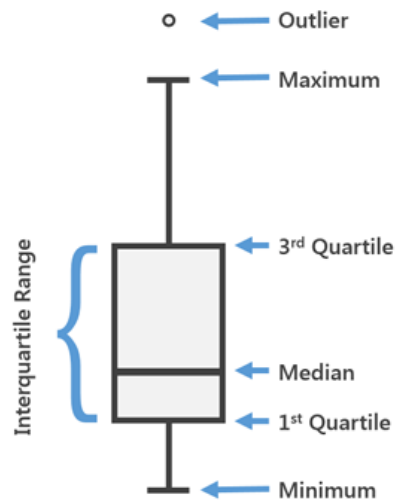
能使我们指出与平均数的距离在某个特定个数的标准差之内的数据值所占的比例

定理

与平均数的距离在 z 个标准差之内的数据值所占比例至少为 $(1 - 1/z^2)$,其中 z 是大于1的任意实数。

经验应用

- 至少75%的数据值与平均数的距离在 $z=2$ 个标准差之内
- 至少89%的数据值与平均数的距离在 $z=3$ 个标准差之内
- 至少94%的数据值与平均数的距离在 $z=4$ 个标准差之内

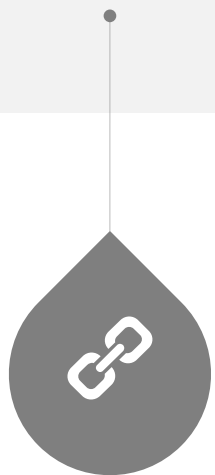


定义

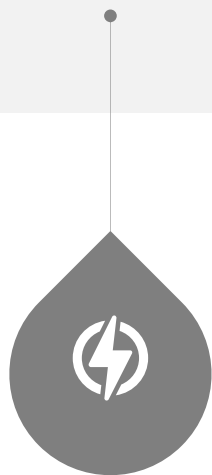
一个或者多个数值异常大或异常小的观测值，这样的极端值被称为异常值。异常值的原因：
错误记录数值、一个错误的数、反常的数据值

计算方式

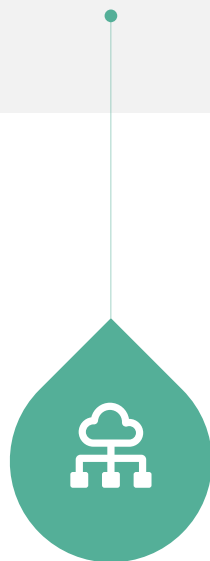
- 下限 = $Q_1 - 1.5 \times \text{IQR}$
- 上限 = $Q_3 + 1.5 \times \text{IQR}$
- 观测值大于上限或者小于下限就被归类为异常值



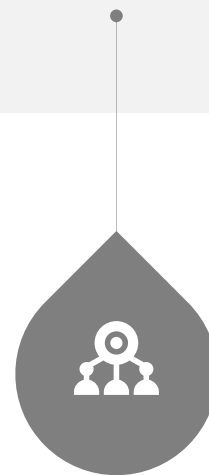
最大值



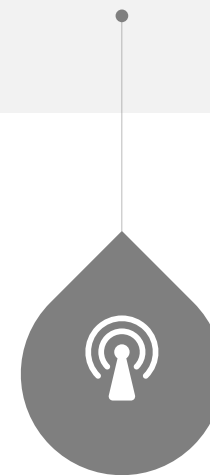
第一四分位数



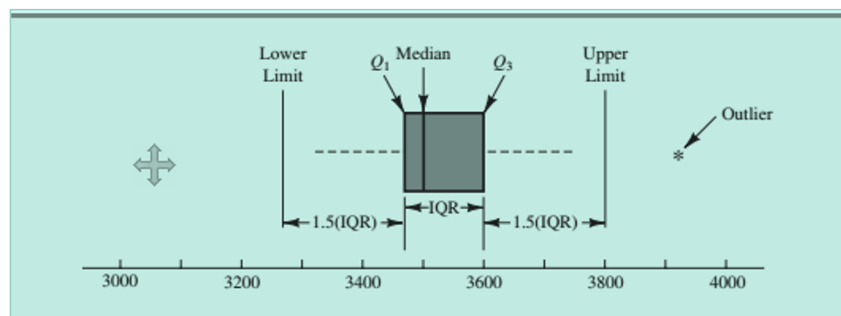
中位数



第三分位数



最小值

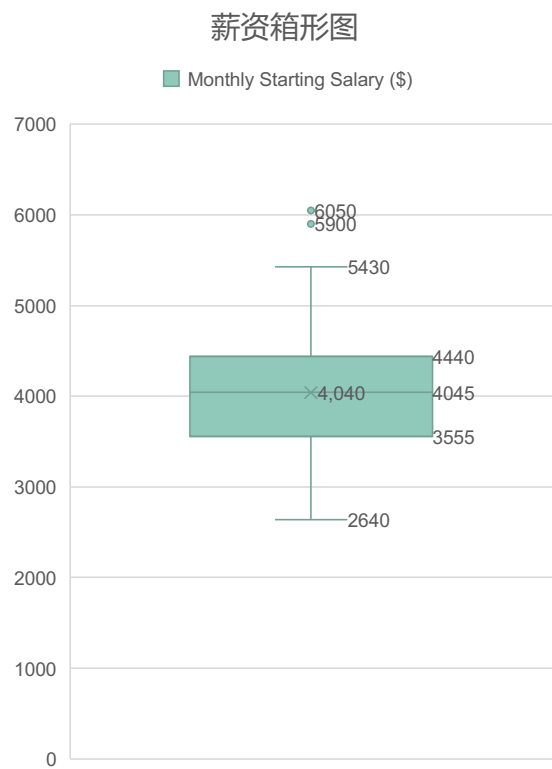


箱形图

是基于五数概括法的数据图形汇总。关键是计算四分位数： $IQR = Q_3 - Q_1$ 。箱体中代表的是中位数、 Q_1 、 Q_3

异常值

超过 Q_1 或者 Q_3 的1.5倍IQR 就被认为是异常值，如果所示。



解 读

可以看出薪资的中位数是4040，异常值为5900与6050。

各专业起始月薪的箱形图



比较

针对不同专业毕业的学生的月薪进行对比。每一个箱形图出现在对应专业的纵轴上

分析

- 会计专业起薪较高，而管理和市场营销的起薪较低
- 会计、金融、和市场营销专业存在异常值
- 根据中位数，会计、信息系统专业有着较高的数值，金融次之，其它专业较低

协方差

$$\text{样本协方差 } s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\text{总体协方差 } s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$$

协方差的解释

表象：一个大的正数表示强的正线性相关关系；一个大的负数表示强的负线性相关关系。

缺点：协方差依赖于x和y的计量单位

广告次数与销量

Week	x	y	x-x̄	y-ȳ	(x-x̄)*(y-ȳ)
1	2	50	-1	-1	1
2	5	57	2	6	12
3	1	41	-2	-10	20
4	3	54	0	3	0
5	4	54	1	3	3
6	1	38	-2	-13	26
7	5	63	2	12	24
8	3	48	0	-3	0
9	4	59	1	8	8
10	2	46	-1	-5	5
合计	30	510	0	0	99

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{99}{10-1} = 11$$

相关系数

皮尔逊相关系数：

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

r_{xy} 相关系数

s_{xy} 协方差

s_x x的标准差

s_y y的标准差

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{1.49 \times 7.93} = 0.93$$

相关系数的解释

相关系数的值在 -1 至 1 之间，其绝对值越大，表明变量之间的相关性越大，正负号代表的相关性的方向。

$r_{xy} = 0.93$ 表明广告次数和销售额之间存在着强的线性关系。

注意：相关性系数提供的是两个变量之间关联性的度量，并不意味着他们之间存在因果关系！