# 假设检验的前世今生

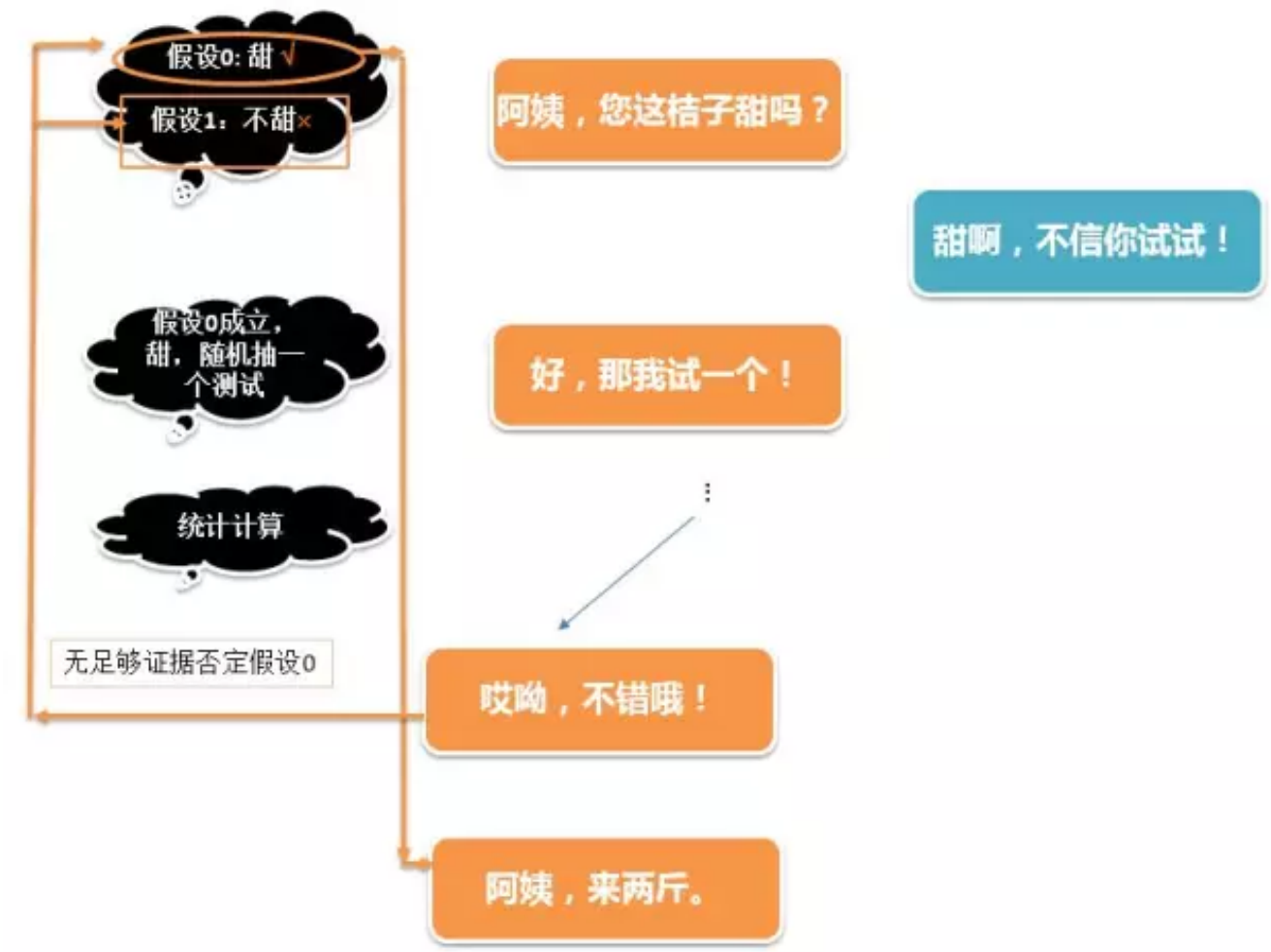
原创：谷鸿秋 统技思维 2016-01-19

> 这篇是上一篇 「统计？我懂个P！」的姊妹篇。

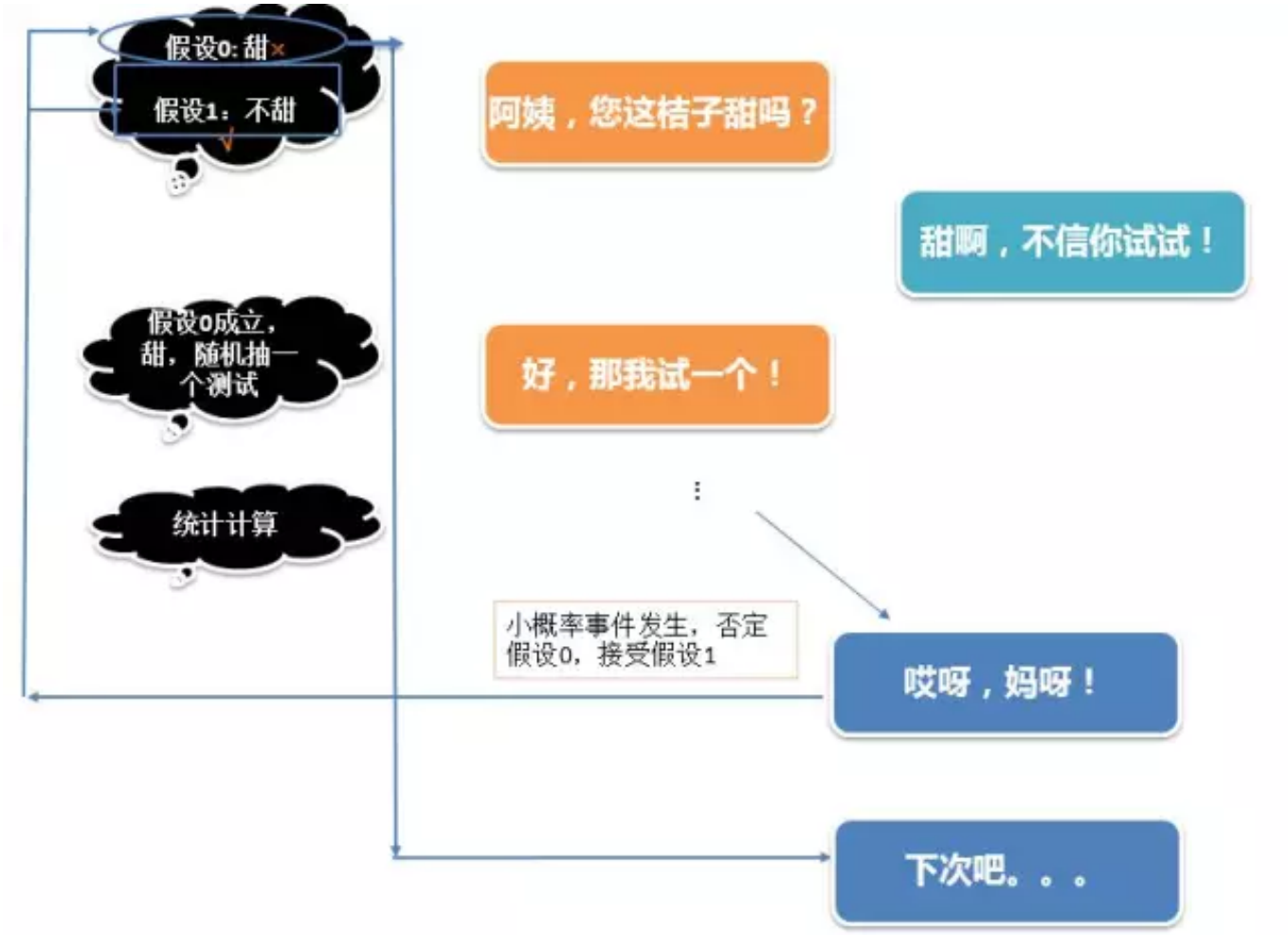
其实，「前世今生」系列的文章我已经看到过好几篇了，比如「正太分布的前世今生」、「Meta分析的前世今生」。不知为何，我个人也很喜欢「前世今生」这个词。今天呢，就聊一聊我知道的一点「假设检验的前世今生」吧。

假设检验是统计学里最重要、最基础的的概念，即便是不知道，不了解这个术语，与统计学毫不相干的人，在日常生活中，也不知不觉地应用了假设检验。比如，我们在街上水果摊闲逛买橘子。

甜的时候，我们的思维过程：




不甜的时候，我们的思维过程：



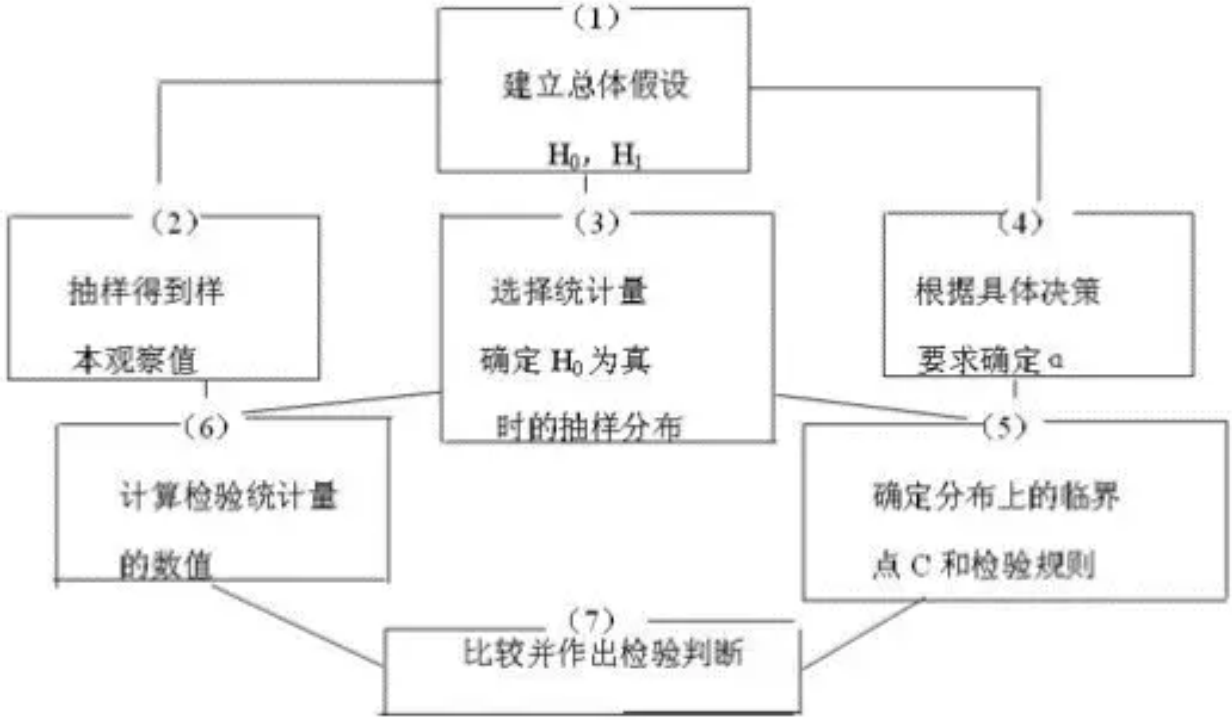

当然，以上只是个简单类比，不必细究。不过，相比一些翻译教材喜欢用老外的「法官定罪」的例子来说，这个场景应该更容易为国人所理解。

现行的假设检验，叫原假设显著性检验( Null Hypothesis Significance Testing，NHST)。其基本思路和框架在现行的统计教材中论述较多，在此仅简要概括：

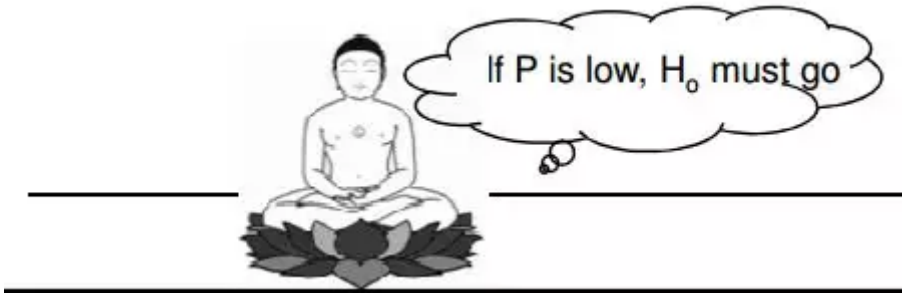
1. 建立假设，确定检验水平。假设包括两种，一种称为原假设、无效假设、零假设（Null Hypothesis，H0）；另一种称为备则假设（Alternative hypothesis, H1），H1是H0的对立面。原假设H0通常是「别担心，啥事也没有」，比如没有差异，没有疗效等。H1 则是「有情况，要留意啊」，比如有差异，有疗效。检验水平alpha，又称显著性水平，这个是预先规定游戏标杆，通常为0.05。

2. 计算检验统计量，计算P值。我们认为手头已有的数据是从H0 为真的总体中的一个抽样，但是这个可能性是多少？这需要计算评估。如何计算评估呢？我们可以计算检验统计量，不过不同的问题，计算的检验统计量不同，如Z值，t值，F值，X2值，这样岂不是比较乱？是的，所以把那些统计量统统对应到P值，统一用P值来解决。

3. 做出统计推断结论。比较P值及alpha值，如果 P<=alpha, 拒绝H0,差异显著，有统计学意义；反之，如果P>alpha, 不拒绝H0,差异不显著，无统计学意义。




不太想了解假设检验的具体流程和细节的，只要记住一条简单粗暴的黄金口诀：If P is low, H0 must go!




以上这一套流程，看起来好像是流畅统一的整体，然而，统计教材没有说明的是，这其实是一道大拌菜，是统计学家Karl Pearson的「拟合优度检验」，Ronald A Fisher的「显著性检验 」和 Jerzy Neyman，Egon Pearson的N-P「假设检验」的大杂烩。

故事的关键点大概是这样的：

## 1. Karl Pearson的「拟合优度检验」

部分文献以为P值是Fisher发明的，但其实最先提出P值的是Karl Pearson。Karl Pearson在其1900年的论文中提出了拟合优度的卡方检验，这其中就包括P值。但是给出了P值的在各种情形下的计算方法的却是Karl Pearson的死对头，Ronald A Fisher。应该说，Karl Pearson的提出了「P值 」，Ronald A Fisher 将「P值 」发扬光大。 而

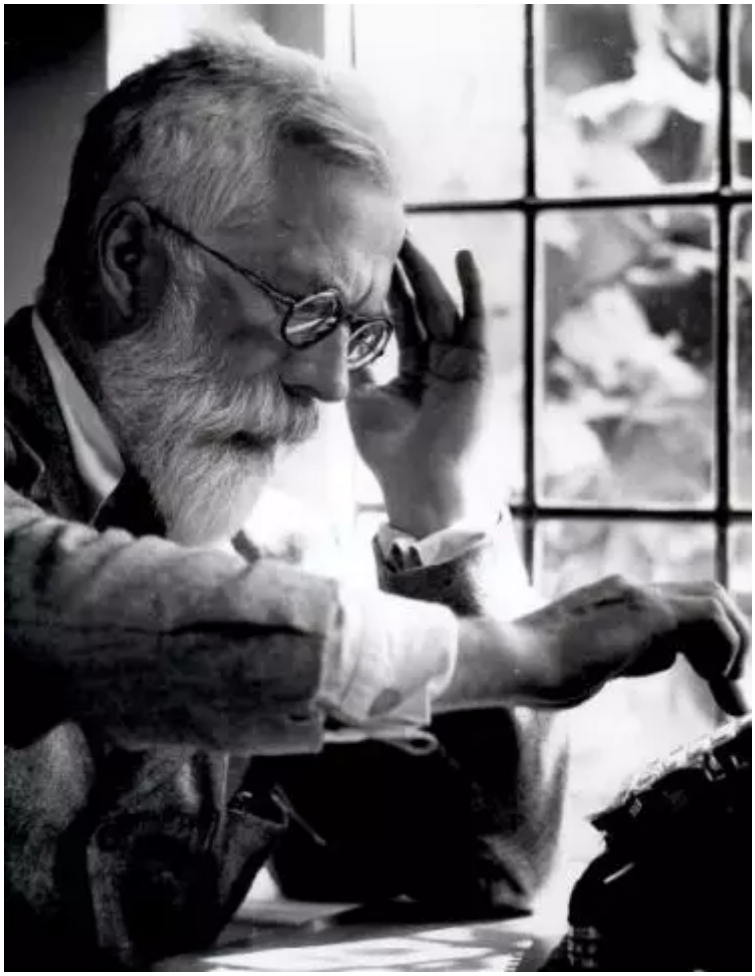
1925年Ronald A Fisher的 经典著作《Statistical Methods for Research Workers》腾空出世奠定了其现代统计学之父的威名。

**Karl Pearson**

## 2. Ronald A Fisher的 「显著性检验 」

1925年，Fisher提出了其显著性检验的思想。Fisher的显著性检验可大概概括为以下5个步骤：

1. 选择合适的检验，如卡方检验，t检验
2. 建立原假设H0
3. 假定H0的条件下计算理论的P值
4. 评估结果是否有统计学显著
5. 对结果的统计学显著性进行解释

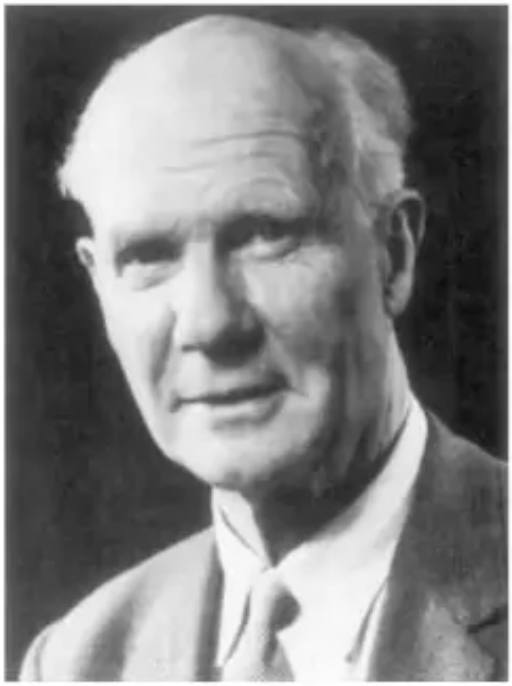


**Ronald A Fisher**

### 3. N-P的「假设检验」

1928年，Jerzy Neyman和Karl Pearson 的儿子 Egon Pearson提出了「假设检验」，「假设检验」思想可大概概括为以下8个步骤：

1. 设立人群中期望的效应值
2. 选择合适的检验
3. 建立主假设Hm
4. 建立备则假设Ha
5. 计算为达到良好的把握度所需的样本量
6. 计算检验的临界值，确定拒绝域
7. 计算研究的检验值（老实说，这条我也没理解）
8. 做出支持Hm或者Ha的决策




Egon Pearson，Jerzy Neyman

简单来看，Ronald A Fisher的 「显著性检验 」是没有备则假设的，而N-P的「假设检验」不仅有备则假设，还有一个主假设Hm(与H0类似)，不仅如此，N-P的「假设检验」还提出了效应值、把握度，I类、II类错误的概念，且采用拒绝域而非P值来做决策。

除了以上形式上的差别，Ronald A Fisher的 「显著性检验 」与N-P的「假设检验」在深层次的统计哲学上也不同。

- Fisher的统计模型的方法论基础是假想无限总体，现有资料可视为是从中抽取的一个随机样本。而N-P则是假想无限抽样。N-P 「假设检验」 的要旨为在限制第一类错误的概率不超过显著性水平 α 的条件下， 谋求第二类错误的概率最小化。虽不期望知晓每个独立的假设是真是假， 但仍可研究指导我们与之相关行为的准则， 以便保证在长远意义上不至错得太多。
- Fisher认为统计学的功用是“归纳推论”( inductive inference)， 而不是做“归纳行动”( inductive behavior)；统计学应当止于归纳结论， 而不涉足行动判断。显著性检验不能给出针对现实的判断， 而只能改变研究者对事实的态度。而在 N-P 看来， 没有任何一种统计推论思想能够不涉及决策过程。他们直接绕过假设检验作为科学推论的适合性的讨论， 而将它作为一种决策方法， 在先行给出决策前提( 控制第一类错误、 然后追求功效最大化) 的前提下， 进行数学上的最优化论证( 错误率最低) 。这种思维方式对实际研究者显然是很有“实际优势” 的， 因为这正符合了他们使用假设检验的最初目的和最终期待

## 4. 原假设显著性检验，NHST

1940年，Lindquist首次对Ronald A Fisher的 「显著性检验 」和N-P的「假设检验」进行了糅合， 提出了原假设显著性检验（Null Hypothesis Significance Testing，NHST）。

NHST 的基本杂合方式是:

1. 采用 N-P 的原假设对备择假设 的假设形式( H0 vs H1)， 而备择假设却是 Fisher 没有使用并且一直反对引入的
2. 同时采用 P值( Fisher 的判断依据) 和拒绝域法( N-P 的判断依据)， 认为两者的判定效果是等价的， 但 Fisher 本人却极其反对拒绝域法， 而 N-P 则并不强调P值的作用
3. 把检验功效 和两类错误作为 NHST 的内在内容加以介绍， 而不提及这只是 N-P 的观点， Fisher本人是反对这些概念的

至此，这就是我们统计教科书里看到的假设检验了。NHST自其诞生以来就饱受质疑和批判，后世的统计学家也一直在呼吁用置信区间，贝叶斯统计来取代NHSTH这种统计推论方式。更多批判NHST的文章和更深层的讨论，好像已经超出我的能力范围了。

这篇写着写着就写岔气了，希望不要掉粉。哎，做饭去了。

P.S. 回复「StatHis」，送给大家一点福利：**统计发展历史年表。**


相关阅读：

1. 统计？我懂个P！
2. 吕小康. Fisher与Neyman-Pearson的分歧与心理统计中的假设检验争议[J]. 心理科学, 2012.
3. Perezgonzalez J D. Fisher, Neyman-Pearson or NHST? A Tutorial for Teaching Data Testing[J]. Frontiers in Psychology, 2015, 6:223.
4. David Jean B, Jolles B M, Rapha?L P. P value and the theory of hypothesis testing: an explanation for new researchers.[J]. Clinical Orthopaedics &

Related Research, 2010, 468(468):885-92.
5. Fienberg S E, Tanur J M. Reconsidering The Fundamental Contributions Of Fisher And Neyman On Experimentation And Sampling[J]. International Statistical Review, 1996, 64(3):237-253.

---