

第四章 双总体比例差

双总体比例差是指两个总体中两种现象的比例之间的差异。如果从满足两种分布的两个整体中提取两个独立的样本，则两个样本比例差异的样本为正态分布，即

(p-hat1 - p-hat2) ~ N(p1 - p2, sqrt(p1(1-p1)/n1 + p2(1-p2)/n2))

我们再次将其标准化，然后我们得到

z = (p-hat1 - p-hat2 - (p1 - p2)) / sqrt(p1(1-p1)/n1 + p2(1-p2)/n2) ~ N(0,1)

当两个总体的比例 p1 和 p2 未知时，当两个完整比例 p1 和 p2 未知时，可以用示例比例和替换。因此，根据正态分布建立的两个总体比例的差异在于 p1-p2 在 1-α 置信度的置信区间是

(p-hat1 - p-hat2) ± z_{α/2} * sqrt(p-hat1(1-p-hat1)/n1 + p-hat2(1-p-hat2)/n2)

首先，对某节目收视率的调查如下表（表 5）所示

表 5 节目收视率

	城市	农村
总人数	500	400
观看该节目的人数	225	128

我们要以 95%的置信度计算城乡收视率差距的置信区间。

这时可以使用 R 语言的 prop.test() 函数对两个总比率差进行区间估计，如下图所示。

```
> prop.test(x=c(225,128),n=c(500,400),correct=F) %将 correct 值置为 F,
说明
没有
使用
连续
```

性修

正

```
2-sample test for equality of proportions without continuity
correction

data:  c(225, 128) out of c(500, 400)
X-squared = 15.754, df = 1, p-value = 7.213e-05
alternative hypothesis: two.sided
95 percent confidence interval:
 0.06682346 0.19317654
sample estimates:
prop 1 prop 2
 0.45   0.32
```

根据输出结果，预计置信区间为（6.68%，19.32%），即 95%的置信区间下城乡收视率差为 6.68%~19.32%。

如果我们不改变 correct 的默认值，计算结果，则会得出下列代码

```
> prop.test(x=c(225,128),n=c(500,400))
2-sample test for equality of proportions with continuity
correction

data:  c(225, 128) out of c(500, 400)
X-squared = 15.214, df = 1, p-value = 9.601e-05
alternative hypothesis: two.sided
95 percent confidence interval:
 0.06457346 0.19542654
sample estimates:
prop 1 prop 2
```

0.45 0.32