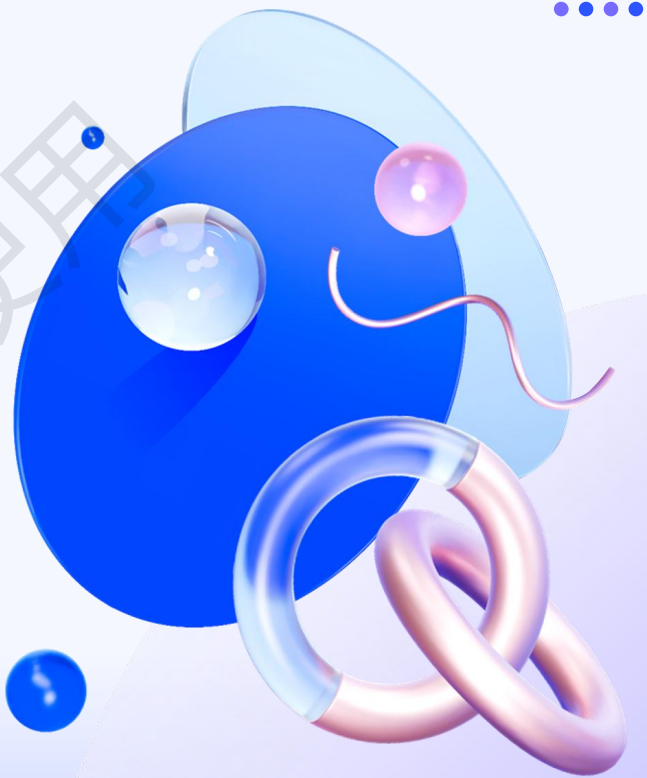
描述统计分析基础

万川



掌握描述统计的分析技能

CONTENTS

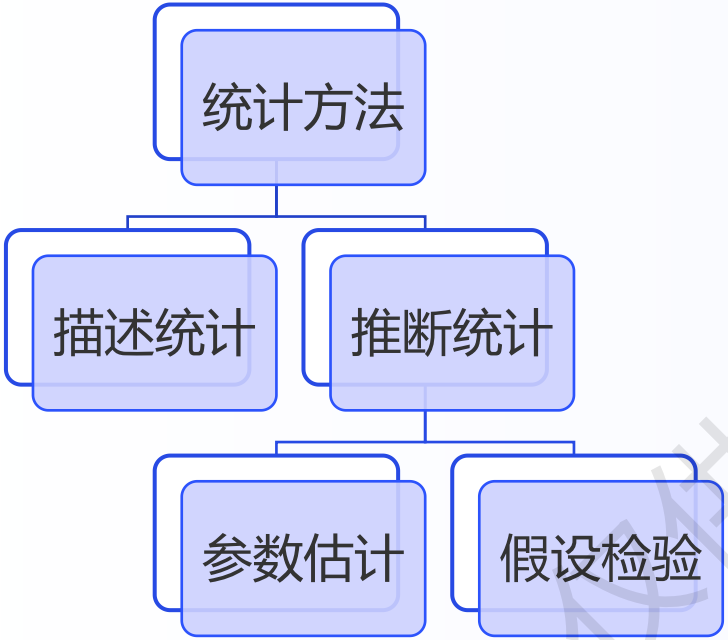
目录

PART1 含义解读

- 了解统计方法的结构要素
- 了解描述统计在跨境电商数据化运营中的作用
- 形成描述统计分析的逻辑思维

◆ 锻炼数据分析逻辑思维能力

1.1 统计方法结构



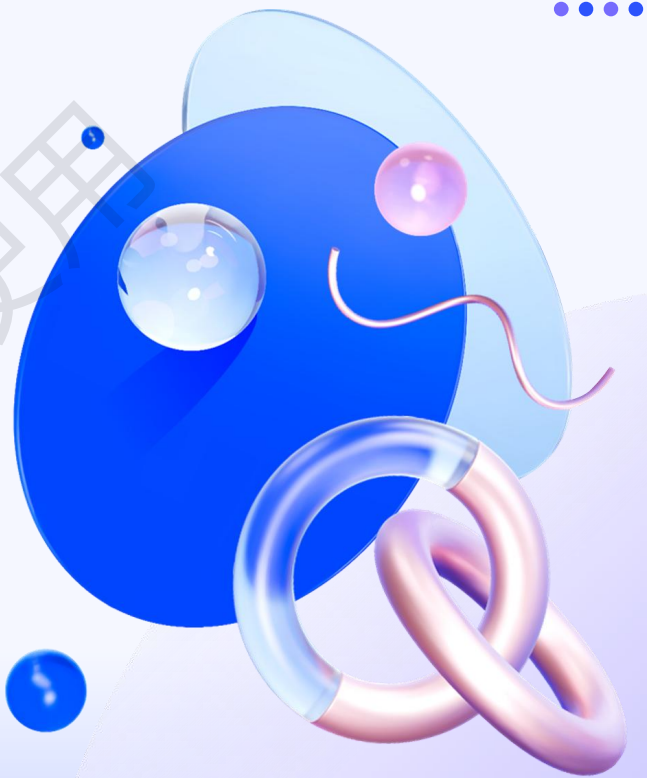
- 统计方法的结构中，主要分为描述统计和推断统计两大类。而推断统计下面又分参数估计与假设检验。
- 描述统计主要关注如何收集、处理、展示数据，并通过这些数据来描述或总结观察量的基本情况。
- 推断统计则使用样本数据来预测和推断总体特征。
- 参数估计涉及从样本数据推断总体的情况，包括点估计和区间估计。
- 假设检验则是通过样本统计量得出的差异作出一般性结论，判断总体参数之间是否存在差异。
- 总结而言，描述统计用于总结和描述现有数据，而推断统计则基于样本数据对总体进行预测和推断。

1.2 描述统计作用

描述统计是统计学的基础，在跨境电商数据分析中具有十分重要的作用：

- 数据概览：提供数据的快速概览，使决策者对业务状况有一个初步的了解。
- 价格分析：分析不同价格点的销售情况，确定最佳定价策略提高利润率。
- 趋势预测：描述统计本身不直接进行预测，但通过分析历史数据，可发现背后的趋势和模式，为预测分析提供基础。
- 风险管理：通过描述统计识别订单取消率、退货率等指标，商家可以评估运营风险，及时调整策略以降低损失。
- 总之，描述统计在跨境电商数据分析中的应用是多方面的，为商家提供了从宏观到微观的全方位数据视角，是跨境电商决策和运营不可或缺的思维。

PART2 数据特征

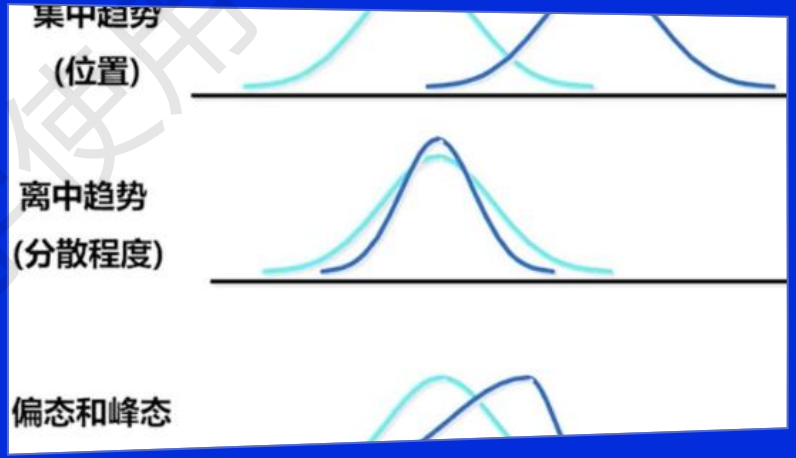


- 了解描述统计中关于数据特征的表现方式
- 熟悉运用EXCEL表格制作各类图表
- 熟悉集中趋势与离散程度的一般指标
- 熟悉EXCEL常用函数的基础操作方法
- 形成熟练进行描述统计的数据分析技能

◆ 锻炼数据分析逻辑思维能力

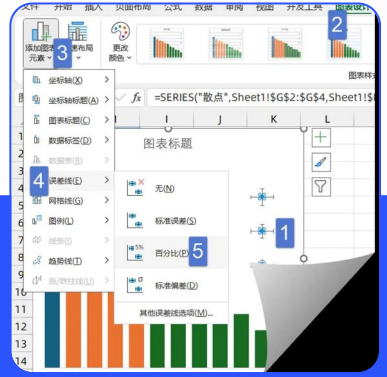
2.1 表现方式

数据可视化，频数分布表、直方图、圆形图、散点图等来直观了解整体数据分布情况



数据的集中趋势（如平均数、中位数、众数）和离散程度（如全距、标准差）的测量

2.2 数据可视化



图表

三线表以及饼状图、柱状图、折线图、雷达图等各类图

工具

Power BI、Tableau、Google Charts等各类制作图表的工具

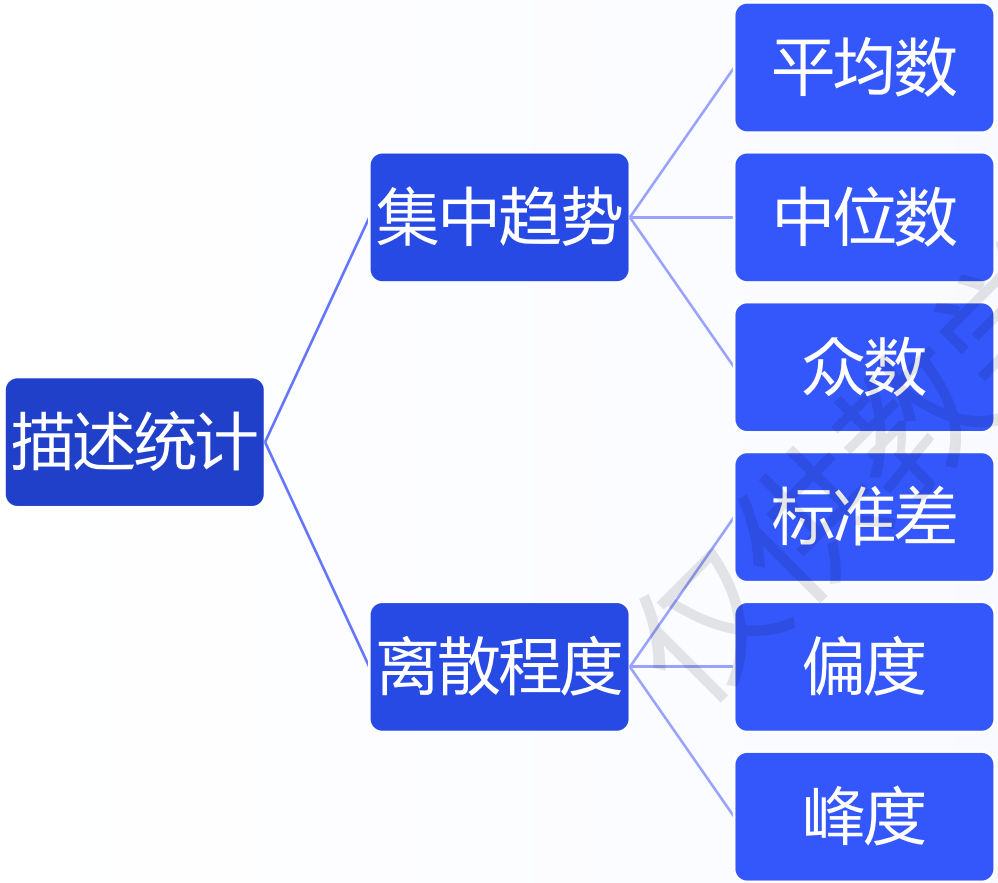
网站

常用的、功能齐全的、开源的网站

https://www.microsoft.com/zh-cn/power-platform/products/power-bi/

https://www.tableau.com/

2.3 集中趋势与离散程度



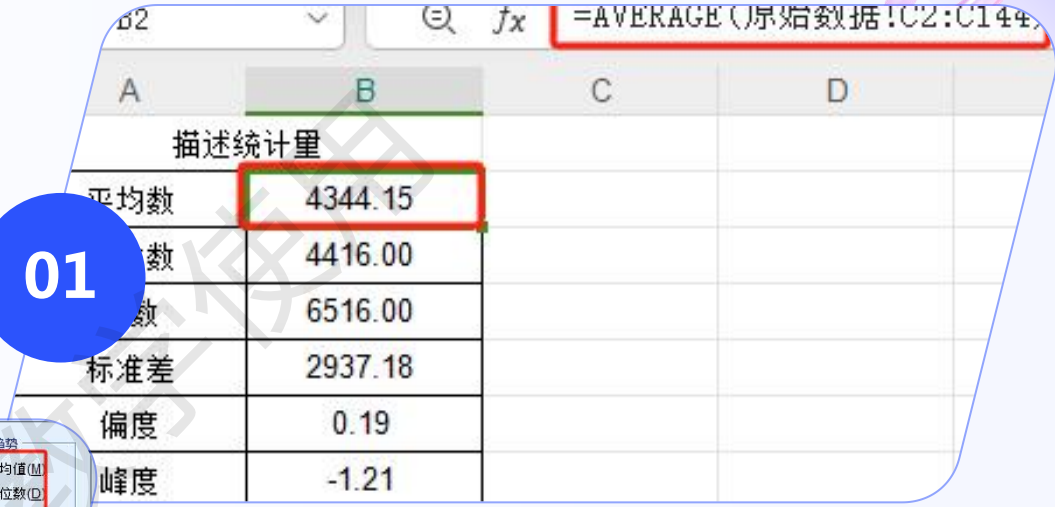
- 平均数：算术平均数，average。
- 中位数：中间的数，median。
- 众数：出现次数最多的数，mode。
- 标准差：观测值与平均数的距离再开方，stdeva。
- 偏度：观测值左右偏离平均数，即偏向走势，skewness。
- 峰度：观测值上下偏离平均值，即极端值情况，kurtosis。

2.4 工具运用

EXCEL，快速计算集中趋势与离散程度的各类指标。

例如，算术平均数、中位数、众数、标准差、偏度、峰度的函数输入方式与使用技巧

01



02

SPSS，快速得出集中趋势与离散程度的各类指标，判断数据的分布特征。

2.5 EXCEL基础函数及应用

01

常用函数

sum/ countif/ if/ datedif/ left/
right/ mid/ int/ round/
randbetween

02

数据处理

行列互换
文本数字转换与提取
缺失值清洗
重复值清洗
时间格式清洗

03

分类汇总与合并计算

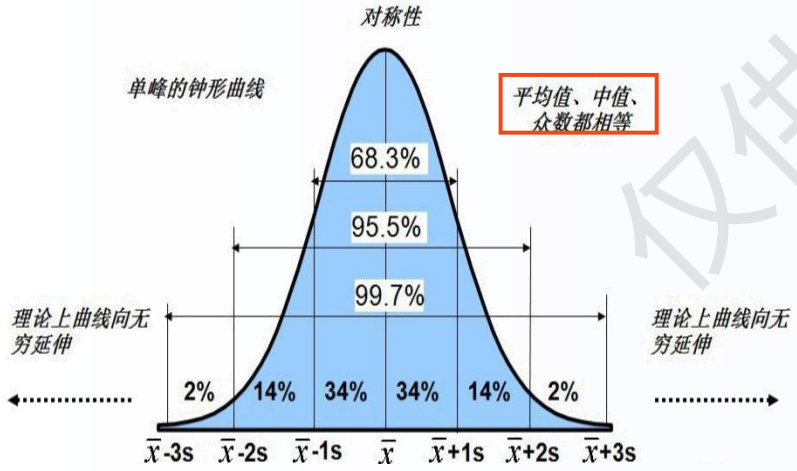
删除分类汇总数据
选择性粘贴数据
合并计算

PART3 数据分布

- 理解正态分布的含义及判断方法
- 理解偏态分布的含义及判断方法
- 形成数据分布的检验分析技能

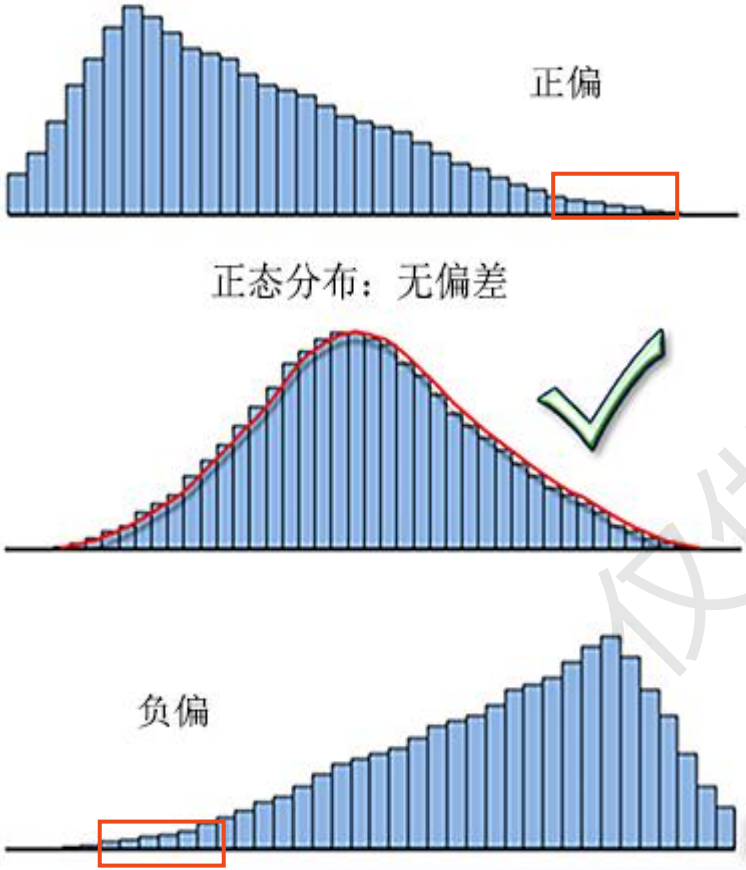
◆ 锻炼数据分析逻辑思维能力

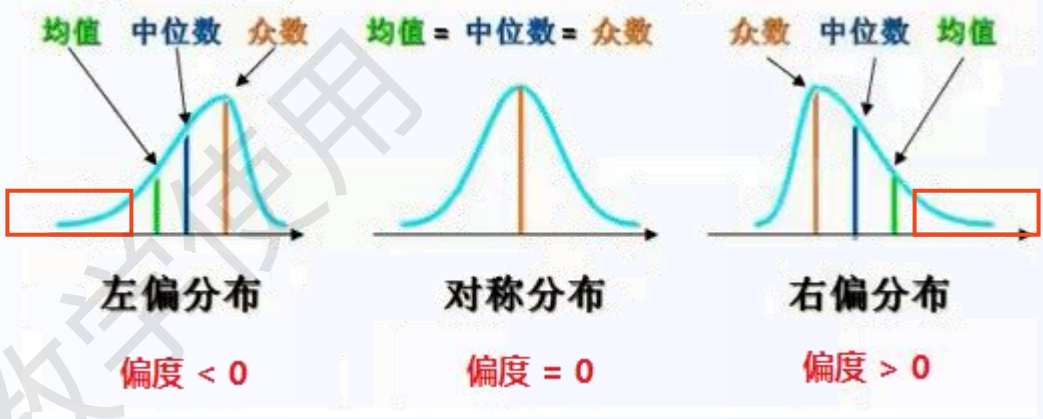
3.1 正态分布



- 德国数学家Gauss（高斯）率先将其应用于天文学研究，故又叫高斯分布。
- 呈现对称的钟形曲线的一种数据分布形式。在平均值附近集中，两边逐渐减少，表示大多数情况或测量结果是平均或典型的，极端值较少出现。（平均值=中位数=众数）
- 标准正态分布，平均值=0，方差=1。
- 推断统计中假设检验提及小概率事件，是指发生的概率小于5%或者1%的事件。它被认为在一次试验中该事件是几乎不可能发生的。
- 描述和预测许多自然和社会现象，是统计学、科学研究和决策制定中重要的工具。例如，消费者的购买力，学生的成绩等等。

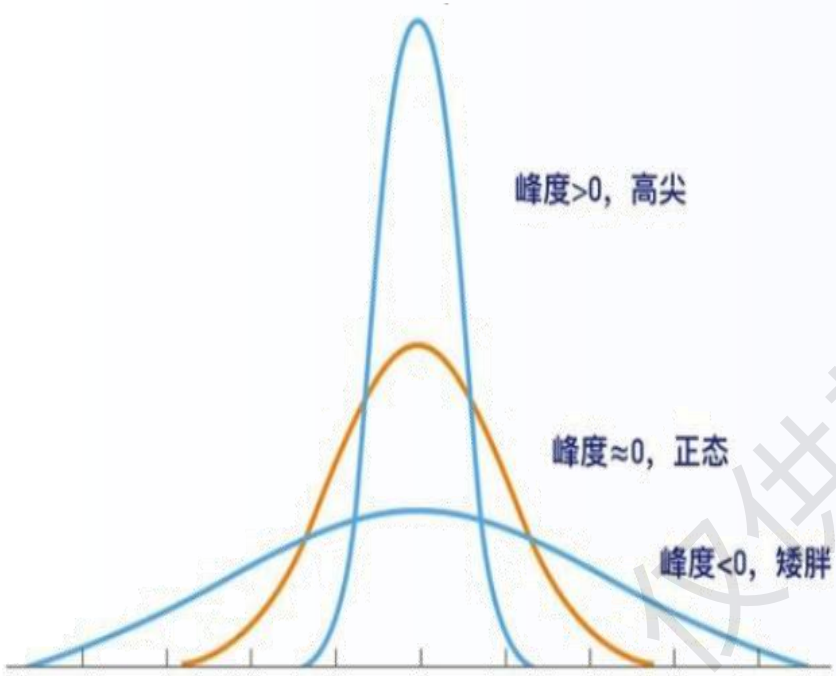
3.2 偏态分布-偏度





- 正偏态与负偏态，左偏态与右偏态。看小尾巴来决定。
- 平均值、中位数、众数三者之间的大小关系。
- 受极端值影响，在敏感性方面，平均值>众数>中位数。
- 数据分布整体偏大，还是偏小，还是近似正态分布（[-1,1]）。

3.2 偏态分布-峰度

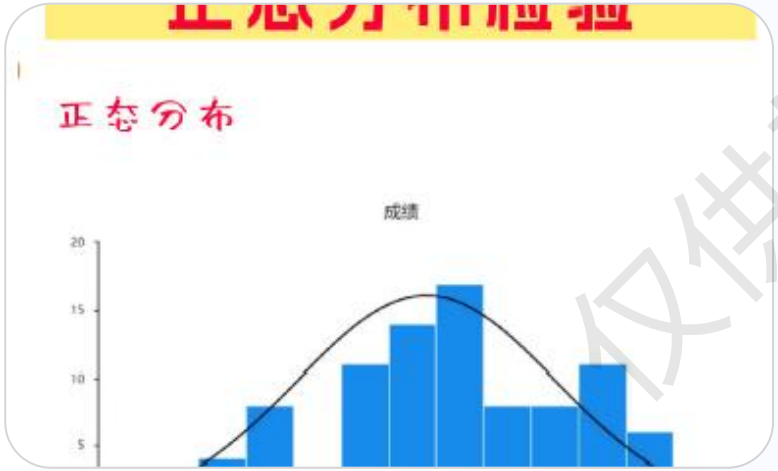


- 标准正态分布的峰度值为3，大于3是高尖，小于3是平矮。（为了方便计算，有时也标准化为0）
- 峰度>3，称为高峰度，极端值出现的频率相对较低，但一旦出现，其影响力较大。
- 峰度<3，称为低峰度。极端值出现的频率相对较高，虽常出现，但也分担了风险，减少了特殊事件的影响程度。
- 各有优缺点，前者优点是预测稳定，风险集中易被识别，但缺点是风险一旦发生，后果影响巨大。后者优点是决策稳定，风险分散，极端事件影响小，但缺点是风险发生频率多，管理难度大。

3.3 分布检验

EXCEL

从集中趋势与离散程度的6大指标来进行快速判断。

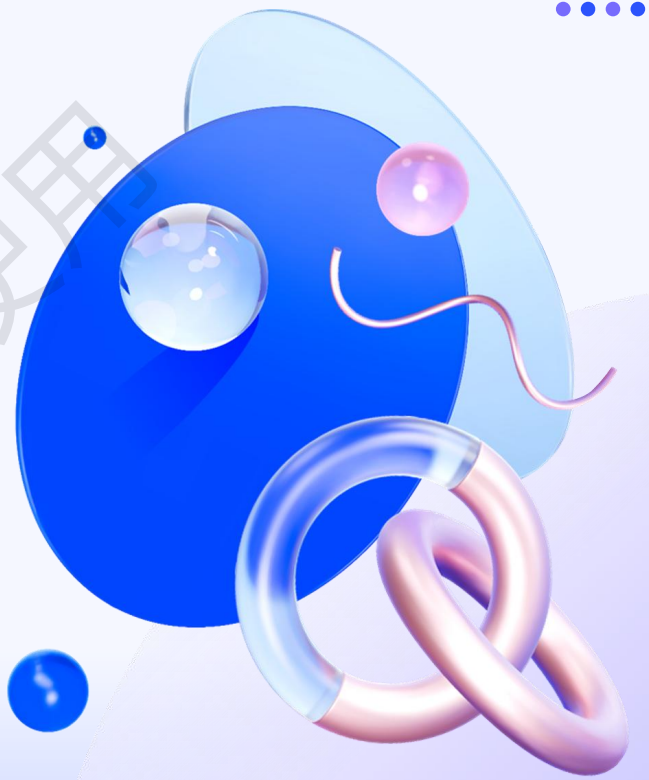


SPSS

分析功能中，描述统计的探索选项，可以直接进行正态图P-P的输出与检验。

PART4 极端值检验

- 了解极端值的含义
- 了解极端值的影响
- 掌握极端值检验的方法
- 形成极端值检验的数据分析技能



◆ 锻炼数据分析逻辑思维能力

4.1 极端值的含义

术语Outliers

那些与其他数据点相比较，显著偏离正常范围或分布模式的数值。

产生的原因

可能是自然的异常点，也可能是由于测量误差、数据输入错误或其他异常情况造成的。

特征

1.异常性：极端值是数据集中不寻常或异常的数据点，它们与大多数其他数据点不一致。

2.稀有性：极端值在数据集中出现的频率通常很低。

4.2 极端值的影响

统计影响

不能准确反映数据的集中趋势和离散程度，可能导致模型不准确。

数据处理

需要被识别并处理，以避免对分析结果的影响，这可能涉及删除、替换或转换极端值。

数据分析

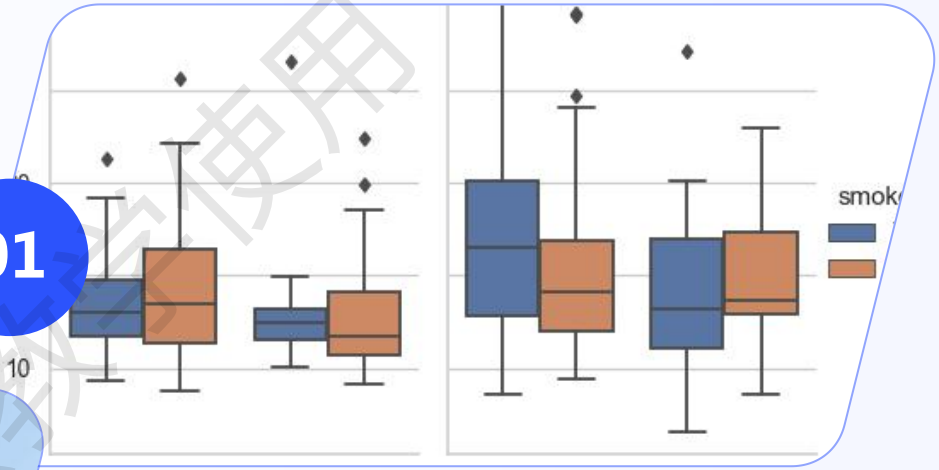
可能违反某些统计方法的基本假设。是数据预处理的一个重要步骤，可以帮助识别数据质量问题或潜在的兴趣点。

业务决策

可能会误导决策者，使他们基于异常情况而非数据的真实趋势做出决策。特别是金融与保险行业，因为它们可能代表极端事件的风险。

4.3 极端值检验方法

箱形图，针对单个数据集，运用EXCEL表格和SPSS软件都可以制作出箱形图，从而标注出极端值的情况。

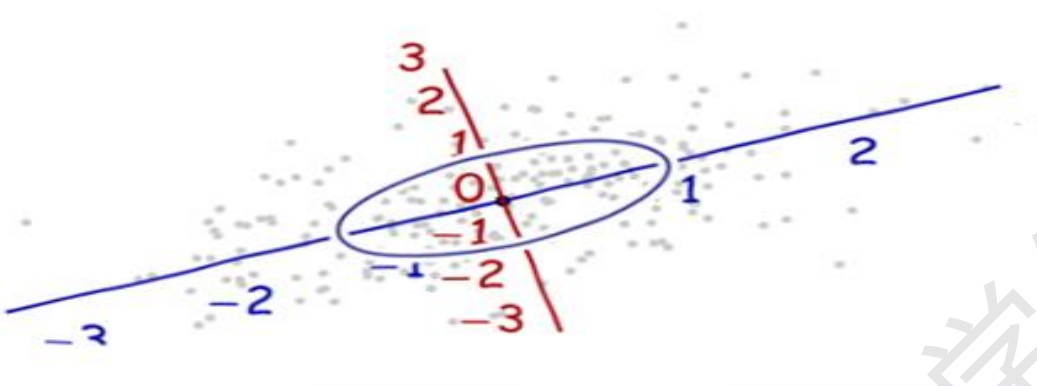


01

02

马哈拉诺比斯距离（Mahalanobis distance,记做d^2），针对多个数据集，运用AMOS软件得出d^2 的值，同时，比较该值和临界阈值之间的大小关系。如大于，说明对应的个案是多元异常值（Tabachnick & Fidell,2014）。

4.4 马哈拉诺比斯距离



- 案例在多元空间中围绕质心形成群落。
- 如果每个案例在群体中表示一个点的话，那么一个多元异常值的案例则会处于群体之外，并与其他案例保持明显的距离差异。
- 在P<0.001的卡方分布临界阈值下，运用公式 =CHISQ.INV(1-0.001,df) 可计算阈值。df是指问卷中题项数。

Observation number	Mahalanobis d-squared	p1	p2
596	73.265	0	0.067
126	66.524	0.001	0.078
610	66.428	0.001	0.012
516	62.046	0.002	0.065
604	61.665	0.003	0.026
192	61.44	0.003	0.009
263	61.252	0.003	0.003
95	61.145	0.003	0.001
204	60.388	0.004	0.001
164	59.603	0.004	0.001
311	58.639	0.005	0.001
22	58.53	0.006	0
10	55.93	0.01	0.019
189	55.193	0.012	0.031

PART5.1 实践练习

想一想

在描述统计中，判断数据集的分布特征十分重要。编写数据分析报告中，如果涉及到推断统计的知识来进行分析，那么正态性检验将是必不可少的步骤。请问我们可以采用哪些方法来进行正态性检验呢？

PART5.2 实践练习二

想一想

在描述统计中，判断数据集的分布特征十分重要。编写数据分析报告中，会涉及很多函数的使用，那么对于常见的函数，你是否熟悉它们的用法呢？

谢谢观看