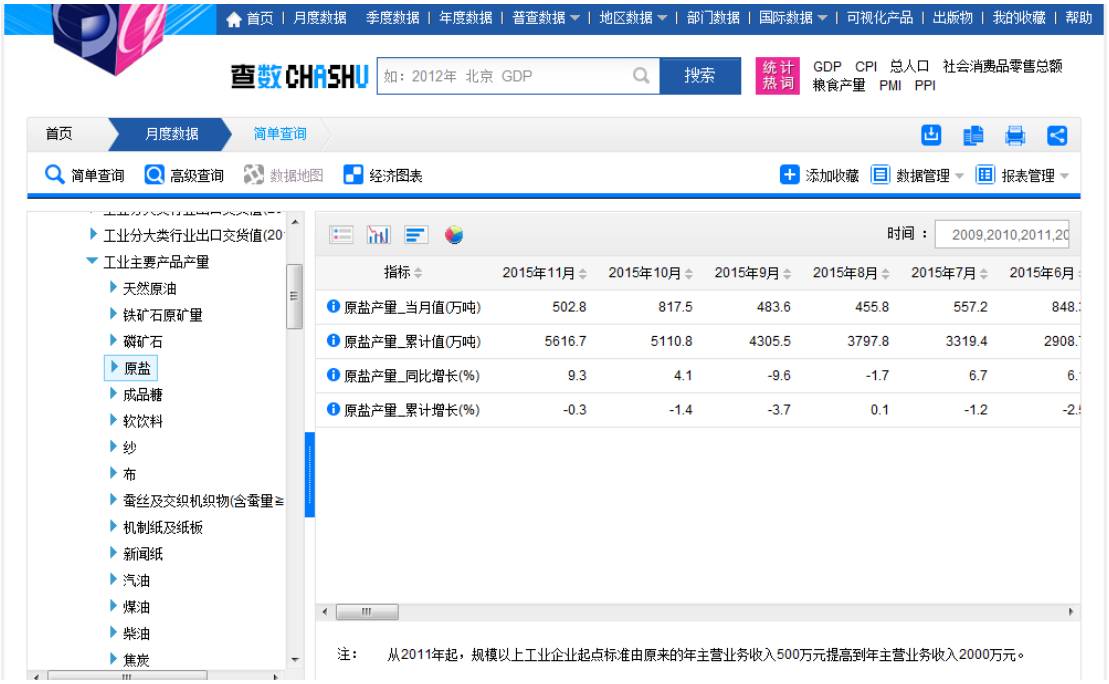
数理统计第二三章作业
15S003005　成坚

数据下载

之前我们从国家统计局网站上下载到了近三年的原盐数据，现在继续下 2009~2015 的产量数据，如下图所示。



示例程序

源代码就是 **scipytst.py**，用 **Python** 实现，利用了 **numpy** 和 **scipy** 库，读取 **2009~2015.xml** 的数据，并进行解析，统计数据。

求近三年数据均值的置信区间

理论基础:课本 **page53-2.3.3**　大样本区间估计-一般总体均值的区间估计

由于方差 σ^2 未知，使用样本标准差 **s** 代替 σ 得到 **u** 的置信区间，当 **n** 充分大时候，有

$\frac{\bar{x}-u}{s}\sqrt{n}$ 近似服从 **N(0, 1)**,故得 **u** 的置信区间为

$$(\bar{x}-u_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\bar{x}+u_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}})$$

对应与代码中的 **ConfidenceIntervalMean(data, ci = 0.975)**
其中 **data** 是待处理的数据，**ci** 是置信度
主要代码如下

```
    data --  待处理的数据
    ci   --  置信度
    """
    # 首先计算均值，方差，标准差
    dataArray = np.array(data)
    mean = dataArray.mean( )            #  计算均值
    print "Mean = ", mean
    #  numpy进行统计时，方差和标准差的自由度默认为N，  但是可通过参数ddof调整ddof : int, optional
    #  "Delta Degrees of Freedom": the divisor used in the calculation is N - ddof,
    #  where N represents the number of elements. By default ddof is zero.
    #  http://docs.scipy.org/doc/numpy/reference/generated/numpy.var.html
    var = dataArray.var(ddof = 1)    #  计算方差，自由度为N - 1
    print "Var  = ", var
    std = dataArray.std(ddof = 1)    #  计算标准差，自由度为N - 1
    print "Std  = ", std

    #(sqrt(n) *(x - u)) / S  近似服从于N(0, 1)
    #normal approximation interval
    #print ss.norm.ppf(0.975) = U0.025 = 1.96
    #q = (ci > 0.5) ? (ci : 1 - ci)
    if ci > 0.5 :
        q = ci
    else :
        q = 1 - ci
    #  为了找到一个分部的中心，我们可以使用分位数函数ppf，其是cdf的逆。
    scaled_crit = ss.norm.ppf(q) * std / np.sqrt(len(dataArray))   #  计算Uci * S / sqrt(N)
    low = mean - scaled_crit                                       #  计算置信下界
    up = mean + scaled_crit                                        #  计算置信上界

    print "The Sample  confidence level %f%% per cent confidence interval is [%f, %f]" % (q * 100, low, up)
```

运行结果如图

```
For three consecutive years(36 month) the monthly production of salt information :
 [502.8, 817.5, 483.6, 455.8, 557.2, 848.3, 669.0, 400.8, 368.2, 368.2, 557.2,
 403.8, 509.9, 820.2, 599.8, 517.4, 570.3, 872.3, 794.8, 447.4, 401.6, 447.4,
 401.6, 438.7, 600.7, 761.9, 578.0, 560.2, 553.3, 849.5, 613.3, 436.6, 355.7,
 436.6, 553.3, 408.9, 502.2, 791.2, 561.1, 518.9, 541.1, 856.2, 716.7, 416.2,
 375.8, 312.8, 375.8, 444.6, 484.5, 713.1, 573.0, 495.2, 699.9, 772.3, 732.8,
 487.4, 380.2, 315.1, 495.2, 533.2, 534.2, 771.9, 631.8, 546.9, 686.0, 789.3,
 534.4, 356.4, 318.8, 258.0, 534.4, 398.3, 532.4, 816.8, 628.3, 470.5, 553.3,
 816.2, 713.1, 349.6, 387.0, 211.3, 534.4]
Mean =  554.494444444
Var  =  24133.0313968
Std  =  155.3480975
The Sample  confidence level 97.500000% per cent confidence interval is
[503.748332, 605.240557]
```

我们得到数据的置信区间如下

置信度	置信区间
90.000000%	[521.313345, 587.675544]
95.000000%	[511.906964, 597.081925]
97.500000%	[503.748332, 605.240557]
99.500000%	[487.802747, 621.186141]
99.900000%	[474.484160, 634.504729]

显著性差异比较

比较近三年（**2013-2015**）和前几年（**2009-2012**）的原盐产量数据有无明显差异。

使用 t 检验，代码对应于 **StudentTest** 函数和 **SignificantDifference** 函数，两者都是进行 t 检验，只不过实现方式不同。

我们选择的统计量 $t=\dfrac{\bar{x}-\bar{y}}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$，其中 $S_w=\sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$

拒绝域为 $t\geq t_\alpha(n_1+n_2-2)$

主要代码如下

```
105     #  如果H0成立，那么P()
106     #  计算t统计量
107     statSw = np.sqrt(((lenOld - 1.0) * varOld + (lenNew - 1.0) * varNew) / (lenOld + lenNew
108     statT = (meanOld - meanNew)  / statSw / np.sqrt((1.0 / lenOld) + (1.0 / lenNew))
109     #print statSw
110     #print np.sqrt((1.0 / lenOld) + (1.0 / lenNew))
111     #print (meanOld - meanNew) / statSw / np.sqrt((1.0 / lenOld) + (1.0 / lenNew))
112     if ci > 0.5 :
113         q = ci
114     else :
115         q = 1 - ci
116     ppfT = ss.t.ppf(q, df = lenOld + lenNew - 2)
117     print "置信度%f%%拒绝域为 ta(%d + %d - 2) >= %f" % (q * 100, lenOld, lenNew, ppfT)
118 
119     if statT >= ppfT :
120         print "t统计量%f > %f%%分位数%f" % (statT, q * 100, ppfT)
121         return True
122     else :
123         print "t统计量%f < %f%%分位数%f" % (statT, q * 100, ppfT)
124         return False
125 
```

运行结果

```
2015年数据  [502.8, 817.5, 483.6, 455.8, 557.2, 848.3, 669.0, 400.8, 368.2, 368.2, 557.2]
2014年数据  [403.8, 509.9, 820.2, 599.8, 517.4, 570.3, 872.3, 794.8, 447.4, 401.6, 447.4, 401.6]
2013年数据  [438.7, 600.7, 761.9, 578.0, 560.2, 553.3, 849.5, 613.3, 436.6, 355.7, 436.6, 553.3]
2012年数据  [408.9, 502.2, 791.2, 561.1, 518.9, 541.1, 856.2, 716.7, 416.2, 375.8, 312.8, 375.8]
2011年数据  [444.6, 484.5, 713.1, 573.0, 495.2, 699.9, 772.3, 732.8, 487.4, 380.2, 315.1, 495.2]
2010年数据  [533.2, 534.2, 771.9, 631.8, 546.9, 686.0, 789.3, 534.4, 356.4, 318.8, 258.0, 534.4]
2009年数据  [398.3, 532.4, 816.8, 628.3, 470.5, 553.3, 816.2, 713.1, 349.6, 387.0, 211.3, 534.4]
-0.246900588054 0.807382814034
2015年的原盐产量与2014年的原盐产无显著性差别
The 2009~2011 data : Length = 35, Mean = 558.654286, Var = 24201.550202, Std = 155.568474
The 2012~2015 data : Length = 48, Mean = 539.097917, Var = 27001.024038, Std = 164.319883
置信度97.500000%拒绝域为 ta(35 + 48 - 2) >= 1.989686
t统计量0.547489 < 97.500000%分位数1.989686
2013~2015年的原盐产量与2009~2012年的原盐产无显著性差别
```

我们先统计了 2015 年原盐产量与 2014 年原盐产量的差异，然后又比较了 2013~2015 和 2009~2012 几年的原盐产量的差异。

我们发现均无显著性差别，但是 2015 年和 2014 年原盐产量的差异更小