



華東師範大學

EAST CHINA NORMAL UNIVERSITY

教育实证研究中量化分析新方法

顾昕

华东师范大学教育心理学系

2018年10月27日

第四届全国教育实证研究论坛



概要

- 传统量化分析统计方法的缺点与不足
- 贝叶斯统计方法的优点
- 实证研究实例：青少年反社会行为与家庭环境的关系
- 贝叶斯统计方法的应用与发展现状及热点问题

第四届全国教育实证研究论坛



实证研究的特征

量化分析是教育实证研究的基本方法，是促进包括教育学在内的社会科学取得突破性进展的关键。

袁振国教授在《实证研究是教育学走向科学的必要途径》一文中将实证研究的基本特征与要求归纳为如下四点：

- 第一是客观，以确凿的事实和证据为基础。
- 第二是量化，努力获得对事物特征和变化的“度”的把握。
- 第三是有定论，有确切的发现和结论。
- 第四是可检验，使用共同方法、工具，可以获得相同的结果。



量化研究基本步骤

统计推断是量化分析中至关重要的一环。传统的统计推断方法能否满足实证研究的特征与要求，表达实证研究的精神？

量化研究基本步骤：

- 研究理论
- 提出假设
- 收集数据
- 统计分析
 - 统计模型
 - 统计推断--假设检验
- 报告结论

在统计推断中，研究人员通常使用基于 p 值的零假设显著性检验，来推断总体间的差异 ($p < 0.05$)。但是，该方法在表达实证研究的四个特征与要求时是存在缺陷与不足的。



传统假设检验

研究理论：母亲受教育程度对孩子学业成绩有显著影响

| | 母亲受教育程度 | | |
|------|-----------|-----------|-----------|
| | 初中 | 高中 | 本科 |
| 学业成绩 | $\mu_{初}$ | $\mu_{高}$ | $\mu_{本}$ |

根据研究理论，构建零假设与备择假设

$$H_0: \mu_{初} = \mu_{高} = \mu_{本}$$

$$H_a: \mu_{初}, \mu_{高}, \mu_{本}$$

收集数据，建立统计模型，计算 p 值。如果 $p < 0.05$ ，则拒绝零假设 H_0 ，得出母亲受教育程度对孩子学业成绩有显著影响的结论。

该统计分析过程与结论是否满足教育实证研究的特征与要求？



p值的缺点与不足

1. 实证研究特征一：以确凿的事实和证据为基础

- 基于 p 值的统计推断所得到的结论，能否为教育研究提供“确凿”的证据？

如上面的例子，当 $p < 0.05$ 时拒绝零假设，得到显著性差异的结论。

- 为什么是 $p < 0.05$ ，不是 0.06 或 0.04 ？有无严格的统计理论支持？
God loves the 0.06 as much as the 0.05 (Rosnow and Rosenthal, 1989)
(上帝并不偏爱0.05)

- 依据 $p < 0.05$ 的统计推断是草率的 (Ioannides, 2005)
更为严格的标准 $p < 0.01$ 或者 $p < 0.005$? (Benjamin et al., 2017)

$p < ?$ 可以得到“确凿”的证据？并无定论。



p值的缺点与不足

2. 实证研究特征二：获得对事物特征和变化的“度”的把握

- p 值不能衡量数据支持假设的程度

比如在上面的例子中，我们通过收集数据，建立统计模型（方差分析），计算得到 $p = .04 < .05$ ，得出不同受教育程度的母亲，其孩子的课业成绩存在显著性差异的结论。

但是具体存在多少差异？即差异的程度是多少？并不可知。

因此，基于p值的统计推断不能很好的把握事物特征（差异）的“度”。

第四届全国教育实证研究论坛



p值的缺点与不足

3. 实证研究特征三：有确切的发现和结论

- 首先，基于p值的显著性检验不能同时比较多个研究理论（假设）。

$$H_0: \mu_{\text{初}} = \mu_{\text{高}} = \mu_{\text{本}}$$

$$H_a: \mu_{\text{初}}, \mu_{\text{高}}, \mu_{\text{本}}$$

比如，母亲初中高中文化水平，孩子学业成绩无差异，但与母亲为本科文化水平的有差异。

$$H_{a1}: \mu_{\text{初}} = \mu_{\text{高}}, \mu_{\text{本}} \quad \checkmark$$

$$H_{a2}: \mu_{\text{初}}, \mu_{\text{高}} = \mu_{\text{本}}$$

因此，传统零假设检验在部分实证研究问题中，获得确切的发现和结论是有困难的。



p值的缺点与不足

3. 实证研究特征三：有确切的发现和结论

- 再者，基于p值的显著性检验通常不能准确检验研究理论，仅能够用来拒绝零假设。

如在前面的例子中，实际研究理论为母亲受教育程度越高，其孩子的课业成绩越好。而零假设并不能直接表达这一研究理论。

$$~~H_0: \mu_{初} = \mu_{高} = \mu_{本}~~$$

$$H_1: \mu_{初} < \mu_{高} < \mu_{本} ?$$

因此，传统假设检验方法较难直接获得确切的发现和结论。



p值的缺点与不足

- 4. 实证研究特征四：可检验，使用共同方法、工具，可获得相同的结果
 - 基于p值的统计推断存在严重的发表偏差（**publication bias**），显著性的实证研究结果更容易被期刊发表，所以研究人员都希望得到 $p < .05$ 。
 - 但是，当研究人员对发表的实证研究做可重复性实验时，发现很多实证研究的结果是不能被重复实现的。这违背了实证研究“可检验”的特征，即用共同方法、工具，无法获得相同的结果。

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration^{1,2}
+ See all authors and affiliations

Science 28 Aug 2015
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716



Science

Vol 349, Issue 6251
28 August 2015

Table of Contents
Print Table of Contents
Advertising (PDF)
Classified (PDF)
Masthead (PDF)

= 0.188), representing a substantial decline. Ninety-seven percent of original studies had significant results ($P < .05$). Thirty-six percent of replications had significant results; 47% of

97%的实证研究得到显著的结果
36%的重复性研究得到显著结果



p值的缺点与不足

由于以上缺点，近年科学界来对于 p 值的批评不断升温，也引起了教育与心理学领域相关机构与杂志的重视。

美国教育研究协会（AERA）

美国心理协会（APA）

谨慎使用p值作为统计推断的依据

Basic and Applied Social Psychology

拒绝接收使用p值的文章

August 2017

AERA, along with the American Psychological Association, is co-sponsoring the 2017 American Statistical Association (ASA) Symposium on Statistical Inference, October 11-13, in Bethesda, Md. The conference follows up on the ASA Statement on P-Values and Statistical Significance, which called for moving statistical analysis and evidence-based decision making beyond "bright line rules" toward a "post- $p < 0.05$ era."



ASA SYMPOSIUM ON
STATISTICAL
INFERENCE
OCTOBER 11-13, 2017 BETHESDA, MARYLAND

Scientific Method for the 21st Century: A World Beyond $p < 0.05$

nature

International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

Archive > Volume 519 > Issue 7541 > Research Highlights: Social Selection > Article

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans P values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015



p值的缺点与不足

是否有替代传统假设检验的新统计推断方法，满足实证研究的特征与要求？

- 信息假设—替代零假设
- 贝叶斯因子—替代p值

第四届全国教育实证研究论坛



信息假设

信息假设：准确反映教育研究理论，根据理论构建模型参数间的复杂关系。它既包括传统的零假设，也包含不等式假设，约等式假设，区间假设等。

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad (\text{等式假设, 零假设})$$

$$H_1 : \mu_1 > \mu_2 > \mu_3 \quad (\text{不等式假设, 次序假设})$$

$$H_2 : \mu_1 - \mu_2 \approx 0.5 \quad (\text{约等式假设})$$

$$H_3 : 0 < \mu_1 < 0.5 \quad (\text{区间假设})$$

。 。 。

根据具体研究理论构建并检验信息假设，能够在教育实证研究中，获得确切的发现和结论。

信息假设中的模型参数包括均值，相关系数，回归系数，因子负荷的等



例1：学习能力的比较



儿童



青少年



成年人

儿童学习能力强于青少年，强于成年人

$$H_1: \mu_{\text{儿童}} > \mu_{\text{青少年}} > \mu_{\text{成年}}$$



例2：影响学业成绩的因素



学习动机



课堂注意力



考试焦虑

学业成绩 = 学习动机 + 课堂注意力 + 考试焦虑

$H_2: \beta_{\text{学习动机}} > 0;$

$\beta_{\text{课堂注意力}} > 0;$

$\beta_{\text{考试焦虑}} < 0;$

$\beta_{\text{学习动机}} > \beta_{\text{课堂注意力}}$



例3：儿童抑郁症的发展

儿童抑郁症的发展

| | 测量指标 | | | |
|----|---------|---------|---------|---------|
| | 8岁 | 12岁 | 16岁 | 20岁 |
| 男孩 | μ_1 | μ_2 | μ_3 | μ_4 |
| 女孩 | μ_5 | μ_6 | μ_7 | μ_8 |



抑郁症的性别差异随着年龄的增长而逐步增大

$$H_3: \mu_5 - \mu_1 < \mu_6 - \mu_2 < \mu_7 - \mu_3 < \mu_8 - \mu_4$$



信息假设

信息假设通过构建准确描述研究理论的假设，帮助教育实证研究获得确切的发现和结论。

如何检验信息假设，以确凿的证据来支持或拒绝研究理论，获得对事物特征和变化的“度”的把握。

- p值 x
- 贝叶斯因子 ✓



贝叶斯因子

贝叶斯因子衡量数据支持假设的程度（统计理论支持）

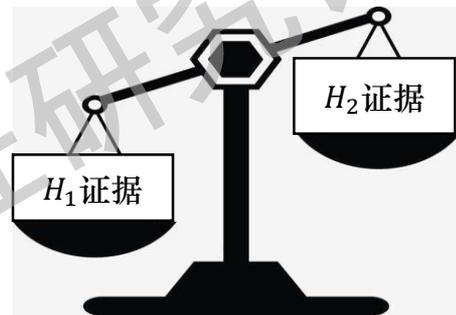
$$\text{贝叶斯因子BF} = \frac{\text{假设1的证据}}{\text{假设2的证据}} = 5$$

$$H_1: \mu_{\text{儿童}} > \mu_{\text{青少年}} > \mu_{\text{成年}}$$

$$H_2: \mu_{\text{儿童}}, \mu_{\text{青少年}}, \mu_{\text{成年}}$$

BF=5，则数据支持假设1的程度是假设2的5倍

相比于p值，贝叶斯因子把握了证据的“度”





贝叶斯因子

在实证研究中，贝叶斯因子是如何判断“确凿”的证据的？

$$H_1: \mu_{\text{儿童}} > \mu_{\text{青少年}} > \mu_{\text{成年}}$$

$$H_2: \mu_{\text{儿童}}, \mu_{\text{青少年}}, \mu_{\text{成年}}$$

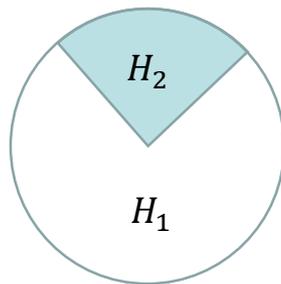
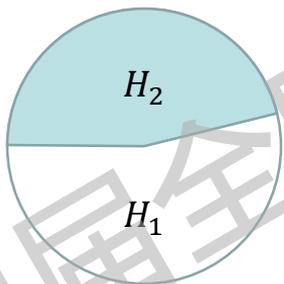
- $1 < BF_{12} < 3$, 数据支持假设1, 但并无确凿的证据。
- $3 < BF_{12} < 20$, 数据支持假设1, 且存在令人信服的证据
- $20 < BF_{12} < 150$, 数据支持假设1, 且存在强力的证据
- $BF_{12} > 150$, 数据支持假设1, 且存在非常强力的证据



贝叶斯因子

与p值不同，贝叶斯因子存在非决定性区间，即当贝叶斯因子在1附近时，数据并未明确告知我们哪个假设为真。如左图所示，未得到确凿证据。

而当贝叶斯因子等于3时，从右图可以看到，假设1得到的数据支持明显大于假设2





贝叶斯因子

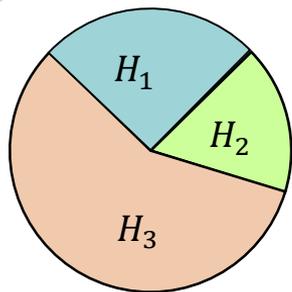
贝叶斯因子可以同时检验多个假设

$$H_1: \mu_{\text{儿童}} = \mu_{\text{青少年}} = \mu_{\text{成年}}$$

$$H_2: \mu_{\text{儿童}} < \mu_{\text{青少年}} < \mu_{\text{成年}}$$

$$H_3: \mu_{\text{儿童}} > \mu_{\text{青少年}} > \mu_{\text{成年}}$$

将贝叶斯因子转换为概率的形式。如下图可知假设3得到的数据支持最多。





实例

- 如何根据实际实证研究问题、研究理论，构建信息假设？
- 如何使用贝叶斯因子检验信息假设？
- 如何报告基于贝叶斯因子的量化分析结果？

Gu, X., Mulder, J., Dekovic, M., Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, 19, 511-527.



实例

青少年反社会行为与家庭环境的关系



家庭环境对儿童与青少年的成长有着非常重要的作用。家庭环境包括家长社会经济地位，婚姻状况，青少年与家长的关系等。

研究发现青少年与家长的关系是影响青少年每日行为的最重要因素。良好的家庭教育方式能够减少青少年的反社会行为。



研究理论

家长与青少年的关系包括

- 积极关系：家长对青少年的关心程度
- 消极关系：家长与青少年的分歧程度
- 自我披露：青少年告诉家长自己的行为 and 想法

青少年成长过程中，与同伴相处的时间逐渐增多

- 偏差同伴：同样有反社会行为的同伴数量

研究理论：

- 青少年反社会行为与三种家长与青少年的关系相关。
- 青少年自我披露在三种关系中对反社会行为的影响最大。
- 偏差同伴对反社会行为的影响大于家长与青少年的关系。





数据与测量

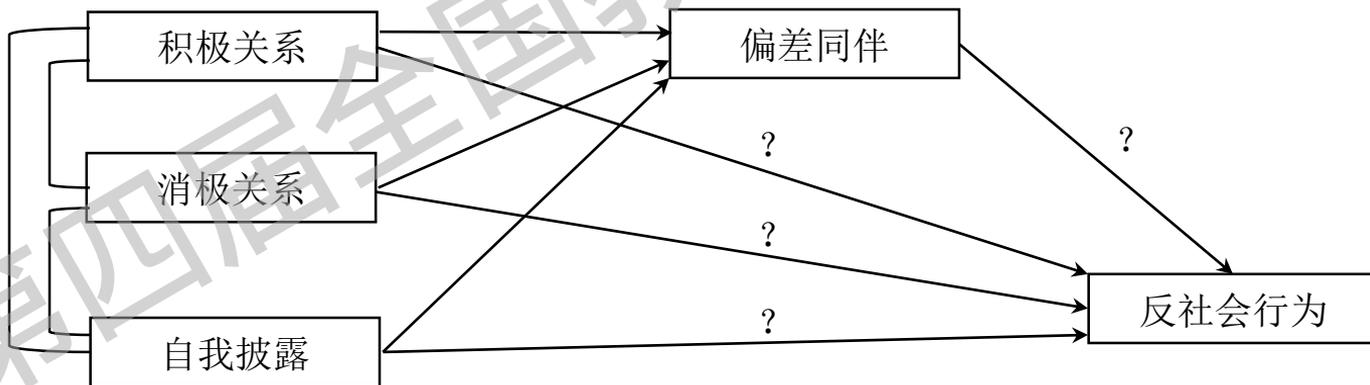
数据： 603名荷兰中学二年级的青少年（14-16岁），平均年龄为14.43，318女性，285男性。青少年问卷调查

- 反社会行为： 一年内反社会行为的频率
(1: 从不, 2: 一次, 3: 两到三次, 4: 四到十次, 5: 大于十次)
- 积极关系： 家长对青少年的关心程度
(1: 十分不关心, 2: 不关心, 3: 一般, 4: 关心, 5: 十分关心)
- 消极关系： 家长与青少年的分歧, 争吵程度
(1: 从不, 2: 很少, 3: 一般, 4: 很多, 5: 总是)
- 自我披露： 告诉家长自己的想法感受
(1: 从不, 2: 很少, 3: 很多; 4: 全部)
- 偏差同伴： 同样有反社会行为的同伴数量
(1: 无 —— 5: 所有)



统计描述与模型

| 变量 | 均值 | 方差 |
|-------|------|------|
| 积极关系 | 3.03 | 0.92 |
| 消极关系 | 2.00 | 0.70 |
| 自我披露 | 2.76 | 0.64 |
| 偏差同伴 | 1.71 | 0.68 |
| 反社会行为 | 1.44 | 0.56 |





传统方法

理论1：青少年反社会行为与三种家长与青少年的关系相关。

报告p值，青少年反社会行为

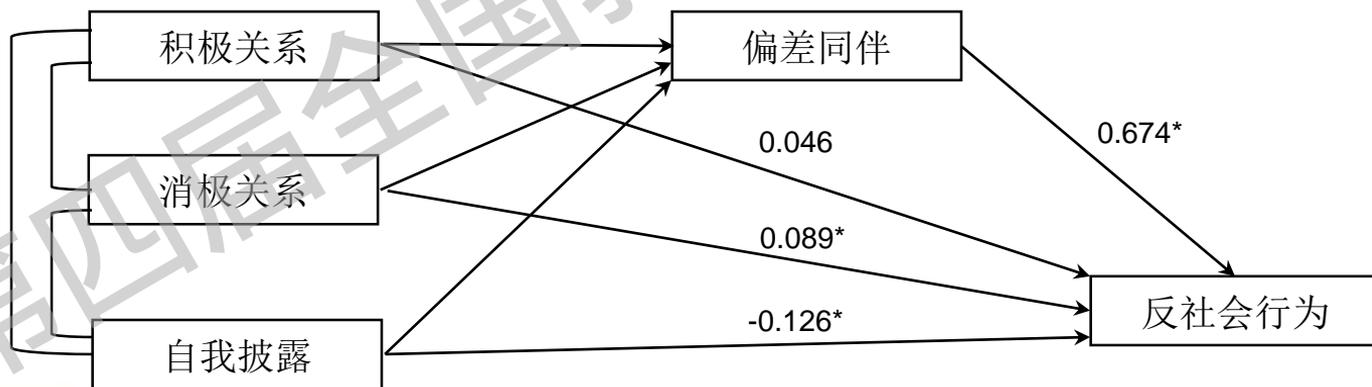
- 与家长的积极关系不相关 ($p = 0.156 > 0.05$)
- 与家长的消极关系相关 ($p = 0.003 < 0.05$)
- 与自我披露相关 ($p = 0.000 < 0.05$)

相关的具体程度？

MODEL RESULTS

| | | Estimate | S. E. | Est. /S. E. | Two-Tailed P-Value |
|-----|-----|----------|-------|-------------|--------------------|
| ZY1 | ON | | | | |
| | ZX1 | 0.082 | 0.044 | 1.839 | 0.066 |
| | ZX2 | -0.320 | 0.040 | -7.954 | 0.000 |
| | ZX3 | -0.201 | 0.044 | -4.592 | 0.000 |
| ZY2 | ON | | | | |
| | ZX1 | 0.046 | 0.032 | 1.418 | 0.156 |
| | ZX2 | 0.089 | 0.031 | 2.922 | 0.003 |
| | ZX3 | -0.126 | 0.032 | -3.911 | 0.000 |
| | ZY1 | 0.674 | 0.030 | 22.692 | 0.000 |

Mplus软件





贝叶斯方法

报告贝叶斯因子 BF ，青少年反社会行为

- 与家长的积极关系不相关 ($BF_{10} = 0.113 < 1$)
- 与消极关系相关的证据是不相关证据的2.79倍 ($BF_{10} = 2.79 < 3$, 证据不够确凿)
- 与自我披露相关的证据是不相关证据的84.1倍 ($BF_{10} = 84.1$, 证据确凿)

不仅判断是否相关，同时给出了相关的证据。

| | Estimate | SE | BF10 |
|------------|----------|-------|--------|
| positive | 0.046 | 0.032 | 0.113 |
| negative | 0.089 | 0.031 | 2.789 |
| disclosure | -0.126 | 0.032 | 84.077 |

R软件



实例

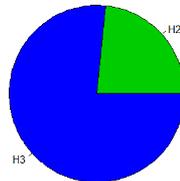
- 研究理论2: 自我披露在三种家长与青少年的关系中对反社会行为影响最大。
 H_2 : 自我披露 > {积极关系, 消极关系}
- 研究理论3: 偏差同伴对反社会行为的影响大于家长与青少年的关系。
 H_3 : 偏差同伴 > 自我披露 > {积极关系, 消极关系}

这些研究理论无法用零假设描述, 但构建不等式假设。

基于p值的传统方法不能检验不等式假设, 因此只考虑贝叶斯方法。通过软件计算得到假设 H_2 的贝叶斯因子为 $BF_2 = 6.16$, 假设 H_3 的贝叶斯因子为 $BF_3 = 29.2$ 。即有确凿证据表明自我披露在三种家长与青少年的关系中对青少年反社会行为影响最大, 而偏差同伴对反社会行为的影响大于家长与青少年的关系。

```
> test_result
```

| | fit | complexity | BF | PMPa | PMPb |
|----|-------|------------|--------|-------|-------|
| H2 | 0.789 | 0.378 | 6.159 | 0.235 | 0.211 |
| H3 | 0.794 | 0.117 | 29.190 | 0.765 | 0.688 |





p值vs贝叶斯因子

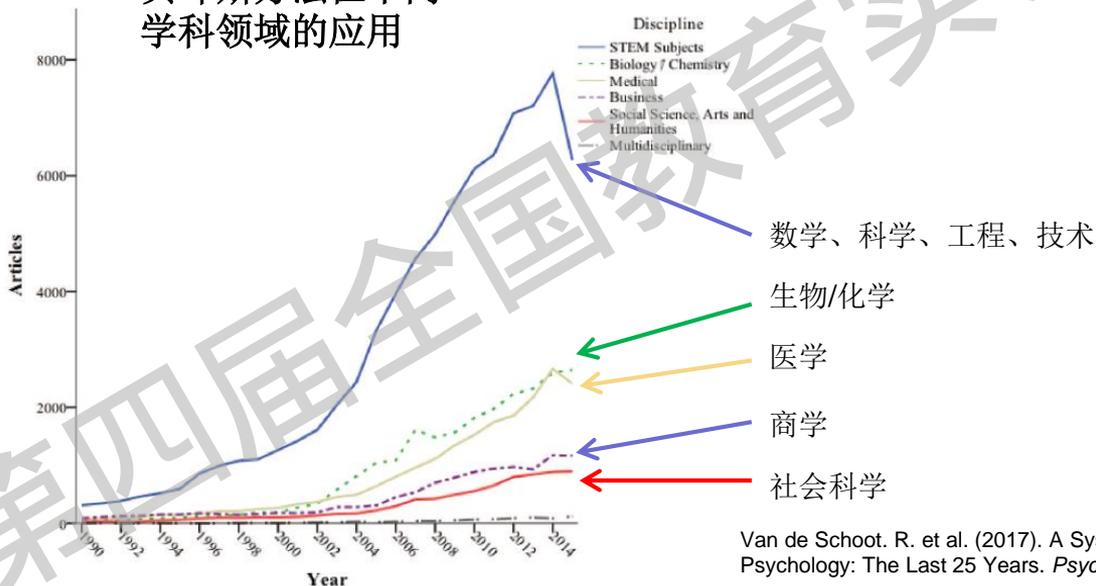
| | p值 | 贝叶斯因子 |
|----------|----|-------|
| 检验零假设 | √ | √ |
| 检验信息假设 | x | √ |
| 比较多个假设 | x | √ |
| 衡量数据支持程度 | x | √ |
| 结论可检验、重复 | x | ? |



贝叶斯方法的应用

尽管贝叶斯方法较传统p值方法有诸多优点，在社科领域的实证研究中，基于p值的统计推断仍然占据主导地位。

贝叶斯方法在不同学科领域的应用



Van de Schoot, R. et al. (2017). A Systematic Review of Bayesian Articles in Psychology: The Last 25 Years. *Psychological Methods*, 22, 217-239.



贝叶斯方法的应用

可能的原因：

1. 贝叶斯方法统计原理相对复杂，社科研究人员主要关注的是解决实质性的研究问题，往往不在意统计方法的优劣，因此更偏爱简单的统计方法。
2. 研究人员所学的统计方法是基于 p 值的统计推断，并实际应用与推广。学习与应用新统计方法，需要研究人员投入精力。
3. 研究人员担心使用新的统计方法将会偏离“标准”，即偏离多数人使用的传统方法。因此更愿意留在原有的、“安全的”的地方。
4. 几乎所有的统计软件都可以计算 p 值，而提供贝叶斯因子计算的软件相对较少。

但是以上困难正在被社科统计学者（教育统计、心理统计、社会统计）解决。



贝叶斯方法的应用

1. 近年来，教育与心理统计方向的研究人员发表了一系列文章，讲述如何在实证研究中使用贝叶斯统计方法，为教育和心理实证研究提供更好的统计分析范式。

Hoijtink, H., Gu, X., Mulder, J. & van Lissa, C. (in press). A Tutorial on Testing Hypotheses Using the Bayes Factor. *Psychological Methods*.

Gu, X., Mulder, J. & Hoijtink (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, 71, 229-261.

Konig, C. and van de Schoot, R. (2017). Bayesian statistics in educational research: a look at the current state of affairs. *Educational Review*, 70, 486-509

Levy, R. (2016). Advances in Bayesian modeling in educational research. *Educational Psychologist*. 51, 368-380.

2. 同时，国际上越来越多的大学为教育学与心理学专业，开设贝叶斯统计课程。如美国威斯康辛-麦迪逊大学，荷兰阿姆斯特丹大学，乌特勒支大学等。



贝叶斯方法的应用

3. 贝叶斯方法在实证研究中的应用逐年增长。

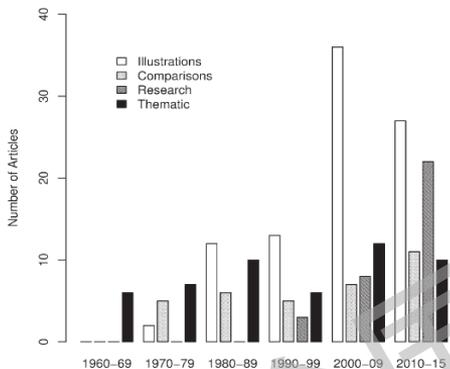
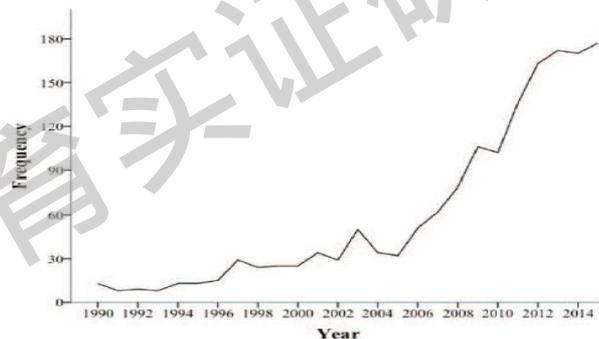


Figure 3. General prevalence of Bayesian statistics in educational research. Absolute frequencies of articles mentioning the term bayes*

教育学



A. Absolute growth of Psychology papers mentioning "Bayesian"

心理学

Konig, C. and van de Schoot, R. (2017). Bayesian statistics in educational research: a look at the current state of affairs : The Last 25 Years. *Educational Review*, 70, 486-509.

Van de Schoot, R. et al. (2017). A Systematic Review of Bayesian Articles in Psychology: The Last 25 Years. *Psychological Methods*, 22, 217-239.

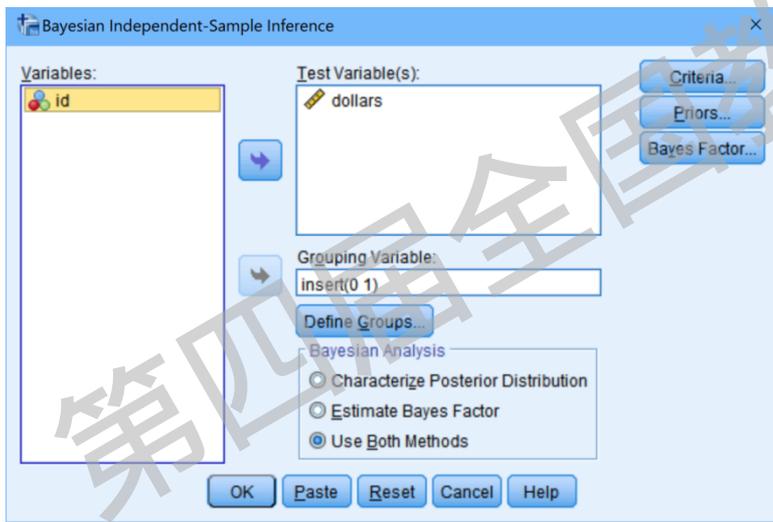


贝叶斯方法的应用

4. 支持贝叶斯方法的统计软件越来越多

➤ SPSS

SPSS 25版本提供贝叶斯统计分析，计算贝叶斯因子。包括 t 检验，相关性检验，线性回归分析，方差分析等。



但只能在简单模型下检验零假设！

Bayes Factor Independent Sample Test (Method = Rouder)^a

| | Mean Difference | Pooled Std. Error Difference | Bayes Factor ^b | t | df | Sig. (2-tailed) |
|------------------------------------|-----------------|------------------------------|---------------------------|-------|-----|-----------------|
| \$ spent during promotional period | 71.1110 | 31.45914 | 1.150 | 2.260 | 498 | 0.024 |



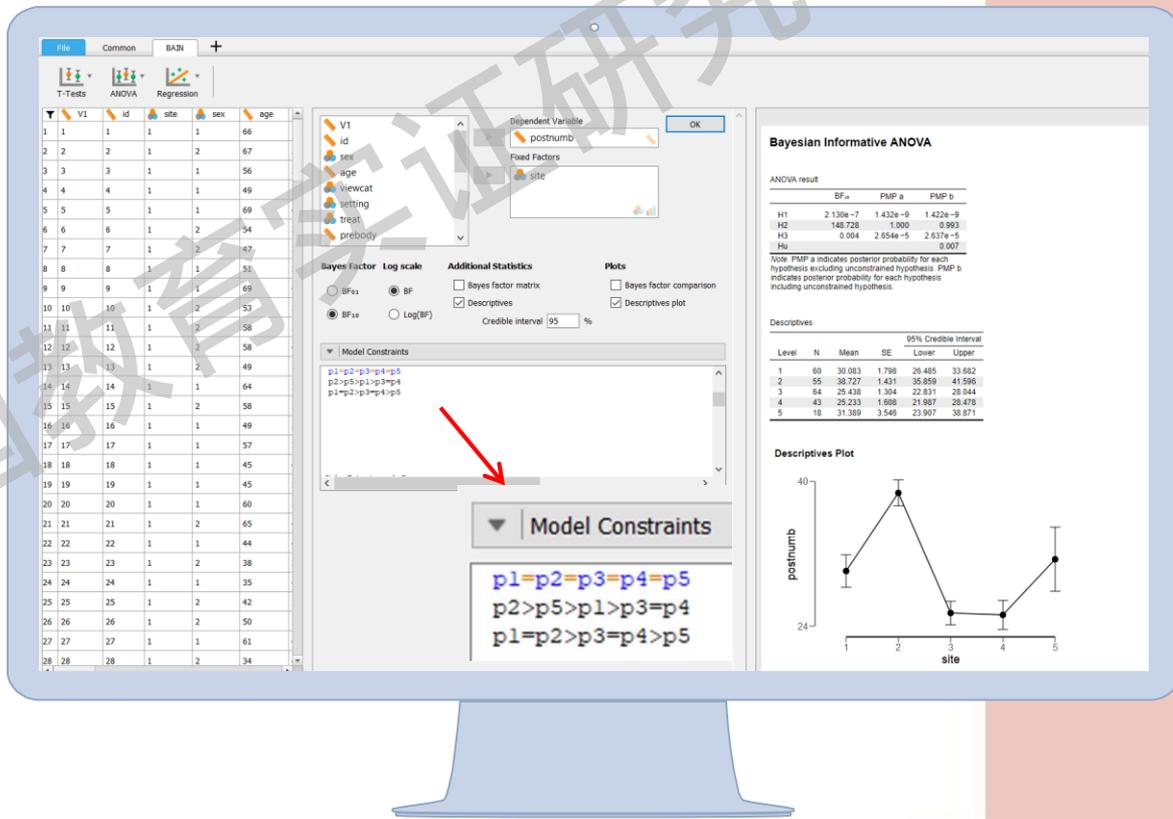
贝叶斯方法的应用

➤ JASP

JASP是一款类似于SPSS的图形界面操作统计软件

- 免费
- 直观的用户界面
- 完善的贝叶斯统计模块

JASP-Bain模块可以用来检验信息假设





发表偏差与可重复性议题

依据贝叶斯方法所得到的结论是否存在发表偏差，能否被可重复性实验所检验？

贝叶斯因子存在非决定性区间，及当 $\frac{1}{3} < BF < 3$ 时，不做研究理论被支持或不被支持的判断，从一定程度上降低了发表偏差与不可重复的问题。

但是依据贝叶斯因子 $BF > 3$ ，推断数据支持研究理论，且存在令人信服的证据。这一过程同样存在发表偏差与研究不可被重复的问题，即研究人员期望得到 $BF > 3$ 的结果。

可能尝试“修补”数据，如

- 去掉异常值？
- 去掉部分变量？
- 选择性报告结论？



发表偏差与可重复性议题

如何避免发表偏差与可重复性问题？

1. 不依据临界值做判断，即不使用 $BF > 3$ 作为临界值来判断数据支持假设与否。如果 $BF = 3.5$ ，则报告结论为：数据支持假设（研究理论）的程度是不支持的3.5倍。如母亲受教育程度对孩子学业成绩有影响的证据是无影响的3.5倍。
2. 研究预注册（preregistration）
Center for Open Science. <https://cos.io/rr/>



发表偏差与可重复性议题

2. 研究预注册

- 研究人员将其研究理论，实验设计，分析计划撰写成文，在收集数据、分析数据之前，将文章投稿。同行评审根据文章研究理论的重要程度、新颖程度，以及实验设计和分析计划的优劣，来决定文章是否能发表。如果支持发表，则研究人员再去收集数据，分析数据，报告结论。





发表偏差与可重复性议题

2. 研究预注册的优点

- 有效解决了发表偏差的问题，因为数据分析结果在同行评审过程中并不可知。
- 由于文章接收与否不依赖于数据分析，研究人员会报告更准确而不是“更漂亮”的结果。
- 目前越来越多的期刊鼓励这一评审过程，比如心理学权威期刊。

Psychological Science.

"Registered Reports eliminates the bias against negative results in publishing because the results are not known at the time of review."

-- Daniel Simons, Professor at University of Illinois, Urbana-Champaign, co-editor of Registered Replication Reports at Perspectives on Psychological Science, and incoming chief editor of Advances in Methods and Practices in Psychological Science

"Because the study is accepted in advance, the incentives for authors change from producing the most beautiful story to the most accurate one."

--Chris Chambers, Professor at Cardiff University, Section Editor for Registered Reports at Cortex, European Journal of Neuroscience and Royal Society Open Science, Chair of the Registered Reports Committee supported by the Center for Open Science



第四届全国教育实证研究论坛

谢谢!