5 非参数检验方法

5.1 非参数方法的基本概念

5.2 成组样本比较的非参数方法

5.2.1 两样本比较

scipy.stats.median_test()

中位数检验，两组或者多组时均可使用

scipy.stats.ranksums(a, b)

wilcox秩和检验，相对使用较少

scipy.stats.mannwhitneyu(a, b, use_continuity, alternative)

Mann-Whitney U检验，实际使用中一般直接代替wilcox秩和检验

scipy.stats.ks_2samp(data1, data2)

两样本KS检验

例5.1

在分析中，已经确认了已婚和未婚人群的总信心指数均值存在差异，现需要进一步分析究竟是哪些构成指标导致了总信心指数出现差异。

构成指标只有五种分值，因此按照有序分类变量加以分析更为稳妥。

In [ ]:

```
print(ss.mannwhitneyu(ccss.Qa3[ccss.s7 == '未婚'], ccss.Qa3[ccss.s7 == '已婚']))
print(ss.mannwhitneyu(ccss.Qa4[ccss.s7 == '未婚'], ccss.Qa4[ccss.s7 == '已婚']))
print(ss.mannwhitneyu(ccss.Qa8[ccss.s7 == '未婚'], ccss.Qa8[ccss.s7 == '已婚']))
print(ss.mannwhitneyu(ccss.Qa10[ccss.s7 == '未婚'], ccss.Qa10[ccss.s7 == '已婚']))
print(ss.mannwhitneyu(ccss.Qa16[ccss.s7 == '未婚'], ccss.Qa16[ccss.s7 == '已婚']))
```

In [ ]:

```
ss.ranksums(ccss.Qa3[ccss.s7 == '未婚'], ccss.Qa3[ccss.s7 == '已婚'])
```

In [ ]:

```
ss.median_test(ccss.Qa3[ccss.s7 == '未婚'], ccss.Qa3[ccss.s7 == '已婚'])
```

In [ ]:

```
ss.ks_2samp(ccss.Qa3[ccss.s7 == '未婚'], ccss.Qa3[ccss.s7 == '已婚'])
```

5.2.2 多样本比较

scipy.stats.kruskal(sample1, sample2, ... , nan_policy = 'propagate')

nan_policy : {'propagate', 'raise', 'omit'}
多样本的Kruskal-Wallis H检验

例5.2

前面的分析中已经发现北京消费者的总信心值在不同时点有差异，现需要进一步分析究竟是哪些构成指标导致了总信心指数出现差异。

In [ ]:

```
ss.kruskal(ccss.query("s0 == '北京' & time == '200704'").Qa3,
           ccss.query("s0 == '北京' & time == '200712'").Qa3,
           ccss.query("s0 == '北京' & time == '200812'").Qa3,
           ccss.query("s0 == '北京' & time == '200912'").Qa3
          )
```

In [ ]:

```
ss.kruskal(ccss.query("s0 == '北京' & time == '200704'").Qa4,
           ccss.query("s0 == '北京' & time == '200712'").Qa4,
           ccss.query("s0 == '北京' & time == '200812'").Qa4,
           ccss.query("s0 == '北京' & time == '200912'").Qa4
          )
```

In [ ]:

```
print(ss.mannwhitneyu(ccss.query("s0 == '北京' & time == '200704'").Qa3,
                      ccss.query("s0 == '北京' & time == '200712'").Qa3))
print(ss.mannwhitneyu(ccss.query("s0 == '北京' & time == '200704'").Qa3,
                      ccss.query("s0 == '北京' & time == '200812'").Qa3))
print(ss.mannwhitneyu(ccss.query("s0 == '北京' & time == '200704'").Qa3,
                      ccss.query("s0 == '北京' & time == '200912'").Qa3))
```

5.3 配对/配伍样本比较的非参数方法

5.3.1 配对样本

scipy.stats.wilcoxon(a, b, zero_method='wilcox', correction=False)

两配对样本的Wilcox符号秩检验，实际工作中很少用到
zero_method : {'pratt', 'wilcox', 'zsplit'}
检验中包括0差值、丢弃0差值、将0差值对半分入两组

例5.3

前面的分析中发现受访者的预期指数在2007年12月有下降，那么在构成预期指数的三个指标中，究竟是哪些指标出现了下降呢？

In [ ]:

```
ccss_p = pd.read_excel("CCSS_sample.xlsx", sheet_name = 'CCSS_pair')
ccss_p.head()
```

In [ ]:

```
ss.wilcoxon(ccss_p.Qa4, ccss_p.Qa4n)
```

In [ ]:

```
ss.wilcoxon(ccss_p.Qa8, ccss_p.Qa8n)
```

In [ ]:

```
ss.wilcoxon(ccss_p.Qa10, ccss_p.Qa10n)
```

In [ ]:

```
ccss_p.describe()
```

5.3.2 配伍样本

scipy.stats.friedmanchisquare(measurements1, measurements2, measurements3...)

friedman卡方检验，至少需要提供三组数据

In [ ]:

```
ss.friedmanchisquare(ccss.query("s0 == '北京' & time == '200704'").Qa4[:10],
           ccss.query("s0 == '北京' & time == '200712'").Qa4[:10],
           ccss.query("s0 == '北京' & time == '200812'").Qa4[:10],
           ccss.query("s0 == '北京' & time == '200912'").Qa4[:10]
          ) # 此处仅为结果演示
```

5.4 秩变换分析

scipy.stats.rankdata(

a : 需要编秩的数值类数组结构。
method = 'average' : 对结的处理方式。
'average': 取对应秩次的平均值。
'min'/'max' : 取对应秩次的最小/最大值。
'dense': 所有相同的数值只赋予一个秩次，随后继续流水编号。
'ordinal': 按照数值出现的顺序依次赋予不同的秩次。

)

例5.4

利用秩变换分析方法完成例5.2（哪些指标的变化导致了北京消费者的总信心值在不同时点有差异）中的多组比较和事后两两比较操作。

In [ ]:

```
# 取出所需数据，便于后续操作
dfrank = ccss.loc[ccss.s0 == '北京', ['time', 'Qa3']]
dfrank.head()
```

In [ ]:

```
dfrank['qa3r'] = ss.rankdata(dfrank.Qa3)
dfrank.head()
```

In [ ]:

```
ss.f_oneway(dfrank[dfrank.time == 200704].qa3r,
            dfrank[dfrank.time == 200712].qa3r,
            dfrank[dfrank.time == 200812].qa3r,
            dfrank[dfrank.time == 200912].qa3r
            )
```

In [ ]:

```
# 尝试使用其他编秩方法
dfrank['qa3r2'] = ss.rankdata(dfrank.Qa3, method = 'dense')
ss.f_oneway(dfrank[dfrank.time == 200704].qa3r2,
            dfrank[dfrank.time == 200712].qa3r2,
            dfrank[dfrank.time == 200812].qa3r2,
            dfrank[dfrank.time == 200912].qa3r2
            )
```

In [ ]:

```
# 进行两两比较
import scikit_posthocs as sp

sp.posthoc_conover(dfrank, val_col='qa3r', group_col='time',
                   p_adjust = 'bonferroni')
```

5.5 实战练习

对CCSS数据中不同城市受访者的年龄进行比较，尝试使用变量变换、秩和检验、秩变换分析等方法完成该任务，比较相应的分析结果，并思考各种方法的优缺点。

6 卡方检验

6.1 卡方检验的基本原理

6.2 行*列表的卡方检验

例6.1

在CCSS的分析报告中，所有受访家庭会按照家庭年收入被分为低收入家庭和中高收入家庭两类，现希望考察不同收入级别的家庭其轿车拥有率是否相同。