

## 描述算法

### 目录

缺失分析 .....	1
全局过滤 .....	2
相关指标 .....	2

描述分析用于计算数据的平均值、标准差、分位数等，并且提供‘全局过滤’即过滤出分析标题均无缺失数据后的描述分析结果，在 SPSSAU 中支持：

- ✓ 批量对‘标题’进行分析；
- ✓ 提供平均值、标准差、分位数等近 20 个指标；
- ✓ 提供‘缺失分析’，即探索分析项是否有缺失数据；
- ✓ 提供‘全局过滤’后的描述分析结果。



### 缺失分析

选中该参数后，SPSSAU 会提供分析项的缺失样本量情况，某分析项是否有数字恒定，以及‘全局过滤’后样本量信息等。

## 全局过滤

Title1	Title2	Title3	Title4
3	4	3	5
5	2	5	4
4	4	1	1
4	5		4
1	3	2	3
3	1	3	5
4		2	5
1	4	2	2
	5	3	1
5	2	2	3

比如分析 4 个标题，共有 10 行数据，但个别标题（Title1、Title2 和 Title3）有缺失数据，全局过滤后仅为 7 行数据。

## 相关指标

描述分析时相关指标计算公式说明如下：

样本标准差：

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

其中： $n$ 是样本数量， $\bar{x}$ 是样本的平均值

样本方差：

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

标准误：

$$SE = \frac{s}{\sqrt{n}}$$

平均值的 95% 置信区间：

$$CI = \bar{x} \pm t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}}$$

其中： $t_{\frac{\alpha}{2}, n-1}$ 是 t 分布的临界值。

四分位距：

$$IQR = Q_3 - Q_1$$

其中： $Q_1$ 是 25% 分位数， $Q_3$ 是 75% 分位数。

峰度 (Kurtosis)：

$$K = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^4 - 3$$

偏度 (Skewness):

$$G = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3$$

变异系数 (Coefficient of Variation, CV):

$$CV = \frac{s}{\bar{x}} \times 100\%$$

分位数:

假定一个有序数据集  $X = \{x_1, x_2, \dots, x_n\}$ , 其中  $x_1 \leq x_2 \leq \dots \leq x_n$ , 分位数  $Q_p$  的计算公式如下:

✓  $Q_p$  的位置可以通过以下公式确定:  $k = (n + 1) \times p$

其中:

$k$  是分位数的位置

$n$  是数据点的总数

$p$  是所需的分位数 (比如中位数  $Q_2$  对应  $p = 0.50$ )

✓ 分位数计算

当  $k$  为整数时:  $Q_p = x_k$

当  $k$  不是整数时:  $k = [k] + d$ , 其中  $[k]$  是  $k$  的整数部分,  $d$  是小数部分 (即  $d = k - [k]$ ), 则分位数的计算公式为:  $Q_p = x_{[k]} + d \times (x_{[k]+1} - x_{[k]})$ 。

提示:

SPSSAU 计算分位数时不包括边界 0 和 1