第四章 概率、分布与随机模拟

实验5 简单描述统计分析及R语言实现(1)

目 录

5.1 实验目的

- 1. 掌握使用R对数据作描述性统计分析的方法；
- 2. 掌握R语言中生成随机数及进行随机抽样模拟的方法；
- 3. 掌握蒙特卡洛模拟的方法。

5.2 实验原理

1.数据的集中趋势分析

数据的集中趋势分析是用来反映数据的一般水平，常用的指标有平均值、中位数和众数等。

设n个观测值为$x_1, x_2, \ldots, x_n$，其中n称为样本容量。

均值即是$x_1, x_2, \ldots, x_n$的平均数，表示数据的集中位置：$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

中位数是描述数据心位置的数字特征，计算公式：$$M = \begin{cases} x_{\frac{n+1}{2}}, n\text{为奇数} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\left(\frac{n}{2}+1\right)}\right), n\text{为偶数} \end{cases}$$

众数是指在数据中发生频率最高的数据值。

如果各个数据之间的差异程度较小，用平均值就有较好的代表性；而如果数据之间的差异程度较大，特别是有个别的极端值的情况，用中位数或众数有较好的代表性。

对于对称分布的数据，均值和中位数较接近；对于偏态分布的数据，均值和中位数不同。

2.数据的离散程度分析

数据的离散程度分析主要是用来反映数据之间的差异程度，常用的指标有方差和标准差。方差是标准差的平方，根据不同的数据类型有不同的计算方法。

方差是描述数据取值分散性的一个度量，它是数据相对于均值的偏差平方的平均

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

方差的开方称为标准差。方差的量纲与数据的量纲不一致，它是数据量纲的平方，而标准差的量纲与数据量纲一致，标准差为

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

方差或标准差的值越小，说明数据的离散程度越低。

关于数据的统计性描述还有诸如：变异系数（$C_v = \frac{\sigma}{\mu}$）、峰度（$C_k = \frac{\mu_4}{\mu_2^2} - 3$）、偏度（$C_s = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$）等等。

3.数据的分布

在统计分析中，通常要假设样本的分布属于正态分布，因此需要用偏度和峰度两个指标来检查样本是否符合正态分布。偏度衡量的是样本分布的偏斜方向和程度；而峰度衡量的是样本分布曲线的尖峰程度。一般情况下，如果样本的偏度接近于0，而峰度接近于3，就可以判断总体的分布接近于正态分布。

5.3 实验过程

1.样本均值

在R中用mean()函数计算样本均值，其调用格式为：

mean(x,trim = 0,na.rm = F)

- x ：表示要计算均值的对象；
- trim ：表示介于0到0.5之间的数(默认值是0)，表示在计算均值之前，去掉两端数据的百分比；
- na.rm ：为逻辑值，当取T时，允许样本中有缺失值。

```
x <- c(1:10,50)
xm <- mean(x)
c(xm,mean(x,trim = 0.1))
```

2.样本方差

在R语言中，用var()计算样本方差，其调用格式为：

var(x,y = NULL,na.rm = F)

- x ：数值向量、矩阵或数据框；
- y ：为NULL(默认值)，此时计算样本方差；y为数值向量、矩阵或数据框时，计算样本协方差；
- na.rm ：逻辑变量，当取值为T时，可处理缺失数据。

另外，cov()和corr()分别可计算样本协方差矩阵和相关系数矩阵，其使用方法与var()一样。样本方差的开方称为样本均方差，在R中，sd()函数计算样本均方差，其调用格式为sd(x,na.rm = F)

3.顺序统计量

将n个数据(观测值)按从小到大的顺序排列后,称其为顺序统计量。

在R中用sort(x)计算样本x的顺序统计量；order()给出排序后的下标；

rank()给出了样本x的秩次统计量。它们的调用格式及参数说明如下。

sort(x, decreasing = F, na.last = NA, method = c("auto", "shell", "quick", "radix"),...)

- x ：数值向量、矩阵或数据框；
- y ：为NULL(默认值)，此时计算样本方差；y为数值向量、矩阵或数据框时，计算样本协方差；
- na.rm ：逻辑变量，当取值为T时，可处理缺失数据。

- x ：数值向量；
- decreasing ：逻辑变量，取值为T，返回值为降序排列；取值为F（默认）时，返回值为升序排列；
- na.last：控制缺失数据的参数，当取值为NA（默认）时，不处理缺失值；当取值为T时，缺失数据排在最后；当取值为F时，缺失数据排在最前面；
- method ：指定排序所用的算法，当取值为"auto"时，表示使用的是"radix"即基数排序算法；当取值为"shell"时，表示使用的是希尔排序算法；当取值为"quick"时表示使用的是快速排序算法；
- … ：附加参数。

order(x, na.last = TRUE, decreasing = FALSE, method = c("auto", "shell", "radix"))

其中各参数的使用方法与sort()一样。

rank(x, na.last = TRUE, ties.method = c("average", "first", "last", "random", "max", "min"))

其中参数x，na.last的含义与sort()一样，ties.method是x中存在重复分量时，秩的确定方法；当取值为"average"时取下标的平均值；当取值为"first"时，取该值出现的所有下标按升序排列；当取值为"last"时，取该值出现的所有下标按降序排列；当取值为"random"时，随机取一个下标；当取值为"max"时，取该值对应的下标中的最大者；当取值为"min"时，取该值对应的下标中的最小者。

《第四章 概率、分布与随机模拟》

```
x <- c(75,64,47.4,66.9,62.2,62.2,58.7,63.5)
sort(x)
# 注意order()与rank()
order(x)
rank(x)
```

```
> sort(x)
[1] 47.4 58.7 62.2 62.2 63.5 64.0 66.9 75.0
> order(x)
[1] 3 7 5 6 8 2 4 1
> rank(x)
[1] 8.0 6.0 1.0 7.0 3.5 3.5 2.0 5.0
```

```
# 比较rank()的不同ties.method的区别
x <- c(1,1,2,2,2,3,3,3,3)
rank(x,ties.method = "average")
rank(x,ties.method = "max")
rank(x,ties.method = "last")
```

```
> rank(x,ties.method = "average")
[1] 1.5 1.5 4.0 4.0 4.0 7.5 7.5 7.5 7.5
> rank(x,ties.method = "max")
[1] 2 2 5 5 5 9 9 9 9
> rank(x,ties.method = "last")
[1] 2 1 5 4 3 9 8 7 6
```

4.中位数

函数median()给观测量的中位数。其调用格式为

median(x,na.rm = F)

- x ：表示是要计算中位数的数据对象；
- na.rm ：逻辑值，当取T时，允许样本中有缺失值。

```
x <- c(75,64,47.4,66.9,62.2,62.2,58.7,63.5)
median(x)
```