南 京 师 范 大 学

毕 业 设 计（论 文）

（ 2015 届）



题 目： 基于R语言的非参数统计p值计算中的复杂度

学 院： 数科院

姓 名： 韩亮亮

学号 ： 06150326

专业 ： 统计学

指导教师： 李启才

南京师范大学教务处制

独 创 性 声 明

秉承学校严谨的学风与优良的科学道德，本人声明所呈交的论文是我个人在导

师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，不包含本人或他人已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了致谢。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

论文作者签名： 日期：

摘要

随着很多科学领域与计算机领域的不断结合，对复杂的数据集的处理成为了统计学领域之外的人比如生物学家，医生，地质学家，等必须面临的问题，大量的非统计相关的专业的人员对统计方法的使用和统计人员对统计方法的不完全理解，带来了大量的对统计方法得来的结论的担忧。 许多已经发表过的科学结论的基础是“统计显著性”的概念，统计显著性经常被误用和被错误的陈述，这使得一些科学杂志不再鼓励p-值的使用。从p值诞生开始，就有观点不断地质疑p值的使用。这篇研究另辟蹊径，通过讨论p值的使用过程，用信息论的信息度量工具，来讨论究竟p值的使用为什么是复杂的。

通过理论分析，和信息量分析，本文认为p值的复杂性是来源于自身的，伴随着p值计算的本身，并不能够通过约束p值的使用和p值的解释规则来解决。

针对统计显著性和p值带来的问题，本文认为，尽可能的提高数据本身的透明度，是解决统计问题的一个真正有效的方法。面对同样的数据集，不同的人，或者是相同的人在不同的时间，可能做出来相差很大的结论。即便是同样的统计方法，经过不同的预处理也能得出差异很大的结论。而根据本文的研究，不管是通过软件工具包进行黑盒操作计算p值，还是通过R语言的基本数学方法计算p值，都有很大的不确定性，导致 p-值 在一个很大的范围内波动。

本文的建议，对于一些比较简单的，不需要太多预处理的统计模型，可以使用p值和其他的置信区间等信息联合判断论文的成果。对于复杂的模型，或者需要大量数据处理步骤的情况，不建议使用统计显著性作为科学发现的判断标准。

关键词：

统计审核，复杂度分析， R语言与P值，非参数统计方法

Abstract

With the continuous integration of many scientific fields and the computer field, the processing of complex data sets has become a problem that must be faced by people outside the field of statistics such as biologists, doctors, geologists, etc., and a large number of non-statistics. The use of statistical methods by relevant professional personnel and the incomplete understanding of statistical methods by statisticians have brought a lot of concerns about the conclusions drawn from statistical methods. Many of the scientific conclusions that have been published are based on the concept of “statistical significance”, where statistical significance is often misused and misrepresented, which has led some scientific journals to no longer encourage the use of p-values. Since the birth of the p-value, there has been a constant questioning about the use of p-values. This study takes a different approach by discussing the use of p-values ​​and using information-based information measurement tools to discuss why the use of p-values ​​is complex.

Through theoretical analysis and information analysis, this paper considers that the complexity of p-value is derived from itself, and the calculation of p-value itself cannot be solved by constraining the use of p-values ​​and the interpretation rules of p-values.

In view of the problems caused by statistical significance and p-value, this paper believes that improving the transparency of the data itself is a truly effective way to solve statistical problems. Faced with the same data set, different people, or the same person, at different times, may make a big difference. Even with the same statistical methods, different pre-processing can lead to very different conclusions. According to the research in this paper, whether the p-value is calculated by the software toolkit for the black box operation or the p-value is calculated by the basic mathematical method of the R language, there is great uncertainty, resulting in a large p-value. Fluctuations within the range.

The recommendations in this paper, for some simple statistical models that do not require too much preprocessing, can be used to jointly judge the results of the paper using information such as p-values ​​and other confidence intervals. For complex models, or where large data processing steps are required, statistical significance is not recommended as a criterion for scientific discovery.

keyword: Statistical review, complexity analysis, R language and P value, nonparametric statistical method

目录

[第一章 绪论 6](#_Toc7956277)

[第一节 选题背景与研究意义 6](#_Toc7956278)

[第二节 国内外研究综述 6](#_Toc7956279)

[第三节 研究思路，框架与方法 6](#_Toc7956280)

[第二章 概念界定与理论基础 7](#_Toc7956281)

[第三章 7](#_Toc7956282)

[R语言的p值计算案例 8](#_Toc7956283)

[第一节 8](#_Toc7956284)

[R语言的例子1 判断组函数的中位数是否是一个数 8](#_Toc7956285)

[说明 需要人为判断的环节，以及复杂度分析 8](#_Toc7956286)

[总结 9](#_Toc7956287)

[第二节 9](#_Toc7956288)

[R语言中的例子2 判断一个函数的0.25分位点是不是一个数值 9](#_Toc7956289)

[说明 9](#_Toc7956290)

[第三节 10](#_Toc7956291)

[R语言中的例子3 判断一个函数的0.25分位点是不是小于一个给定的数值多少 10](#_Toc7956292)

[说明 10](#_Toc7956293)

[总结 10](#_Toc7956294)

[第四章 复杂度增加导致p值被错误计算的机理 11](#_Toc7956295)

[论文的总结 12](#_Toc7956296)

[结论和展望 12](#_Toc7956297)

[第一节 研究结论 12](#_Toc7956298)

[第二节 难点与不足 12](#_Toc7956299)

[第三节 展望 12](#_Toc7956300)

[参考文献 12](#_Toc7956301)

[致谢 13](#_Toc7956302)

# 第一章 绪论

## 第一节 选题背景与研究意义

目前关于p值的使用和误用已经有了大量的讨论，并且在近两年，讨论甚嚣尘上，在美国国家统计协会的参考链接中，可以找到大量的这样的讨论，比如Benjamin, Daniel J, and Berger, James O: A simple alternative to p-values， Benjamini, Yoav: It’s not the p-values’ fault，因为在这些主张里有着不同的意见和大量的不同的观点， 而这些讨论本身更多的是基于经验的人文性质的讨论，本研究的目的，就是用更为科学的信息学的方法，来系统性的检查p值的问题，以及做一份对未来p值使用的建议。

## 第二节 国内外研究综述

国内目前研究这个课题的数量并不多，不做讨论。

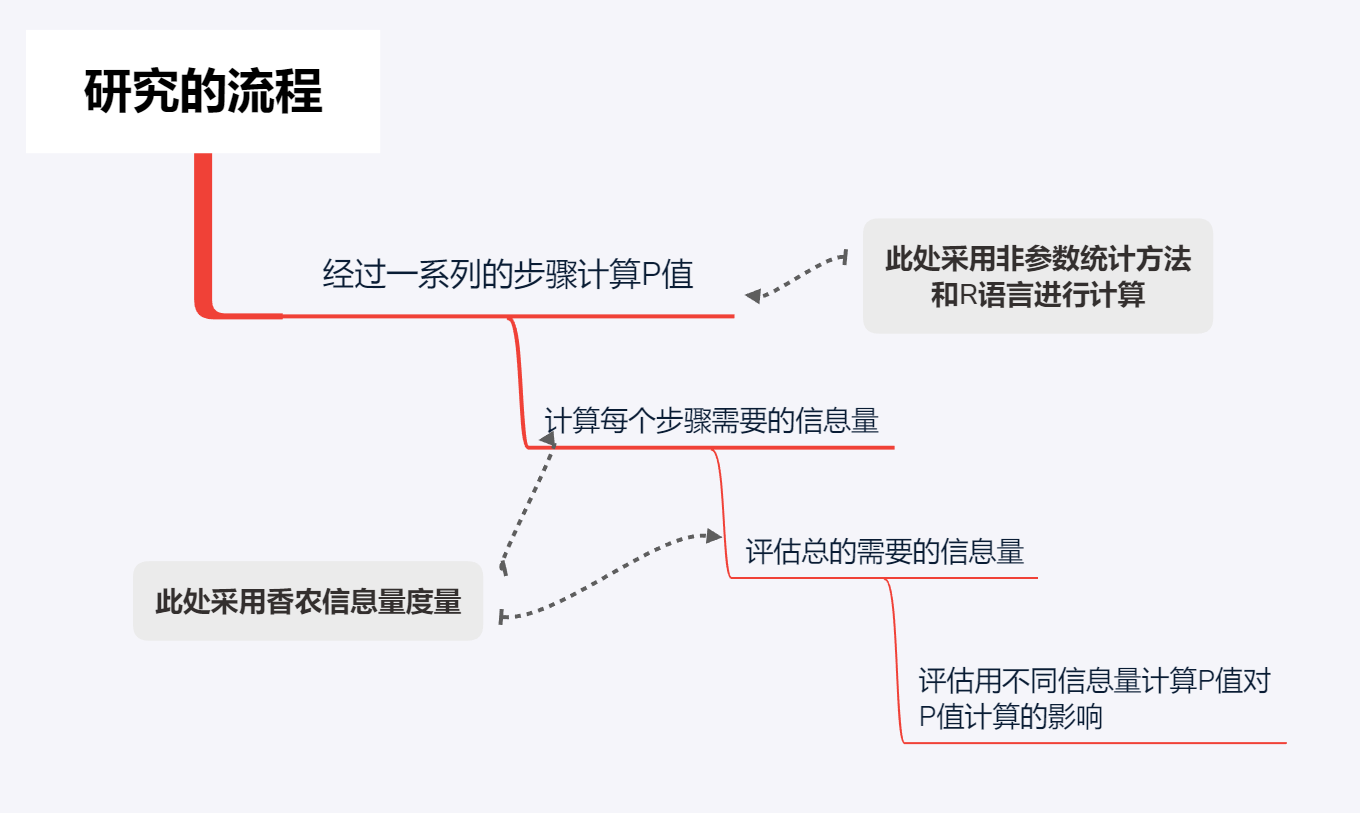
国外的话，面对p值的广泛误用，由相当多的讨论在各种论坛中出现，但是并没有给出使用p值的一致性意见（绝大部分关于p值使用的建议出现在论坛当中，而并非文献之中），论坛的大部分声音是站出来指出p值的大量缺陷，转而反对p值的使用[2]。考虑到信息来源的一手性，绝大多数在这些论坛发声的都是一手的统计工作人员。虽然他们讨论没有公开发表出来，但是作为工具使用人员的主流态度，值得高度重视。

美国国家统计协会（ASA）在The ASA's Statement on p-Values: Context, Process, and Purpose给出了一个使用p值的统一的参考性声明。这个声明给出了一些能够规范p值使用的准则。主要的态度是不鼓励抛弃p值的使用，但是希望规范p值的使用和更高度的信息透明化。

因为关于这个问题有着不同的立场，而研究更多的是在进行人文性质讨论。这篇文章不对市面上不同立场做任何讨论，转而希望通过信息论的度量方法分析，到底哪些时候p值容易理解和接受，那些地方则几乎不可能正确使用。

## 第三节 研究思路，框架与方法

本文通过使用R语言计算p值的三个最简单常用的例子来展开论述。因为在用R语言计算P值的整个计算过程中需要大量的需要人为处理的环节，每个计算处理环节有多少种选择。本文使用信息论的信息度量方法度量了把所有的这些环节进行正确需要的信息量，从而估计出正确的计算出p值 需要的信息量。 最后讨论不同信息量需求对计算p值有什么影响。

主要的流程如下图

# 第二章 概念界定与理论基础

#### 第一节 p值的概念界定

非正式地，p值是指定统计模型下数据的统计汇总（例如，样本）的概率,两个比较组之间的平均差异）将等于或高于其观察值。[1]

一个更为简单的解释：在假设检验问题中有一个非常基本的问题。每一个假设检验都有一个原假设和一个备择假设。假定原假设是事实，那么备择假设就有了一个可以在原假设为真的前提下的发生概率，如果这个发生概率非常小，但是实际上却发生了，我们就拒绝原假设。并把这个发生概率叫做p值。

#### 第二节 信息论与复杂度的概念界定

这篇文章中运用的信息度量方法主要来自于信息论，当要以不同的概率在几个选项中选择出来一个结果的时候，假设不同的概率是 ，那么需要人为决定的所有选择的信息量是

Entropy 代表理清这件事情需要的信息量

代表第i个选项的概率。

# 第三章

# R语言的p值计算案例

## 第一节

使用非参数方法计算P值的基本原理

使用二项分布来计算p值 的原理。在我们假定一个数组的中位数是x的时候，如果假设为真，小于x的数应该占了整个数组的一半左右，设整个数组的数的数量为n，那么小于x的数量应该为n/2左右，实际上我们得到的小于x的数量为a，如果a远远小于（或者远远大于）n/2那么我们就有理由怀疑原来的假设（x是这组数的中位数）是错的。下面的两个案例同理，用这样的原理来构造p值的计算公式。

数组的数的数量 n

如果 很小，比如小于0.05，（在要求显著性水平为0.05的情况下）

或者 很大，比如大于0.95 （在要求显著性水平为0.05的情况下）

那么我们拒绝原假设，认为中位数不是x

### R语言的例子1 判断组函数的中位数是否是一个数

x=c(274,279,290,326,329,341,378,405,436,500, 515,541,558,566,618,708,760,867,868,869, 888,915,932,942,960,975,976,1014,1025,1095, 1118,1166,1193,1194,1243,1277,1304,1327,1343,1398, 1407,1409,1417,1467,1477,1512,1530,1623,1710,1921)

#x 代表一个有多个数值的样本

#检验中位数是否是1200   
 n=length(x) # 计算一共有多少数据   
 median(x)  #计算实际的这一组数据的中位数的数值  
p0=0.5 q0=1200 alpha=0.05  #初始化参数，0假设的中位数， 以及显著性水平0.05  
sfu=sum(x<=q0) #计数所有小于0假设的中位数的数量  
low=qbinom(alpha/2,n,p0) #计算给定显著性水平置信区间的下分位点  
upp=qbinom(1-alpha/2,n,p0) #计算给定显著性水平置信区间的上分位点  
p=2\*min(pbinom(sfu,n,p0), 1-pbinom(sfu-1,n,p0)) #计算p值

说明： 给出一组数据，检验中位数是否是某一个数，解出p值得到结论。 这是一个最简单的例子，也是一个最为常用的例子， 我们来检查一下计算p值的几个环节，以及可能出错的部分。

### 说明 需要人为判断的环节，以及复杂度分析

sfu=sum(x<=q0) #计数所有小于0假设的中位数的数量  
 因为不同人的习惯不同，这个地方会涉及到有一部分人默认计数了大于假定中位数的数量，他的代码可能是这样的。 sfu=sum(x>=q0) 也就是任何审阅代码的人必须切实的理解这一部分， 以检查后来的计算是否出现了稿件的作者将第一种sfu 当成第二种sfu进行计算

p=2\*min(pbinom(sfu,n,p0), 1-pbinom(sfu-1,n,p0)) 计算p值 这里有两个需要判断的环节，一个是前面的默认计数对象sfu， 如果前面的计数sfu是小于假定中位数的数量，那么就应该使用上面的代码，否者应该使用下面的代码 p=2\*min(1-pbinom(sfu,n,p0), pbinom(sfu-1,n,p0)) 这个问题的真正简单之处是，即便作者搞错了前后参数的对应关系， 2\*min() 这个取两个数中较小值的函数也能够防止出错， 但是对于审核的人仍然意味着需要搞清每一个参数的真实意义。  
以及需要放在论文中的位置

总结： 这个计算一共需要做统计计算的人搞清楚两处参数的位置，即便搞错了，也能得到正确的结果，  
对于作者来只需要按着流程使用，不需做反复的检查，0bit的复杂度  
对于统计审核的人容易搞反的地方有两处，需要检查的地方有两个地方，检查的复杂度是2bit

## 第二节

### R语言中的例子2 判断一个函数的0.25分位点是不是一个数值

x=c(274,279,290,326,329,341,378,405,436,500, 515,541,558,566,618,708,760,867,868,869, 888,915,932,942,960,975,976,1014,1025,1095, 1118,1166,1193,1194,1243,1277,1304,1327,1343,1398, 1407,1409,1417,1467,1477,1512,1530,1623,1710,1921)

#检验0.25分位数是否是1000   
 n=length(x) #计算一共有多少数据   
   
 p0=0.25 q0=1000 alpha=0.05  #初始化参数，0假设的中位数， 以及显著性水平0.05  
sfu=sum(x<=q0) #计数所有小于0假设的中位数的数量  
   
  p=2\*min(pbinom(sfu,n,p0), 1-pbinom(sfu-1,n,p0)) 计算p值

### 说明

判断一个函数的0.25分位点是不是一个数值

需要人为判断的环节

sfu=sum(x<=q0) #计数所有小于0.25分位数的数量  
 因为不同人的习惯不同，这个地方会涉及到有一部分人默认计数了大于假定中位数的数量，他的代码可能是这样的。 sfu=sum(x>=q0) 也就是任何审阅代码的人必须切实的理解这一部分， 来检查后来的计算是否出现了稿件的作者是不是将第一种sfu 当成第二种sfu进行计算。

p=2\*min(pbinom(sfu,n,p0), 1-pbinom(sfu-1,n,p0)) #计算p值 这里有两个需要判断的环节，一个是前面的默认计数对象sfu， 如果前面的计数sfu是小于假定0.25分位数的数量，那么就应该使用上面的代码，否者应该使用下面的代码:

p=2\*min(1-pbinom(sfu,n,p0), pbinom(sfu-1,n,p0)) 这个案例并没有类似于上面的检验中位数时的防止出错机制，如果弄错了顺序，就会真正的出错 但是对于审核的人意味着需要搞清每一个参数的真实意义,以及前后的承接关系。

总结： 这个计算一共需要做统计计算的人搞清楚两处参数的位置，如果搞错了参数的位置，结果必然是出错的，对于作者来需要按着流程计算p值，有两个地方需要检查，2bit的复杂度。   
对于统计审核的人容易搞反的地方有两处，需要检查的地方有两个地方，检查的复杂度是2bit

结论：需要人为判断的环节和案例1类似，不过这个案例没有防止出错的机制，如果作者搞错了，就真的搞错了。也就是对于作者来说复杂度提升到了2bit， 对于审核人员来说，仍然维持2bit。

## 第三节

### R语言中的例子3 判断一个函数的0.25分位点是不是小于一个给定的数值多少

x=c(274,279,290,326,329,341,378,405,436,500, 515,541,558,566,618,708,760,867,868,869, 888,915,932,942,960,975,976,1014,1025,1095, 1118,1166,1193,1194,1243,1277,1304,1327,1343,1398, 1407,1409,1417,1467,1477,1512,1530,1623,1710,1921)

#原假设： 0.25分位数就是1000 #备择假设0.25分位数小于1000   
 n=length(x) 计算一共有多少数据   
   
 p0=0.25 q0=1000 alpha=0.05  #初始化参数，0假设的中位数， 以及显著性水平0.05  
sfu=sum(x<=q0) #计数所有小于0假设的中位数的数量  
p = 1-pbinom(sfu,n,p0)  #计算p值

### 说明

判断一个函数的0.25分位点是不是小于一个给定的数值多少

这个案例比较复杂，可以通过逆推法来决定最后的p = 1-pbinom(sfu,n,p0)是如何确认的

p值计算的过程

首先p值一定是过小拒绝原假设，

我们构造一个小的p值，

因为拒绝原假设会导致一个必然很小的p值，

拒绝原假设的情况是 实际0.25分位数小于1000

实际0.25分位数小于1000，也就会造成实际小于1000的数量比较多

根据分布函数的规律，应该用1-pbinom(sfu,n,p0)得到一个小值，

所以 p 值的计算应该取 1-pbinom(sfu,n,p0)

整个过程一共有8个思考步骤，中间没有任何防止出错的机制，和第一个例子和第二个例子不同，简单的改了一下问题要求的内容，需要人为进行检查是否正确的环节变得空前的多，而且整个问题彻底失去了防止出错的机制。 整个个问题的复杂程度是8bit

总结：需要人为判断的环节 和上面类似，不过这个案例没有防止出错的机制，如果作者搞错了，就真的搞错了  
   
也就是对于作者来说复杂度提升到了8bit， 对于审核人员来说，同样是8bit

# 第四章 复杂度增加导致p值被错误计算的机理

人为检查的环节过多，会造成p值在使用过程中耗时比较长，不过也因此造成了另一个更为严重的问题， 采用p值的研究即便有错误，检查起来也难如登天。而通过几个论文审核人员进行统计审核的排错能力也会大大下降。另外的一个后果是，因为p值代表的东西实际上需要很复杂计算过程，研究人员会形成只看p值，而不理解p值本身的含义。错看双边p值和单边p值都是非常常见的状况，而两者实际上数值差了一倍，也就导致了p值失去了真正的效力。

假定一篇论文有5个的逻辑错误的地方，而这五个逻辑错误的地方又相互关联，前后承接，这样一共要在32中情况中找出来一个正确的情况，即便有三个人同时参与这篇论文的统计审核，每个人犯错的概率30%，能够保证这篇论文是对的概率也不过1-.3^3 =0.973， 0.027的概率把事实搞成了另外一种情况，这也就意味着， 即便是三个人检查，也有0.027\*32 = 0.864的概率是另外的一种结果， 换句话说，因为p值的计算没有一个能够自动纠错的机制，统计审核的成本变得高的惊人 在三个人检查的情况下，依然有相当大的犯错概率，而现实中是不可能有这么多人对一篇论文进行检查的。基于p值的结果也就大打折扣。

假定一个人对每个bit信息的检查正确率是95%

假定人为检查一个统计环节需要5分钟，

假定统计审稿人本身不需要额外的信息理解这一篇论文，比如如果统计结果是医学的，数学的统计审稿人不需要医学知识来理解这篇医学文章。

那么可以计算得出下面的表格。



上表是在不同信息量下p值计算情况的简单说明

如果预处理方法很多，每个预处理的概率是 则多一个环节的信息复杂度增加的量为：

比如，我们拿到数据之后有四种预处理方式，每种处理的概率是1/4

那么需要解决这个问题的信息量是 4\* 1/4 \*-log(1/4) = 2bit

除去本文中的简单决策之外，更为复杂的决策，也可以通过转换成信息量的方式进行复杂度度量，从而讨论一个统计问题的构建耗时，和审核耗时。

# 论文的总结

上面的例子是p值计算中最为简单的例子，即便是最简单的p值的统计可靠性，也很依赖人为检查，这个时候因为审阅论文的人对论文本身的不熟悉的。 导致从头到尾进行统计审核，审核的时间会比写论文的时间还要长，如果进行了半年统计审核，到头来发现这个论文根本就站不住脚。也就是会面临大量的沉没成本

我们仅仅拥有p值是如何计算的知识是不够的，在绝大多数的关于p值使用过程中， 都会给出正确的使用方式， 但是在正确的道路上前行需要检查众多的参数对应关系，导致的出错概率很大。而大量的预处理，附带结果是基于p值的统计结果可靠性大大下降，一些网站已经开始禁止使用p值，因为p值中的一些不清楚的逻辑，会让同样的一个事实计算出来的p值相差好几倍。这也就让一些论文网站严重的下调p值的接受值，比如原来需要0.05的显著性水平，结果下调到了0.001，这样就会直接拒绝大量研究的发表。   
带来的影响  
 对于同样的数据，不同的计算方法会对p值有很大的浮动。所以对于绝大多数复杂的情况，P值并不可信。

# 结论和展望

## 第一节 研究结论

对于比较复杂或者步骤比较多的显著性计算，p值只有很小的概率是对的，这种情况不管统计人员对自己的结论多自信，应尽量采用更多的指标来衡量统计结果。比如公开数据和计算P值的方法。

如果情况比较简单，则推荐使用现成的统计软件工具包的工具来计算p值，并遵从ASA给出来的使用P值的几个准则。

## 第二节 难点与不足

因为目前用信息论中的方法衡量数学方法的复杂度是一个很少见的课题，所以，并没有很多参考来源，究竟是不是可以使用这样的方式度量，也没有一个普遍公认的结论。 这篇文章差不多是第一篇。

## 第三节 展望

在传统的统计学中，有大量的方法论，也有大量的统计学操作。 但是对于结果的不同处理方式会导致不同的结论，如果想要处理更为复杂的数据，就需要更为复杂的方法，而当统计学方法复杂到一定的成都的时候，就会完全失去可靠性。笔者估计在未来，人类需要面对更为复杂的情况的时候，统计学方法会逐步走向衰落，（实际上，目前世界的绝大多数信息都掌握在少数公司或者金融机构中，而他们已经在采用更先进的方法论来看待问题了）并在2025年左右完全失去市场。会让步于具有更高的容错性的，前后承接更强的神经网络方法，或者其他的方法取代。但是在目前，统计学方法，以及其他的数据处理方法还会有很大的市场份额。

# 参考文献

1. The ASA s Statement on p Values Context Process and Purpose
2. <https://www.r-statistics.com/2016/03/its-not-the-p-values-fault-reflections-on-the-recent-asa-statement/>

[3]《非参数统计》吴喜之 赵博娟

[4] <https://en.wikipedia.org/wiki/Information_theory> 信息论的链接

# 致谢

感谢我的指导老师，李启才给我提供的关于论文的指导和帮助