第二章　单总体参数区间估计

2.1 总体比例的区间估计

我们可以将比例问题比作为一项满足二项分布的试验。，$\hat{p}$为样本比例，n 为样本大小。

所以我们可以得到下面的式子

$$E(\hat{p}) = \frac{1}{n}E(X) = \frac{1}{n} \times np = p$$

并且还得出

$$\operatorname{var}(\hat{p}) = \frac{1}{n^2}\operatorname{var}(X) = \frac{1}{n^2} \times np(1-p) = \frac{p(1-P)}{n}, \quad se(\hat{p}) = \sqrt{\frac{p(1-P)}{n}}$$

由此便已经具备了进行区间估计的必备素材。我们最常用的方法被称为是 Wald 方法。Wald 方法是一种统计学上的假设检验方法和参数估计方法，旨在通过对样本数据进行分析来推断总体参数的值。该方法主要针对二项分布或正态分布中均值或比例的估计。根据中央极限定理（它说明当样本量足够大时，对于任何概率分布的独立同分布随机变量序列，其样本均值的分布会趋近于一个正态分布），当样本 n 足够大时，将会有

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-P)}{n}}\right)$$

我们以历史著名的索尔克随机双盲对照试验为例，我们可以在 rstudio 中使用下面的代码来计算置信区间。输出的置信区间为（0.000 2102390，0.000 3576456）

```
> n<-200745
> (p.hat<-57/n) %样本中发生该事件的比例
[1] 0.0002839423
> p.hat+c(-1.96,1.96)*sqrt(p.hat*(1-p.hat)/n)%计算置信区间
[1] 0.0002102390 0.0003576456
```

我们下面用的 Clopper-Pearson 方法则与 wald 方法截然不同。该方法完全是基于二项分布的，该方法得出的区间一般更加准确。在 rstudio 中可以用 binom.test（）函数来执行 Clopper-Pearson 方法，下面是我们给出的代码：

```
> binom.test(57,200745)

        Exact binomial test


data:  57 and 200745
number of successes = 57, number of trials = 200745, p-value <
2.2e-16
alternative hypothesis: true probability of success is not equal to 0.
5
95 percent confidence interval:
 0.0002150620 0.0003678648
sample estimates:
probability of success
          0.0002839423
```

从上面代码输出的结果显示95%置信水平下之区间估计结果为（0.000 215，0.000 369）。这个数值与wald方法得出的结果相当接近，说明在规定的误差下，两个方法都是可行的

2.2 总体均值的区间估计

对总体均值进行区间估计时，需要分为以下两种情况来讨论：

1. 已知总体标准差：在这种情况下，可以使用正态分布的性质来进行区间估计。假设样本大小为n，样本的平均值为x̄，且总体标准差为σ。则总体均值μ的置信区间为

$$\left(\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$Z=\frac{x-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

其中，zα/2是标准正态分布的分位数，通常取为1.96，表示置信水平为95%

2. 未知总体标准差：在这种情况下，需要使用样本标准差 s 来估计总体标准差σ。此时，可以使用 t 分布的性质来进行区间估计。假设样本大小为 n，样本的平均值为 x̄，且样本标准差为 s。则总体均值μ的置信区间为

$$\left(\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}},\bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

其中，tα/2, n-1 是自由度为 n-1 的 t 分布的分位数，通常根据置信水平及样本大小从 t 分布表中查询得到。

需要注意的是，当样本大小足够大时（通常大于 30），t 分布逐渐接近于标准正态分布，因此可以近似使用标准正态分布的性质。

我们下面来举一个例子

例如现在有一家生产零件的机械厂。按规定每个零件规定的重量应该为 100g。为对零件质量进行监测，质检部门从当天生产的一批零件中中随机抽取了 25 个，并测得每个零件的重量数据如下表（表 1）所示。已知零件的重量的分布服从正态分布，且总体标准差为 10g。要我们计算出这天零件平均重量的置信区间，置信水平为 95%。

表 1 零件的重量（g）

数据	112.5	101.0	102.0	100.5
	102.6	107.5	108.8	115.6
	100.0	123.5	101.6	102.2
	116.6	95.4	108.6	105.0
	136.8	102.8	98.4	93.3

在 R 语言里面，暂时还没有函数能够直接计算已知方差情况下的置信区间，所以我们自己编一个函数用来计算置信区间，下面是代码

```
> conf.int<-function(x,n,sigma,alpha){options(digits=5)
+     mean<-mean(x)
+     c(mean-sigma*qnorm(1-alpha/2,mean=0,sd=1,
+                        lower.tail=TRUE)/sqrt(n),
```

```
+       mean+sigma*qnorm(1-alpha/2,mean=0,sd=1,
+                           lower.tail=TRUE)/sqrt(n))}
```

这个函数用于计算一个总体均值的置信区间。x 是样本数据，n 是样本大小，sigma 是总体标准差（已知或通过样本估计得到），alpha 是置信水平。函数使用正态分布的分位数来计算置信区间的上下界。返回的值是一个包含下限和上限的向量。

然后调用上述函数来计算置信区间，代码如下

```
> x <- c(112.5, 101.0, 103.0, 102.0, 100.5,
+           102.6, 107.5, 95.00, 108.8, 115.6,
+           100.0, 123.5, 102.0, 101.6, 102.2,
+           116.6, 95.40, 97.80, 108.6, 105.0,
+           136.8, 102.8, 101.5, 98.40, 93.30)
> n <- 25
> alpha <- 0.05
> sigma <- 10
> result <- conf.int(x, n, sigma, alpha)
> result
[1] 101.44 109.28
```

由代码输出的结果表明这批零件的平均重量 95%的置信区间是（101.44，109.28）

下表（表 2）总结了本小节中关于单总体均值的区间估计方法

表 2

总体分布	样本量	总体方差 σ^2 已知	总体方差 σ^2 未知
正态分布	大样本($n \geqslant 30$)	$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$
	小样本($n < 30$)	$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$
非正态分布	大样本($n \geqslant 30$)	$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$

2.3 总体方差区间估计

下面我们将讨论正态总体方差的区间估计问题。从前面的方差抽样分布可以得出样本方差服从自由度为 n-1 的卡方分布。我们就要考虑到用卡方分布构造总体方差的置信区间。我们现在要找到一个χ2 值，使其满足下列式子

$$\chi^2_{1-\alpha/2} \leqslant \chi^2 \leqslant \chi^2_{\alpha/2}$$

因为

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

我们用上面的式子来置换掉第一个式子的χ2，然后得到

$$\chi^2_{1-\alpha/2} \leqslant \frac{(n-1)s^2}{\sigma^2} \leqslant \chi^2_{\alpha/2}$$

最后就可以推导总体方差σ2 在 1-α 置信水平下的置信区间为

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

因为在 rstudio 中没有上述方法函数的扩展包，所以我们下面将编写一个程序用作方差区间估计的函数

```
>chisq.var.test<-function(x,alpha){
options(digits=4)
result<-list()
n<-length(x)
v<-var(x) %用 var()函数计算 x 的样本方差 v；
result$conf.int.var<-c(
(n-1)*v/qchisq(alpha/2,df=n-1,lower.tail=F),
(n-1)*v/qchisq(alpha/2,df=n-1,lower.tail=T)) %使用 qchisq()函数计算自由度为 n-1 的卡方分布上 alpha/2 分位数和 1-alpha/2 分位数；
result$conf.int.se<-sqrt(result$conf.int.var)
result}
```