统计学导论：R语言实验09

总体数值特征的估计

主讲人：郑盼盼

Outline

1. 总体均值的估计
2. 总体分位数的估计
3. 总体众数的估计
4. 总体方差的估计
5. 总体标准差、标准得分和变异系数的估计
6. 盒型图的绘制

9.1 总体均值的估计

9.1 总体均值的估计

根据大数定律，应该使用

$$\bar{X} \triangleq \frac{1}{n}\sum_{k=1}^{n} X_k$$

来估计总体均值，并称 $\bar{X}$ 为**样本均值**，或**均值**，或**平均数**。当样本是重复观测时，随着样本容量的增加样本均值“收敛于”总体均值。我们可以使用 `mean()` 计算样本均值。

9.1 总体均值的估计

若 $X \sim N(10, 100)$，对其进行重复观测 1000 次，得到样本 $X_i, (i = 1, 2, \ldots, 1000)$，并通过该样本估计总体的均值：

```r
n <- 1000 # 定义样本量
x <- rnorm(n, 10, 10) # 生成观测值
estim_mean <- mean(x) # 计算均样本值
true_mean <- 10 # 真实均值
cat("估计均值为 ", estim_mean,
    " 与总体均值的差异为", abs(true_mean-estim_mean))
```

9.2 总体分位数估计

9.2 总体分位数的估计

α 分位数描述了随机变量概率分布的位置信息，随机变量落在它两边的概率分别接近于 α 和 $1-\alpha$。根据大数定律，应该用满足如下条件的 $\hat{x}_\alpha$ 提取信息

$$\frac{\text{小于}\hat{x}_\alpha\text{的数据个数}}{n} \le \alpha, \ \frac{\text{大于}\hat{x}_\alpha\text{的数据个数}}{n} \le 1-\alpha$$

定义 4.4.8 对于任意 $\alpha \in (0,1)$，记 k 为数 $\alpha n + 0.5$ 的整数部分，并约定 $X_{(0)}$ 等于 $X_{(1)}$，$X_{(n+1)}$ 等于 $X_{(n)}$，称

$$\hat{x}_\alpha = (\alpha n + 0.5 - k)(X_{(k+1)} - X_{(k)}) + X_{(k)}$$

为样本 $X_1, X_2, \ldots, X_n$ 的**样本α分位数**，简称为α **分位数**。

9.2 总体分位数的估计

- `quantile(x, probs,type)` 来计算样本 x 的 `probs` 分位数，`type` 用于指定计算百分位数的方法，例如 `type=4` 对应于上面的计算方法。
- 特别的，对于中位数（即 0.5 分位数）可以使用 `median()` 函数

```r
x <- runif(1000, 0, 2)
quantile(x, c(0.25, 0.5, 0.75))

median(x)
```

9.3 总体众数估计

9.3 总体众数的估计

1. 对于**离散型随机变量**
    1. 使用 `table` 统计各变量出现的频数
    2. 使用 `max.which()` 找出频数最大的那一列
    3. 通过 `names` 返回频数最大那一列对应的数（即众数

9.3 总体众数的估计

1. 对于**离散型随机变量**
   1. 使用 table 统计各变量出现的频数
   2. 使用 max.which() 找出频数最大的那一列
   3. 通过 names 返回频数最大那一列对应的数（即众数

```
x <- rpois(1000, lambda=1)
tmpFreq <- table(x)
tmpId <- which.max(tmpFreq)
names(tmpId)
plot(0:10, dpois(0:10,1), type="h")
```

9.3 总体众数的估计

2. 对于**连续型随机变量**
    1. 使用 `hist` 对数据进行分组，进行频率统计
    2. 使用 `max.which($density)` 找出密度最大的那组
    3. 通过该组的间隔确定该组的中心点，(即众数

9.3 总体众数的估计

2. 对于**连续型随机变量**
   1. 使用 hist 对数据进行分组，进行频率统计
   2. 使用 max.which($density) 找出密度最大的那组
   3. 通过该组的间隔确定该组的中心点，（即众数

```
x <- rnorm(1000, 1, 3)
tmp <- hist(x, plot=F)  # 计算x的分组数据统计结果
tmpId <- which.max(tmp$density) # 计算最大密度所在的区间序
号
tmpInterval <- tmp$breaks[tmpId:(tmpId+1)] # 获得第
tmpId区间的矩形左右横坐标
mean(tmpInterval) # 平均后得到第tmpId区间的中心
```

9.4 总体方差估计

9.4 总体方差的估计

根据强大数定律，应该使用

$$\frac{1}{n}\sum_{k=1}^{n}(X_k-\mathbb{E}(X))^2$$

来近似方差，但实际上，我们无法得到具体的总体期望 $\mathbb{E}(X)$，只能通过样本均值 $\bar{X}$ 来估计总体期望，称

$$S^2=\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\bar{X})^2$$

为**样本方差**。注意此处除以的是 $n-1$，这是为了保证 $\mathbb{E}(S^2)=D(X_1)$，即样本方差是总体方差的无偏估计。

9.4 总体方差的估计

在R语言中，我们可用 `var(x)` (Variance) 来计算样本数据 x 的方差：

```r
n <- 100
x <- rnorm(n, 1, 3)
var(x)
estim_1 <- 1/n * sum((x - mean(x)) ^ 2) # 除以n
estim_1
estim_2 <- 1/(n-1) * sum((x - mean(x)) ^ 2) # 除以(n-1)
estim_2
```

9.5 总体标准差、标准得分和变异系数的估计

9.5 总体标准差、标准得分和变异系数的估计

样本标准差 样本方差的平方根:

$$S = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2}$$

在R语言中，可以直接使用 sd(x) (**s**tandard **d**eviation) 计算数据向量 x 的样本标准差。

```
x <- rnorm(100, 1,3)
sd(x)
sqrt(var(x))
```

9.5 总体标准差、标准得分和变异系数的估计

标准得分的估计 假设 $X_1, X_2, ..., X_n$ 为样本，对于 $1 \leq i \leq n$，称

$$Z_i = \frac{X_i - \bar{X}}{S}$$

为第 i 个样本数据的**标准化**或**标准得分**：称$Z_1, Z_2, ..., Z_n$为相应的**标准化样本**。R语言中可以直接使用 scale 完成分数的标准化。

```
x<-c(79.0,87.0,75.0,94.0,69.0, 81.0,64.0,62.0,71.0,70.0,
     64.0,70.0,69.0,67.0,61.0,71.0,72.0,71.0, 88.0,83.0)
y<-c(86.0,78.0,73.0,90.0,42.0,75.0,71.0,87.0,71.0,41.0,
     75.0,78.0,68.0,73.0,61.0,60.0,70.0,47.0,74.0,70.0)
xy <- matrix(c(x,y), 20,2)
xyScale <- scale(xy)
xyScale[3,]
```

9.5 总体标准差、标准得分和变异系数的估计

变异系数的估计 样本标准差于样本均值的绝对值之比。

$$\frac{S}{|\bar{X}|}$$

```r
tmpX <- rpois(1000, 3)
sd(tmpX) / abs(mean(tmpX))
```

9.6 盒型图的绘制

9.6 盒型图的绘制

五数概括 对于已经获取的样本数据 $x_1, x_2, ..., x_n$ 而言，其四分位数，最小样本值 $x_{(1)}$ 和最大样本值 $x_{(n)}$ 概括了样本中的大部分信息，因此，人们称 $x_1, Q_1, Q_2, Q_3, x_{(n)}$ 为**样本数据的五数概括**，简称**五数概括**，在R语言中可以通过函数`fivenum()`计算得到

```
fivenum(x)
```

9.6 盒型图的绘制

通过盒型图，可以展示样本观测数据中的离群数据，推断总体密度图像的整体特征（对称，U型，左倾或右倾）。

在**改良盒型图**中，下（左）虚线下（左）端的坐标为

$$a = \max\{x_{(1)}, Q_1 - kQ_d\}$$

上（右）虚线上（右）端的坐标为

$$b = \min\{x_{(n)}, Q_3 + kQ_d\}$$

其中 $Q_d = \frac{Q_3 - Q_1}{2}$ 为**四分位距**，一般设置 $k = 1.5$；

9.6 盒型图的绘制

在R语言中，我们可以用 boxplot(x， range=) 函数绘制数据向量 x 的盒型图，其中 range 对应于上面的 k

```r
x1 <- c(42,55,64,70,75,78,80,82,82,82,85,85,85,85,88,90,90,92,95,99)
x2 <- c(39,52,61,68,72,76,77,78,79,78,83,83,81,81,85,87,86,91,91,98)
xy <- data.frame("A class" = x1, "B class" = x2)
boxplot(xy, range=1.5, ylab="score")
```