模型-评价主题-统计类评价-Pearson相关系数检验【czy】

1. 模型名称

Pearson相关系数检验 (Pearson Correlation Coefficient Test)

2. 适用范围

对于两个**定距**或**定比变量**，度量相关性 [1] 最常用的统计量是Pearson相关系数，简称相关系数。

3. 形式

X,Y两个随机变量，n组样本数据(x_1,y_1),(x_2,y_2),...,(x_n,y_n)

4. 求解方法

4.1 概念

通常用**ρ**表示总体的相关系数，而用**r**表示样本之间的相关系数。

Pearson相关系数的取值范围是(−1,1)，0表示两个变量之间没有相关性，相关系数的绝对值越大表示变量之间的相关性就越强。相关系数的符号为正时表示两组变量成正相关，为负时表示成负相关 [2] 。

若X与Y为任意两个随机变量，则其总体相关系数ρ定义如下：

$$\rho=\frac{\operatorname{cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

4.2 步骤

1. **计算**r

从总体中选取n个二维变量$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，样本相关系数r定义如下：

$$\begin{aligned}r&=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}\\&=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n}x_i^2-n\bar{x}^2)(\sum_{i=1}^{n}y_i^2-n\bar{y}^2)}}\end{aligned}$$

若令$S_{AB}=\sum\limits_{i=1}^{n}a_ib_i-n\bar{a}\bar{b}(A,B\in\{X,Y\})$，则上式可简化为

$$r=\frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}$$

2. **显著性检验**

步骤如下：

2.1 提出假设

$$\text{双尾检验}:\begin{cases}H_0:\rho=0\\H_1:\rho\neq 0\end{cases}$$

$$\text{左尾检验}:\begin{cases}H_0:\rho=0\\H_1:\rho<0\end{cases}$$

$$\text{右尾检验}:\begin{cases}H_0:\rho=0\\H_1:\rho>0\end{cases}$$

上面的三种假设分别对应了三种情况，根据实际需求选取一种即可。

2.2 选定显著性水平α（一般为0.05），确定n的值。

2.3 确定检验统计量：

$$T=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

其中T服从自由度为n−2的t分布。

2.4 确定显著性水平α下的拒绝域𝕎：

双尾检验：𝕎={$T\mid T>t_{\alpha/2}(n-2)$或$T<-t_{\alpha/2}(n-2)$}

左尾检验：𝕎={$T\mid T<-t_{\alpha}(n-2)$}

右尾检验：𝕎={$T\mid T>t_{\alpha}(n-2)$}

2.5 根据统计量的值决定拒绝或接受原假设H_0。

4.3 例子

假设我们有一个由10个美国高中毕业生组成的样本，记录下他们SAT（美国学习能力测验）考试中语言和数学部分的成绩，具体数据见表2（每一科的分数都在200~800）。

表2 SAT 考试中语文和数学成绩

学生	1	2	3	4	5	6	7	8	9	10
语文	490	500	530	550	580	590	600	600	650	700
数学	560	500	510	600	600	620	550	630	650	750

根据表中的数据，请问语文成绩和数学成绩是否线性相关？

第一步：求相关系数r

首先计算几个基本量，结果见表3。

表3 基本统计量

n	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} y_i$	$_{i=1}^{n} y_i^2$	$\sum_{i=1}^{n} x_i y_i$
10	5790	3390500	5790	3612500	3494000

接下来把这些值代入计算公式中，可得

$S_{XX} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 = 38090$

$S_{YY} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 = 48410$

$S_{XY} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) = 37370$

$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} = \frac{37370}{\sqrt{38090 \times 48410}} = 0.87$

语文成绩和数学成绩之间的相关系数是0.87，可以认为是很强的正相关关系，表明某一科目分数很高的学生，通常另一科目的分数也会很高。[3]

只看相关系数r，我们还无法完全确定两个变量之间的**关系显著程度**，为此我们还要做一个显著性检验，步骤如下：

第二步：显著性检验

1.提出假设：

$$\text{双尾检验}: \begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

2.选定显著性水平

$$\alpha = 0.05,\ n = 10。$$

3.确定检验统计量：

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.87 \times \sqrt{10-2}}{\sqrt{1-0.87^2}} = 5.02$$

其中T服从自由度为$n-2$的t分布。

4.确定显著性水平α下的拒绝域$\mathbb{W}$，采用**双尾检验**

查表得知$t_{\alpha/2} = t_{0.025} = 2.306$，

因此拒绝域为$\mathbb{W}$={$T \mid T < -2.306$或$T > 2.306$}

5.分析与总结

因为$T = 5.02 > 2.306$，在拒绝域内，所以我们拒绝语文成绩和数学成绩不相关的原假设，也即**二者之间存在线性相关。**

4.4 代码实现

4.4.1 Matlab

```
% 学生的语文成绩
language = [490, 500, 530, 550, 580, 590, 600, 600, 650, 700];
% 学生的数学成绩
math = [560, 500, 510, 600, 600, 620, 550, 630, 650, 750];
% 将两门课的成绩、显著性水平和检验模式（单尾/双尾）输入pearson()函数，得到检验结果
[H, r, T, t] = pearson(language, math, 0.05, 'both')

% function [ H, r, T, t ] = pearson( X, Y, alpha, tailType )
```

```matlab
% 对两个变量进行Pearson相关系数显著性检验的MATLAB程序代码

% H 表示最终所接受的假设。若为0，表示接受原假设（线性无关）；若为1，表示拒绝原假设。
% r 由样本所计算的Pearson相关系数。
% T 由r进一步计算的检验统计量T，服从自由度为n-2的t分布。
% t 查表所得的拒绝域边界值。若为左（右）尾检验，为t_a(n-2)，否则为t_{a/2}(n-2)。
% X,Y 表示两个变量的一组样本。
% alpha 显著性水平，默认0.05。
% tailType 验证类型，可取值：
%           'both'：双尾检验（默认）。
%           'left'：左尾检验。
%           'right'：右尾检验。

% 以下为一般过程
function [ H, r, T, t ] = pearson( X, Y, alpha, tailType )
% function [ H, r, T, t ] = pearson( X, Y, alpha, tailType )
% 对两个变量进行Pearson相关系数显著性检验的MATLAB程序代码

% H 表示最终所接受的假设。若为0，表示接受原假设（线性无关）；若为1，表示拒绝原假设。
% r 由样本所计算的Pearson相关系数。
% T 由r进一步计算的检验统计量T，服从自由度为n-2的t分布。
% t 查表所得的拒绝域边界值。若为左（右）尾检验，为t_a(n-2)，否则为t_{a/2}(n-2)。
% X,Y 表示两个变量的一组样本。
% alpha 显著性水平，默认0.05。
% tailType 验证类型，可取值：
%           'both'：双尾检验（默认）。
%           'left'：左尾检验。
%           'right'：右尾检验。

% 注意事项：
% X,Y均为一维向量，且应有相同数目的元素。

%% 参数初始化。
if nargin<3
    alpha = 0.05;    %默认显著性水平。
end
if nargin<4
    tailType = 'both';  %默认验证类型。
end

%% 计算检验统计量。
r = corr(X(:), Y(:));    %Pearson相关系数。
n = length(X);  %样本数目。
V = n-2;    %自由度。
T = r*sqrt(V)/sqrt(1-r^2);  %检验统计量。

%% 确定拒绝域并给出最终假设。
switch lower(tailType)
    case 'both' %双尾检验。
        P = 1-alpha/2;  %转换概率，用于tinv函数。
        t = tinv(P, V); %根据累积分布函数的反函数求得拒绝域边界值。
        H = T>t || T<-t;    %得出最终结论。
    case 'left' %左尾检验。
        P = 1-alpha;
        t = tinv(P, V);
        H = T<-t;
    case 'right'    %右尾检验。
        P = 1-alpha;
```

```
        t = tinv(P, V);
        H = T>t;
    otherwise
        error('unknown tail type!');
end

%% 调用方式
% language = [490, 500, 530, 550, 580, 590, 600, 600, 650, 700];
% math = [560, 500, 510, 600, 600, 620, 550, 630, 650, 750];
% [H, r, T, t] = pearson(language, math, 0.05, 'both')

end
```

4.4.2 Python

```
from pearson import pearson

# 学生的语文成绩
languageX = [490, 500, 530, 550, 580, 590, 600, 600, 650, 700]
# 学生的数学成绩
mathY = [560, 500, 510, 600, 600, 620, 550, 630, 650, 750]
# 将两门课的成绩、显著性水平和检验模式（单尾/双尾）输入pearson()函数，得到检验结果
H, r, T, t = pearson(languageX, mathY, 0.05, 'both')

print "H=%d\nr=%f\nT=%f\nt=%f\n" % (H, r, T, t)

# pearson()函数的输入输出与matlab版本类似，在此不做多余说明

# 以下为一般过程
# -*- coding: utf-8 -*-
"""
#################################################
# Pearson相关系数显著性检验的PYTHON程序代码。
# By:Tang Jiajun
#################################################
# 注意事项
# 1.将文件放置于你的代码的相同文件夹中，通过import导入。
# 2.X,Y应为相同长度的列表。
# 3.使用了scipy库，需要预先安装。
#################################################
# 调用样例
from pearson import pearson

languageX = [490, 500, 530, 550, 580, 590, 600, 600, 650, 700]
mathY = [560, 500, 510, 600, 600, 620, 550, 630, 650, 750]
H, r, T, t = pearson(languageX, mathY, 0.05, 'both')

print "H=%d\nr=%f\nT=%f\nt=%f\n" % (H, r, T, t)
#################################################
"""

import scipy.stats

def pearson(X, Y, alpha=0.05, tailType='both'):
    """
    用于Pearson相关系数显著性检验的函数。
    H, r, T, t = pearson(X, Y, alpha=0.05, tailType='both')
```

```python
    # H 表示最终所接受的假设。若为0，表示接受原假设（线性无关）；若为1，表示拒绝原假设。
    # r 由样本所计算的Pearson相关系数。
    # T 由r进一步计算的检验统计量T，服从自由度为n-2的t分布。
    # t 查表所得的拒绝域边界值。若为左（右）尾检验，为t_a(n-2)，否则为t_{a/2}(n-2)。
    # X,Y 表示两个变量的一组样本。
    # alpha 显著性水平，默认0.05。
    # tailType 验证类型，可取值：
    #       'both'：双尾检验（默认）。
    #       'left'：左尾检验。
    #       'right'：右尾检验。
    """

    ## 计算检验统计量。
    r = scipy.stats.pearsonr(X, Y)[0]   #Pearson相关系数。
    n = len(X)  #样本数目。
    V = n-2 #自由度。
    T = r*V**0.5/(1-r**2)**0.5  #检验统计量。

    ## 确定拒绝域并给出最终假设。
    tailType = tailType.lower() #转换为小写。
    if tailType == 'both':  #双尾检验。
        P = 1-alpha/2.0 #转换概率，用于ppf函数。
        t = scipy.stats.t.ppf(P, V) #根据累积分布函数的反函数求得拒绝域边界值。
        H = T > t or T < -t #得出最终结论。
    elif tailType == 'left':    #左尾检验。
        P = 1-alpha
        t = scipy.stats.t.ppf(P, V)
        H = T < -t
    elif tailType == 'right':   #右尾检验。
        P = 1-alpha
        t = scipy.stats.t.ppf(P, V)
        H = T > t
    else: raise ValueError
    H = int(H)  #将bool值转换为整数类型。

    return H, r, T, t
```

4.4.3 C++

```cpp
#include "pearson.h"

using namespace std;

int main(){
    // 学生的语文成绩
    alglib::real_1d_array language="[490,500,530,550,580,590,600,600,650,700]";
    // 学生的数学成绩
    alglib::real_1d_array math="[560,500,510,600,600,620,550,630,650,750]";
    // 将两门课的成绩、显著性水平和检验模式（单尾/双尾）输入pearson()函数，得到检验结果
    pearResult res=pearson(language, math, 0.05, "both");

    cout<<"H:"<<res.H<<endl;
    cout<<"r:"<<res.r<<endl;
    cout<<"T:"<<res.T<<endl;
    cout<<"t:"<<res.t<<endl;
    return 0;
}
```

```
// pearson()函数的输入输出与matlab版本类似，在此不做多余说明

// 以下为一般过程
/*
# Pearson相关系数显著性检验的C++程序代码。
# By: Tang Jiajun
*/

//----------------------------------
/*
注意事项：
# 1.将该文件夹所有文件，包括alglib文件夹放置于你的代码的相同文件夹中，通过include引入。
# 2.X,Y应为相同长度的一维向量，可由字符串定义。
# 3.使用了alglib库，已提供在alglib文件夹下。
# 4.函数默认参数值定义在pearson.h头文件的函数声明中。
*/

//----------------------------------
/*
调用样例：
#include "pearson.h"

using namespace std;

int main(){
    alglib::real_1d_array language="[490,500,530,550,580,590,600,600,650,700]";
    alglib::real_1d_array math="[560,500,510,600,600,620,550,630,650,750]";
    pearResult res=pearson(language, math, 0.05, "both");

    cout<<"H:"<<res.H<<endl;
    cout<<"r:"<<res.r<<endl;
    cout<<"T:"<<res.T<<endl;
    cout<<"t:"<<res.t<<endl;
    return 0;
}
*/

//----------------------------------
#include "pearson.h"

using namespace alglib;

pearResult pearson(const real_1d_array& X,const real_1d_array& Y, double alpha,
std::string tailType)
{
    /*
    用于Pearson相关系数显著性检验的函数。
    pearResult res = pearson(const real_1d_array& X,const real_1d_array& Y,
double alpha=0.05, std::string tailType="both")
    pearResult结构体定义如下：
    struct pearResult{
        int H;
        double r, T, t;
    };
    # H 表示最终所接受的假设。若为0，表示接受原假设（线性无关）；若为1，表示拒绝原假设。
    # r 由样本本所计算的Pearson相关系数。
    # T 由r进一步计算的检验统计量T，服从自由度为n-2的t分布。
```

```
    # t 查表所得的拒绝域边界值。若为左（右）尾检验，为t_a(n-2)，否则为t_{a/2}(n-2)。
    # X,Y 表示两个变量的一组样本。
    # alpha 显著性水平，默认0.05。
    # tailType 验证类型，可取值：
    #       "both"：双尾检验（默认）。
    #       "left"：左尾检验。
    #       "right"：右尾检验。
    */

    // 计算检验统计量。
    pearResult res;
    res.r = pearsoncorr2(X,Y);  //Pearson相关系数。
    long long n = (long long)X.length();    //样本数目。
    long long v = n - 2;    //自由度。
    res.T = res.r*std::sqrt((double)v)/std::sqrt(1-res.r*res.r);    //检验统计量。

    // 确定拒绝域并给出最终假设。
    std::transform(tailType.begin(), tailType.end(), tailType.begin(), tolower);
   //转换为小写。
    if (tailType=="both"){  //双尾检验。
        double P = 1 - alpha/2; //转换概率，用于invstudenttdistribution函数。
        res.t = invstudenttdistribution(v, P);  //根据累积分布函数函数的反函数求得拒绝
域边界值。
        res.H = int(res.T > res.t || res.T < -res.t);   //得出最终结论。
    }else if(tailType=="left"){ //左尾检验。
        double P = 1-alpha;
        res.t = invstudenttdistribution(v, P);
        res.H = int(res.T < -res.t);
    }else if(tailType=="right"){    //右尾检验。
        double P = 1-alpha;
        res.t = invstudenttdistribution(v, P);
        res.H = int(res.T > res.t);
    }else{
        std::cerr << "unknown tail type!" << std::endl;
        exit(1);
    }

    return res;
}
```

5.参考资料

1. 数学建模培训营----Pearson相关系数检验
2. 《高晓沨----数据分析模型概览(PPT)》 P54~P56
3. Stata: 快速呈现常用分布临界值表

---

1. 。一般来说，皮尔森相关系数的绝对值越强，说明两个变量的相关程度越大，除了从相关系数直接判断以外，还应掌握皮尔森相关系数的显著性检验。通过皮尔森相关系数检验（Pearson Correlation Coefficient Test）我们能够确定两个变量之间究竟有没有关联。 ↩

2. Pearson相关系数在变量是非线性相关的情况下很容易产生误导结果，这也是我们需要画出散点图作参考的原因。相关性的"强"和"弱"都没有严格的数值定义，但如果说相关性强，那么基本上就比弱相关的数据具有更明显的线性关系，其数据点也会比弱相关的数据点更集中地分布在一条直线附近 ↩

3. 注意到相关性是一种对称的关系，因此我们不轻易认为是一个变量导致了另一个变量，除非我们已经观察到变量之间的这种因果关系 ↩