皮尔逊相关系数假设检验

1.

第一步：提出原假设H_0和备择假设H_1（两个假设是截然相反的哦）

假设我们计算出了一个皮尔逊相关系数r，我们想检验它是否显著的异于0.

那么我们可以这样设定原假设和备择假设：$H_0: r = 0$ ，$H_1: r \neq 0$

第二步：在原假设成立的条件下，利用我们要检验的量构造出一个符合某一分布的统计量

（注1：统计量相当于我们要检验的量的一个函数，里面不能有其他的随机变量）

（注2：这里的分布一般有四种：标准正态分布、t分布、χ^2分布和F分布）

对于皮尔逊相关系数r而言，在满足一定条件下，我们可以构造统计量：

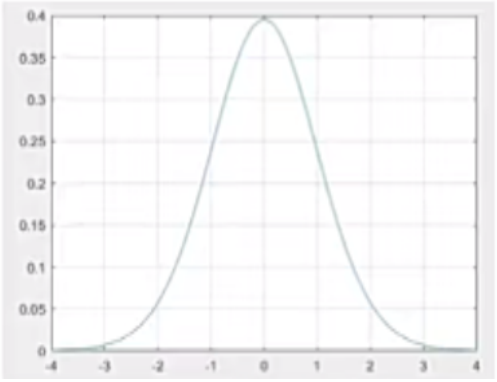
$t = r\sqrt{\frac{n-2}{1-r^2}}$，可以证明$t$是服从自由度为$n-2$的$t$分布

第三步：将我们要检验的这个值带入这个统计量中，可以得到一个特定的值（检验值）。

例如，我们计算出的相关系数为0.5，$n = 30$，那么我们可以得到$t^* = 0.5\sqrt{\frac{30-2}{1-0.5^2}} = 3.05505$

第四步：由于我们知道统计量的分布情况，因此我们可以画出该分布的概率密度函数pdf，并给定一个置信水平，根据这个置信水平查表找到临界值，并画出检验统计量的接受域和拒绝域。

例如，我们知道上述统计量服从自由度为28的t分布，其概率密度函数图形如下：




```
x = -4:0.1:4;
y = tpdf(x,28);
plot(x,y,'-')
grid on   % 在画出的图上加上网格线
```

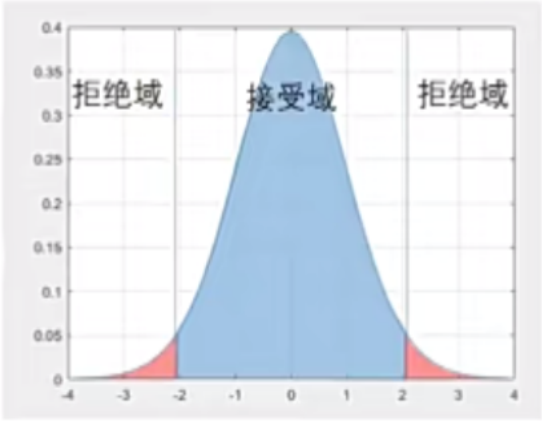
第四步：由于我们知道统计量的分布情况，因此我们可以画出该分布的概率密度函数pdf，并给定一个置信水平，根据这个置信水平查表找到临界值，并画出检验统计量的接受域和拒绝域。

常见的置信水平有三个：90%，95%和99%，其中95%是三者中最为常用的。

因为我们这里是双侧检验，所以我们需要找出能覆盖0.95概率的部分

t分布表：https://wenku.baidu.com/view/d94dbd116bd97f192279e94a.html

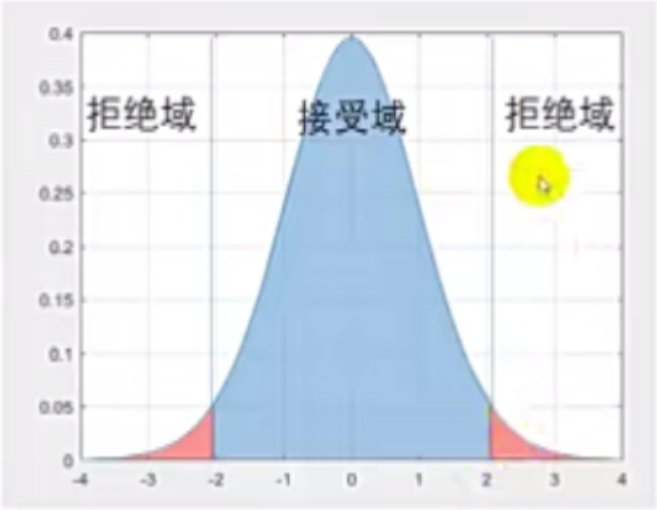
查表可知，对应的临界值为2.048，因此我们可以做出接受域和拒绝域。



第五步：看我们计算出来的检验值是落在了拒绝域还是接受域，并下结论。

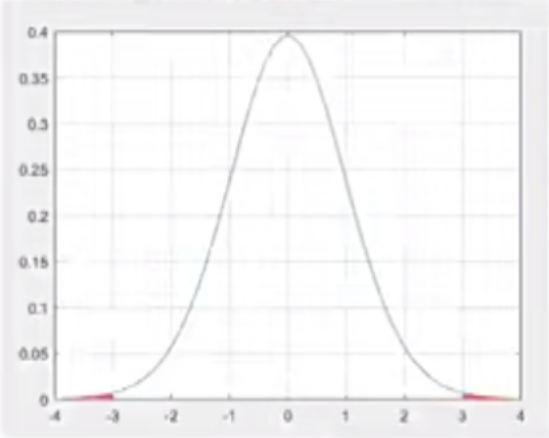
因为我们得到的$t^* = 3.05505 > 2.048$，因此我们可以下结论：

在95%的置信水平上，我们拒绝原假设$H_0: r = 0$，因此r是显著的不为0的。



2.更简便的方法(P值判断法)

我们得到的检验值$t^* = 3.05505$，根据这个值，我们可以计算出其对应的那个概率




```
disp('该检验值对应的p值为：')
disp((1-tcdf(3.055,28))*2)
%双侧检验的p值要乘以2
```

注意这里的函数是tcdf：累计概率密度函数

最后我们计算得到的p值为：0.0049

$p<0.01$，说明在99%的置信水平上拒绝原假设；
$p<0.05$，说明在95%的置信水平上拒绝原假设；
$p<0.10$，说明在90%的置信水平上拒绝原假设；

p值>0.01，说明在99%的置信水平无法拒绝原假设；
$p>0.05$，说明在95%的置信水平上无法拒绝原假设；
$p>0.10$，说明在90%的置信水平上无法拒绝原假设；

在本例中，拒绝原假设意味着皮尔逊相关系数显著的异于0。
小补充：0.5、0.5* 、0.5**、0.5***的含义是什么？（显著性标记）

计算各列之间的相关系数以及p值

一行代码：[R,P] = corrcoef(Test)
R返回的是相关系数表，P返回的是对应于每个相关系数的p值

R =

1.0000	0.0665	-0.2177	-0.1920	0.0440	0.0951
0.0665	1.0000	0.0954	0.0685	0.0279	-0.0161
-0.2177	0.0954	1.0000	0.2898	0.0248	-0.0749
-0.1920	0.0685	0.2898	1.0000	-0.0587	-0.0019
0.0440	0.0279	0.0248	-0.0587	1.0000	-0.0174
0.0951	-0.0161	-0.0749	-0.0019	-0.0174	1.0000

P =

1.0000	0.1061	0.0000	0.0000	0.2859	0.0208
0.1061	1.0000	0.0204	0.0960	0.4978	0.6963
0.0000	0.0204	1.0000	0.0000	0.5469	0.0687
0.0000	0.0960	0.0000	1.0000	0.1542	0.9637
0.2859	0.4978	0.5469	0.1542	1.0000	0.6728
0.0208	0.6963	0.0687	0.9637	0.6728	1.0000

	身高	体重	肺活量	50米跑	立定跳远	坐位体前屈
身高	1.0000	0.0665	-0.2177***	-0.192***	0.0440	0.0951**
体重	0.0665	1.0000	0.0954**	0.0685*	0.0279	-0.0161
肺活量	-0.2177***	0.0954**	1.0000	0.2898***	0.0248	-0.0749*
50米跑	-0.192***	0.0685*	0.2898***	1.0000	-0.0587	-0.0019
立定跳远	0.0440	0.0279	0.0248	-0.0587	1.0000	-0.0174
坐位体前屈	0.0951**	-0.0161	-0.0749*	-0.0019	-0.0174	1.0000

*** p<0.01, ** p<0.05, * p<0.1

```
%% 计算各列之间的相关系数以及p值
[R,P] = corrcoef(Test)
% 在EXCEL表格中给数据右上角标上显著性符号吧
P < 0.01   % 标记3颗星的位置
(P < 0.05) .* (P > 0.01)   % 标记2颗星的位置
(P < 0.1) .* (P > 0.05) % % 标记1颗星的位置
```

3.

皮尔逊相关系数假设检验的条件

第一，实验数据通常假设是成对的来自于正态分布的总体。因为我们在求皮尔逊相关性系数以后，通常还会用t检验之类的方法来进行皮尔逊相关性系数检验，而t检验是基于数据呈正态分布的假设的。

第二，实验数据之间的差距不能太大。皮尔逊相关性系数受异常值的影响比较大。

第三：每组样本之间是独立抽样的。构造t统计量时需要用到。

如何检验数据是否是正态分布?

怎样检验数据是否是正态分布?

1.

正态分布JB检验（大样本 n>30）

雅克-贝拉检验(Jarque-Bera test)

对于一个随机变量 $\{X_i\}$，假设其偏度为S，峰度为K，那么我们可以构造JB统计量：

$$JB = \frac{n}{6}[S^2 + \frac{(K-3)^2}{4}]$$

可以证明，如果 $\{X_i\}$ 是正态分布，那么在大样本情况下JB-$\chi^2(2)$ （自由度为2的卡方分布）

注：正态分布的偏度为0，峰度为3

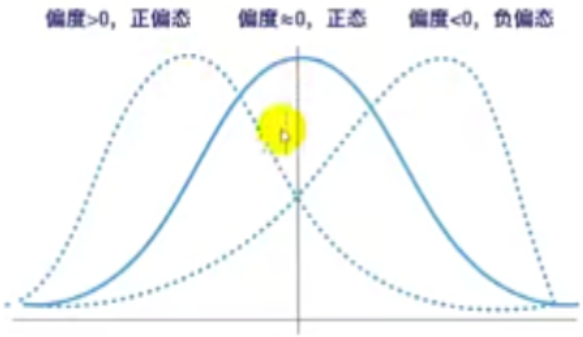
那么进行假设检验的步骤如下：

H_0：该随机变量服从正态分布　H_1：该随机变量不服从正态分布

然后计算该变量的偏度和峰度，得到检验值JB^*，并计算出其对应的p值

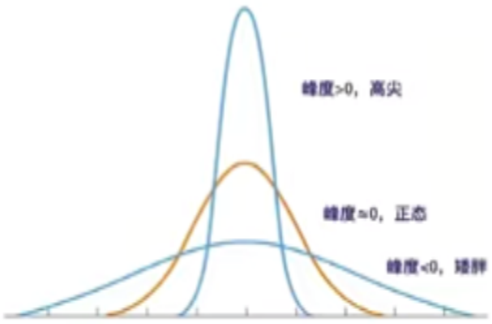
将p值与0.05比较，如果小于0.05则可拒绝原假设，否则我们不能拒绝原假设。

偏度和峰度



$$E[(\frac{X-u}{\sigma})^3]$$

正态分布的偏度为0



$$E[(\frac{X-u}{\sigma})^4] 或 E[(\frac{X-u}{\sigma})^4]-3$$

正态分布的偏度为3
(有些地方定义的正态分布偏度为0)
Matlab软件中使用的是第一种定义

```
x = normrnd(2,3,100,1);
% 生成100*1的随机向量，每个元素是均值为2，标准差为3的正态分布
skewness(x)   %偏度
kurtosis(x)   %峰度
```

MATLAB中进行JB检验的语法：[h,p] = jbtest(x,alpha)
当输出h等于1时，表示拒绝原假设；h等于0则代表不能拒绝原假设。
alpha就是显著性水平，一般取0.05，此时置信水平为1-0.05=0.95
x就是我们要检验的随机变量，注意这里的x只能是向量。

```
%% 正态分布检验
% 检验第一列数据是否为正态分布
[h,p] = jbtest(Test(:,1),0.05)

% 用循环检验所有列的数据
n_c = size(Test,2);  % number of column 数据的列数
H = zeros(1,6);
P = zeros(1,6);
for i = 1:n_c
    [h,p] = jbtest(Test(:,i),0.05);
    H(i)=h;
    P(i)=p;
end
disp(H)
disp(P)
```

2.

小样本3≤n≤50：Shapiro-wilk检验

Shapiro-wilk夏皮洛-威尔克检验

H_0：该随机变量服从正态分布　H_1：该随机变量不服从正态分布

计算出威尔克统计量后，得到相应的p值

将p值与0.05比较，如果小于0.05则可拒绝原假设，否则我们不能拒绝原假设。

3.QQ图方法

在统计学中，Q-Q图（Q代表分位数Quantile）是一种通过比较两个概率分布的分位数对这两个概率分布进行比较的概率图方法。

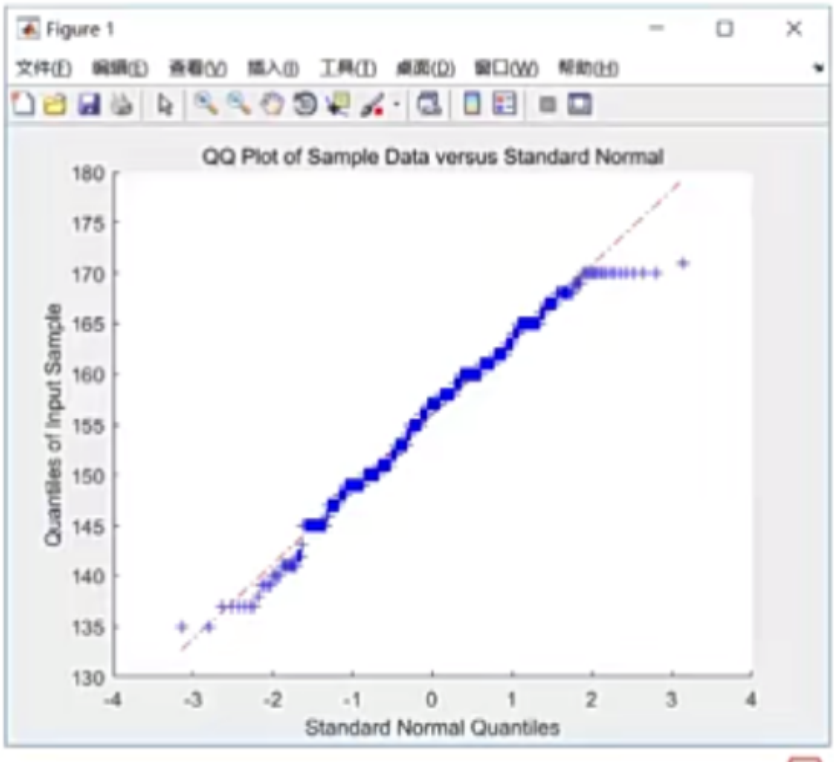
首先选定分位数的对应概率区间集合，在此概率区间上，点(x,y)对应于第一个分布的一个分位数x和第二个分布在和x相同概率区间上相同的分位数。

这里，我们选择正态分布和要检验的随机变量，并对其做出QQ图，可想而知，如果要检验的随机变量是正态分布，那么QQ图就是一条直线。

要利用Q-Q图鉴别样本数据是否近似于正态分布,只需看Q-Q图上的点是否近似地在一条直线附近。

第一列数据和正态分布的Q-Q图

```
qqplot(Test(:,1))
```



QQ图方法要求数据量非常大，数据量小的话，结果不够精准。