描述统计分析

统计学及描述性统计

讲师：萨缪尔 Samuel

萨缪尔老师

网易数据分析教研负责人；

前盛大游戏战略规划总监、前腾讯游戏商业智能中心Leader

上海交通大学高金硕士 、《哈佛管理导师》外部导师

- **知乎大V：「萨缪尔」**主要聚集与商业分析、行业与战略分析、Python数据分析等
- **BAT**互联网巨头商业洞察分析 & 咨询公司战略咨询背景
- 擅长行业趋势研究和战略管理咨询工具，**为20多家上市公司提供战略发展决策建议**

课程亮点

1 统计学的两个分支

2 描述性统计1：集中趋势

3 描述性统计2：波动性

统计学的两个分支

Descriptive Statistics & Inferential Statistics

Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data. The mathematical theories behind statistics rely heavily on **differential and integral calculus, linear algebra, and probability theory**.

The two major areas of statistics are known as **descriptive statistics**, which describes the properties of sample and population data, and **inferential statistics**, which uses those properties to test hypotheses and draw conclusions.

What Are Descriptive Statistics?

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

KEY TAKEAWAYS

- Descriptive statistics summarizes or describes the characteristics of a data set.
- Descriptive statistics consists of two basic categories of measures: measures of central tendency and measures of variability (or spread).
- Measures of central tendency describe the center of a data set.
- Measures of variability or spread describe the dispersion of data within the set.

Inferential Statistics

Inferential statistics are tools that statisticians use to draw conclusions about the characteristics of a population, drawn from the characteristics of a sample, and to decide how certain they can be of the reliability of those conclusions. Based on the sample size and distribution statisticians can calculate the probability that statistics, which measure the central tendency, variability, distribution, and relationships between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which the sample is drawn.

Inferential statistics are used to make generalizations about large groups, such as estimating average demand for a product by surveying a sample of consumers' buying habits or to attempt to predict future events, such as projecting the future return of a security or asset class based on returns in a sample period.

描述性统计的主要内容

Descriptive Statistics

Measures of Central Tendency

集中趋势

- Mean
- Median
- Mode

Measures of Dispersion

波动性

- Range
- Standard deviation

示例：Excel计算描述性统计

步骤1：Excel-数据-数据分析-描述统计-确定

步骤2：逐项选择输入输出选项即可

得出结果

曝光数	
平均	12438.6175
标准误差	136.365594
中位数	12479.95
众数	11052.1
标准差	1493.81024
方差	2231469.02
峰度	-1.3764918
偏度	0.02414384
区域	4909.8
最小值	10100.2
最大值	15010
求和	1492634.1
观测数	120
最大(1)	15010
最小(1)	10100.2
置信度(95.0%	270.017496

描述性统计—集中趋势

平均数、中位数、众数

平均数

平均数用于衡量某个数据集的中心位置，它的缺点是易受异常值影响，因此，求平均数之前需要首先处理异常值。

中位数/四分位数

将数据集排序后分为两部分，处于正中间位置的数据就是中位数；将数据集划分为四部分，这种划分的临界点就是四分位数，包括下四分位数（25%）和上四分位数（75%）。

中位数和四分位数不易受极端值影响，因此，当有极端值存在时，中位数能比平均数更有效地反映出数据的中心位置。

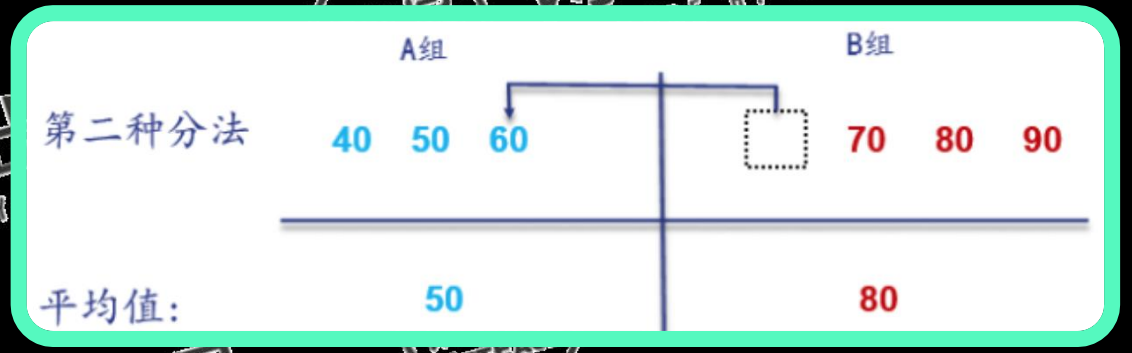
众数

众数是一组数据中出现最多的数，不易受到极端值影响。

分组与平均值

罗杰斯悖论
Will Rogers Phenomenon

Will Rogers悖论，是指将某些事物从一个组移到另一个组，两组的平均值增大，虽然其中没有值变大。之所以会出现这种情况，是因为当数据点从一个组重新归类到另一组的时候，如果这个点在原来组的平均线以下，但是在新组的平均线之上，那么这两个组的平均线都会提升。



描述性统计—波动趋势：方差（标准差）、离差、Z分值

方差或标准差 Variance or Standard deviation

方差和标准差所反映的是一组数据与其均值为代表的中心的平均离散水平。因为标准差的计算应用到每一个变量值，所以，会受到极端值的影响，当数据中有较明显的极端值（outlier)时不宜使用。必须知道这一点，所有方差/标准差分析的前提是：样本总体服从正态分布，如果不服从，就要有补救措施，比如数据转换。

z-score

z分数用于衡量数据项在数据集中的相对位置，用人类语言来说，它描述的是数据项距离均值有几个标准差。

对序列 $x_1, x_2, \ldots, x_n$ 进行变换：

$$y_i = \frac{x_i - \bar{x}}{s}，\text{这里} \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i，\quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

则新序列 $y_1, y_2, \ldots, y_n$ 的均值为 0，而方差为 1，且无量纲。

离差（min-max标准化）

极差就是数据集中最大值与最小值的差值，极差是衡量离散程度的统计两种最容易计算的，但它并不常用，因为它只由两个数据项决定，易受极端值影响。

对序列 $x_1, x_2, \ldots, x_n$ 进行变换：

$$y_i = \frac{x_i - \min\limits_{1 \le i \le n}\{x_j\}}{\max\limits_{1 \le i \le n}\{x_j\} - \min\limits_{1 \le i \le n}\{x_j\}}$$

则新序列 $y_1, y_2, \ldots, y_n \in [0,1]$ 且无量纲。一般的数据需要时都可以考虑先进行规范化处理。

其他

变异系数CV(Coefficient of Variance)

偏态系数(Skewness)

峰态系数(Kurtosis)

课程总结

1. 统计学的两个分支：描述性统计和推断性统计，并用Excel计算描述性统计
2. 描述性统计的集中趋势：平均数、中位数、众数
3. 描述性统计的波动性：方差（标准差）、离差、Z分值

谢谢观看