In [ ]:

```
sp.binom_test(1, 600)
```

In [ ]:

```
sp.proportions_chisquare([5, 3, 2], 10)
```

6.6.3 近似Z检验

statsmodels.stats.proportion.proportions_ztest(

```
    count : 成功次数，单一数值/类数组结构列表
    nobs : 总样本量，单一数值/类数组结构列表
    value = None : H0所对应的总体率/率差
    alternative = 'two-sided'
    prop_var = False : 指定方差分配比例，默认按照样本比例进行计算
```

) # 输出：Z统计量、P值

In [ ]:

```
sp.proportions_ztest(30, 100, 0.2)
```

In [ ]:

```
sp.proportions_ztest([30, 65], [100, 200], 0)
```

6.7 实战练习

计算北京、上海、广州三地的汽车拥有率可信区间。

考察不同收入级别的受访者其职业分布有无差异。 提示：需要考虑两两比较。

在上面分析的基础上，在控制城市的影响之后，考察不同收入级别的受访者其职业分布有无差异。

7 相关分析

7.1 相关分析的指标体系

7.2 相关分析的实现

相关分析作为比较简单的方法，在statsmodels中并未作进一步的完善，因此主要使用scipy实现

两个连续变量，且符合双变量正态分布：Pearson相关系数

```
scipy.stats.pearsonr(a, b)
```

两个连续变量，不符合双变量正态分布：Spearman等级相关系数

```
scipy.stats.spearmanr(a, b)
```

两分类变量 vs. 连续变量：Point-biserial相关系数

```
scipy.stats.pointbiserialr(a, b)
```

两个有序变量：Kendall's Tau

```
scipy.stats.kendalltau(a, b, initial_lexsort=None, nan_policy='omit')
```

考察年龄和总信心指数间的关系

In [ ]:

```
ccss.plot.scatter('s3', 'index1')
```

In [ ]:

```
ccss.groupby('s3').index1.mean().plot()
```

In [ ]:

```
ss.pearsonr(ccss.s3, ccss.index1)
```

In [ ]:

```
ss.spearmanr(ccss.s3, ccss.index1)
```

考察当前家庭经济状况与一年后家庭经济状况感受值之间的关联

In [ ]:

```
pd.crosstab(ccss.Qa3, ccss.Qa4)
```

In [ ]:

```
ss.kendalltau(ccss.Qa3, ccss.Qa4)
```

In [ ]:

```
ss.spearmanr(ccss.Qa3, ccss.Qa4)
```

7.3 相对危险度与优势比

7.3.1 OR和RR的基本概念

7.3.2 scipy的实现方式

scipy.stats.fisher_exact()中可以计算OR值，相应的检验P值则是确切概率法的P值

In [ ]:

```
OR, P = ss.fisher_exact(pd.crosstab(ccss.Ts9, ccss.O1))
OR
```

7.3.3 statsmodels的实现方式

statsmodels.stats.contingency_tables.Table类可以直接提供分块2X2表OR的估计值

statsmodels.stats.contingency_tables.Table2x2类可以直接提供OR、RR的估计和检验结果

class statsmodels.stats.contingency_tables.Table2x2(

```
    table
    shift_zeros = True
```

)

Table2x2类的属性

```
    log_oddsratio / log_oddsratio_se    lnOR / lnOR的标准误
    oddsratio    OR值
    riskratio    RR值
    log_riskratio / log_riskratio_se    lnRR / lnRR的标准误
```

Table2x2类的方法

```
    * 注意：有些方法尚未开发完成

    summary([alpha, float_format, method])    汇总输出OR、RR的估计和检验结果

    symmetry([method])
    test_nominal_association()
    test_ordinal_association([row_scores, ...])

    log_oddsratio_confint([alpha, method])    CI
    log_oddsratio_pvalue([null])    P-value
    log_oddsratio_se()

    log_riskratio()
    log_riskratio_confint([alpha, method])    CI
    log_riskratio_pvalue([null])    p-value
    log_riskratio_se()

    oddsratio()
    oddsratio_confint([alpha, method])    CI
    oddsratio_pvalue([null])    P-value

    riskratio()
    riskratio_confint([alpha, method])    CI
    riskratio_pvalue([null])    p-value
```

In [ ]:

```
import numpy as np
import statsmodels.stats.contingency_tables as tbl

# 这里必须使用np.asarray函数进行转换，否则后续计算可能报错
table = tbl.Table2x2(np.asarray(pd.crosstab(ccss.Ts9, ccss.O1)))
table
```

In [ ]:

```
table.oddsratio
```

In [ ]:

```
table.summary()
```

7.4 实战练习

使用适当的指标表述职业和总信心指数之间的关联性。

使用适当的指标表述职业和汽车拥有情况之间的关联性。

8 线性回归模型入门

8.1 线性回归模型的基本原理

8.1.1 相关与回归的区别和联系

8.1.2 线性回归模型概述

8.1.3 线性回归模型的适用条件

8.1.4 线性回归模型的标准建模步骤

8.2 线性回归模型的Python实现

8.2.1 scipy的实现方式

scipy.stats.linregress(

x, y ： 类数组格式的自变量、因变量，均为一维，也可以直接以k*2的二维数组格式提供
注意：该命令的参数格式是自变量x在前！

)

返回结果：