20200302 在线研讨

Q：老师，听录音的时候有个问题是，有时候不知道在讲ppt的哪个位置，还有板书也看不到。

A：这是我给大家设置的一个训练，通过自己的思索整合，得到的内容更深入一些。如有不理解的问题我们现在就可以讨论。板书通常在PPT的最后一页有拍照。

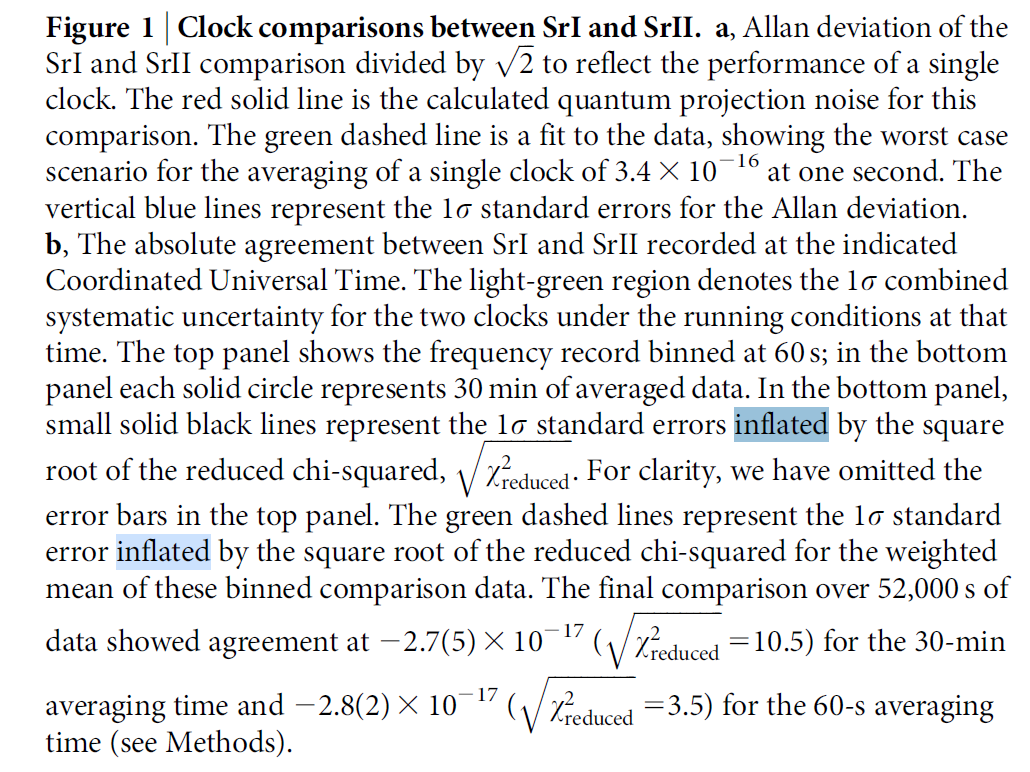
可以先看图，把PPT课件过一遍后，有不懂的再听录音。其实PPT是自给自足的，但听录音可以收获更多。

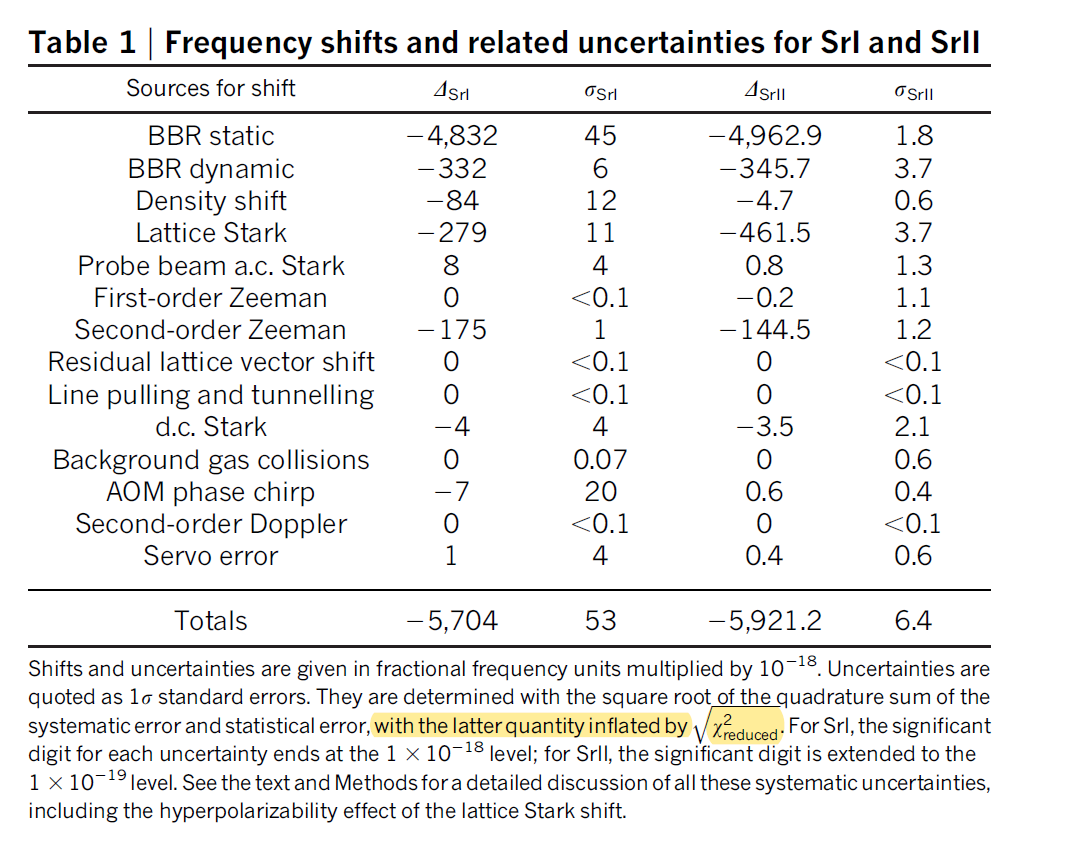
Q：老师，我想问下 乘以reduced chi square这个操作，我理解就是当我们强烈的相信一个模型是正确的，但数据的单点errbar又确实太小以至于给出了非常糟糕（不确定度太大）的拟合时，去把errbar进行一个放大的操作。这个理解对吗？它背后的统计学原理是什么？为什么可以这样做，以及chi square为什么这么定义？谢谢。

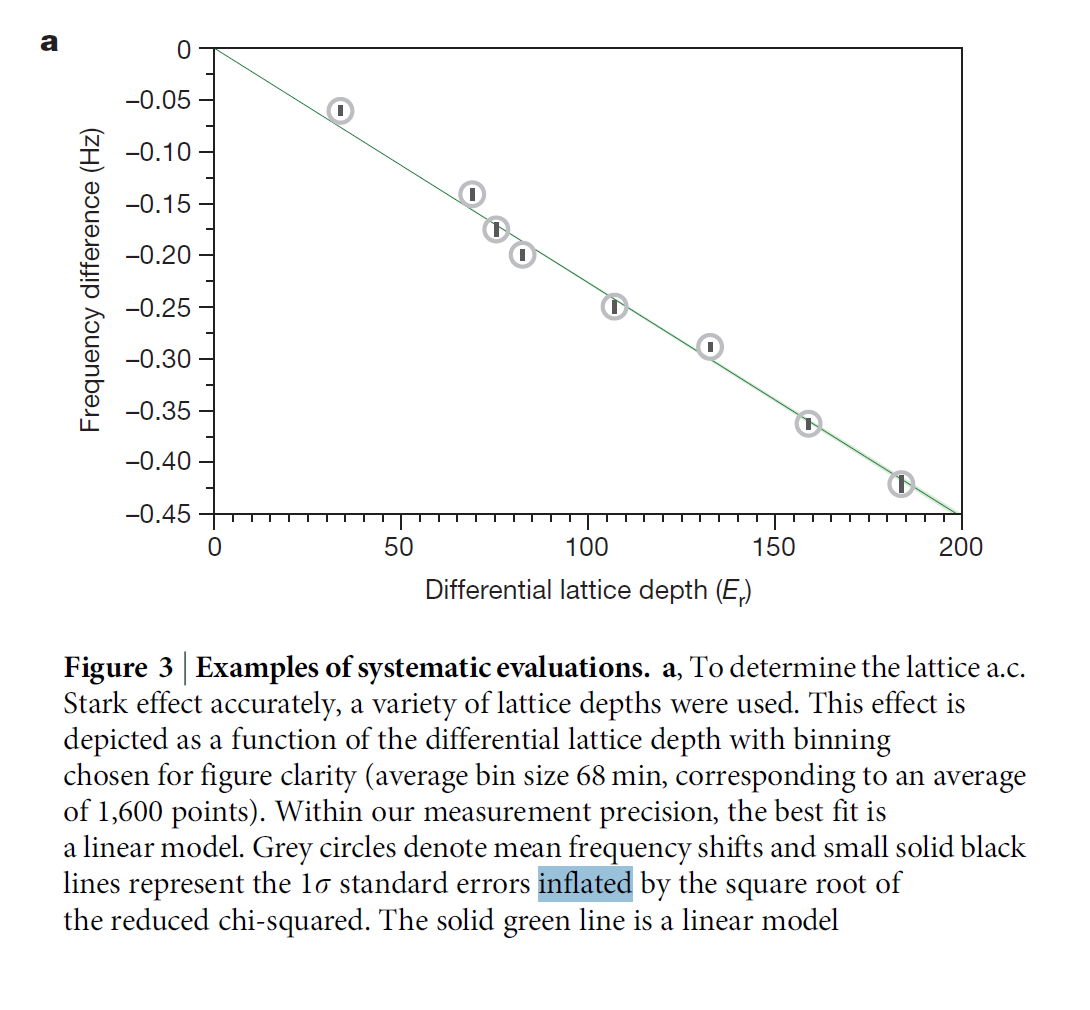
对应这张slide的例子：



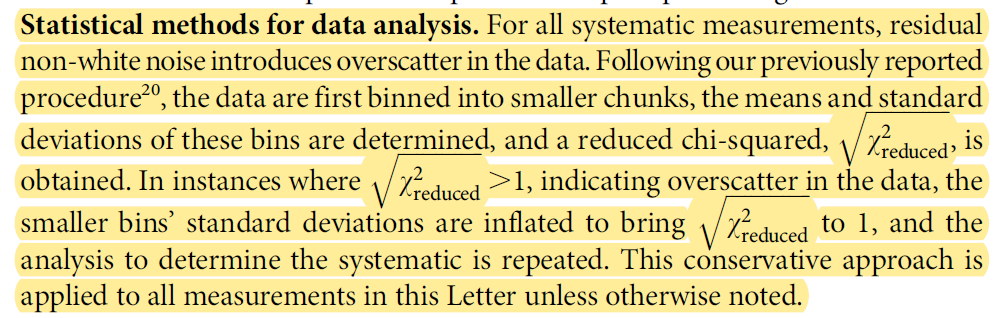
A：这个“乘以the square root of the reduced chi squared”操作，称为inflation。我发一篇文章先（Ref006）。







上面列出了三个例子，是在精密测量文献中的典型用法。你在读文献时，会在Methods，或者Supplemental materials找到更近一步的技术说明。其中还引用了更早一些的工作，你可以一篇一篇地追溯下去。





那么为什么要这样做呢？的确，就如你列出的那张slide上面说，这是理论理解和实验测量不完全一致时的一种解决方案。

我们一直强调物理学以实验为基础，但我们也要说明，这里的实验是所有物理实验的合集，甚至包括对实验室之外的自然数据的测量（如天体物理等）。任何一个实验，都只包含有限的信息，因而是不完备的。好的实验可以提供一颗靠谱的螺丝钉，使我们组装起科学的大飞机；但是不好的实验可能有错误，误导科学研究。介于中间的实验，它可能结论大体是合理的，但是不确定度的测算估计有问题，如果过大，那还好；如果过小，就会出现你上面提到的情况，明明理论和实验看着差不多，但是定量地就是不相符合。

Chi squared 是一个统计学的工具，它基于一定的统计假设（某个分布），来定量地判断数据是否符合某个模型，这个工具和我们说的3 sigma原则类似，只是更加一般性。

我们来回顾一下3 sigma原则，在多次测量求平均值的时候，我们看到有些点偏得特别多，如果按照“除此点外其他所有数据”分析得到一个standard error，然后弄一个正态分布的模型代进去算的话，会发现出现那么大的偏离，其概率小到不现实，因此估计，该点可能是个错误。

错误的测量是不符合统计规律的，但是，如果我们怎么查都查不出错误，操作一切正常呢？如果没有错误，但是我们又相信，这个物理量应该是不会变化的，那就出来第二种可能，就是不确定度估计低了。

比如我拿着螺旋测微器去测长度，估读到0.001微米那一位，如果我的状态比较稳定，观测的角度等也不变，我可能会读出下面的数字来：8.123 mm，8.124 mm，8.125 mm。这种的数据，其离散程度很小，如果我们不考虑螺旋测微器的允差的话，就会得出很小的standard error（远小于0.004 mm），考虑允差之后，就会得到一个更合理的不确定度估计。这个例子里，允差是已知的；在更复杂的实验里，一部分的允差是商用仪器公司给的，类似于螺旋测微器的允差。

但另一部分是自己搭建的仪器，这里面就可能出现“把各个零部件的允差合计起来”，得到的结果小于、甚至显著小于整个仪器允差的现象，这叫组合过程也可能产生相应的不确定度。

Q: 那这种自己组装的整体仪器的不确定度要怎么估计呢？

如果我们naively 只是把各个商用零件的允差加起来，就可能低估不确定度，使得测出来的点都在3 sigma以外。这种不确定度，就需要通过实际测量来估计，这正是inflation做的事情。如果我们坚信物理模型没有错，那错的就可能是对仪器整体不确定度的估计过低，为了测量出这种偏差“到底这种估计过低有多低”，就用inflation factor来进行数据分析和处理。

我再举一个例子，在原子钟测量里面，人们爱说，中性原子钟的好处是你可以用很多很多的原子，像同时抛很多很多硬币一样，这个叫做量子投影噪声（Quantum projection noise, QPN），和原子数的平方根成反比。按理说，你用10000个原子测出来的结果，应该比100个原子测出来的不确定度要小10倍，但是人们发现，好多时候并没有改进，这里面有些东西被忽视了，这些重要因素，首先是激光（local oscillator）的相位噪声，如果激光不改进，用再多的原子也没有用；其次是原子-原子相互作用，如果原子之间老是碰撞，就会产生额外的退相干，所以2017年叶军实验组把原子装进了3D光晶格，把每个原子孤立起来，而不是之前的一维光晶格。把这些因素都做到了，就发现你用更多的原子、更长的测量时间时，钟的稳定度确实大大提高了。

Q：可是有的时候不是“过于相信理论而忽视了实验的声音”吗？感觉这里面主观性比较大？

A：可以从两方面看：任何数据分析都基于一定的理论，同样的数据用不同的理论可以给出不同的解释。刚才的inflation是一种做法，它不一定是最高明的，但却是自洽的、合理的，不是错误。

但你也可以说，这里面有问题，你的仪器为什么会有额外的不确定度？我一定要搞清楚。你花很大功夫去研究，结果真的发现了，是哪个锁频模块不够稳定，或者是光路里面有一些应该减弱的etalon干涉效应，你把这些都去掉了，仪器的整体不确定度也就进一步减小了，更接近于一开始的零件误差之和。这是第二种做法，研究仪器，改进仪器，做测量科学的研究。

第三种就是，我的仪器没有问题，怎么查都查不出来，那你的理论一定有问题；或者反过来说，“如果没有Higgs粒子，这个数据点就是一个5~6 sigma的偏差，统计学上高度不可能”。 这样可以否定一个理论，或者用反证法证实一个理论。

在不同的领域，人们使用不同的策略。如果是作精密测量，那么你宣称一个非常小的不确定度时，一定是非常谨慎的，如果你宣称了一个小的误差，却发现实际上的偏差甚至错误远大于该不确定度，你就在学术同行面前犯了诚实但尴尬的错误，这是人人都尽力避免的；与此同时，如果你的数据很一致，不确定度确实很小，但你通过一定程序估计出的不确定度比数据的离散程度大50%，没有人会批评你什么，最多是说“这个人是个很谨慎的科学家，他一向采取保守估计。error inflation就是类似于上面这种谨慎和保守的思路，因为不确定度本身就是精密测量的目标，宁可把自己的结果说得保守一些，也不能over-claim。

在其他领域，尤其是做多体理论的，他们习惯于提出各种各样的新概念，新模型，那么通过chi square评估模型的合理性，就是一个普遍而合理的做法了

上面的讨论也可以在概念上与其他领域的一些方法联系起来，比如“黑盒测试”与“白盒测试”



我们说的error inflation，从最终结果出发，不管你是因为什么原因低估了测量不确定度，我最后都根据结果来加以修正，这很类似于黑盒测试。而“去做仪器科学，改进仪器”是白盒测试的思路。

我在这里再补充一点，chi squared是基于一些常见的分布，如高斯分布和一些派生的分布，而发展出来的工具。我们做error inflation，基于的其实是两条假设：1. 所测量的物理量保持不变，也就是“实验平台和仪器一直正常运转”+“该控制的因素都控制好了”=>“答案应该不变”；2. 不确定度背后的统计分布满足我们熟悉的规律，比如高斯分布及其派生分布。

如果这两点都没有问题，那么离散度相对于白盒分析得到的原始不确定度过大，就说明原始不确定度被低估了。反过来，如果我们“心与实验合”，一拍脑袋发现“哎呀，钟激光锁频信号现在质量很差”，那就不是做error inflation，而是要评估一下，哪些数据可用（其为在激光出问题之前采集的数据），哪些数据是错误了。

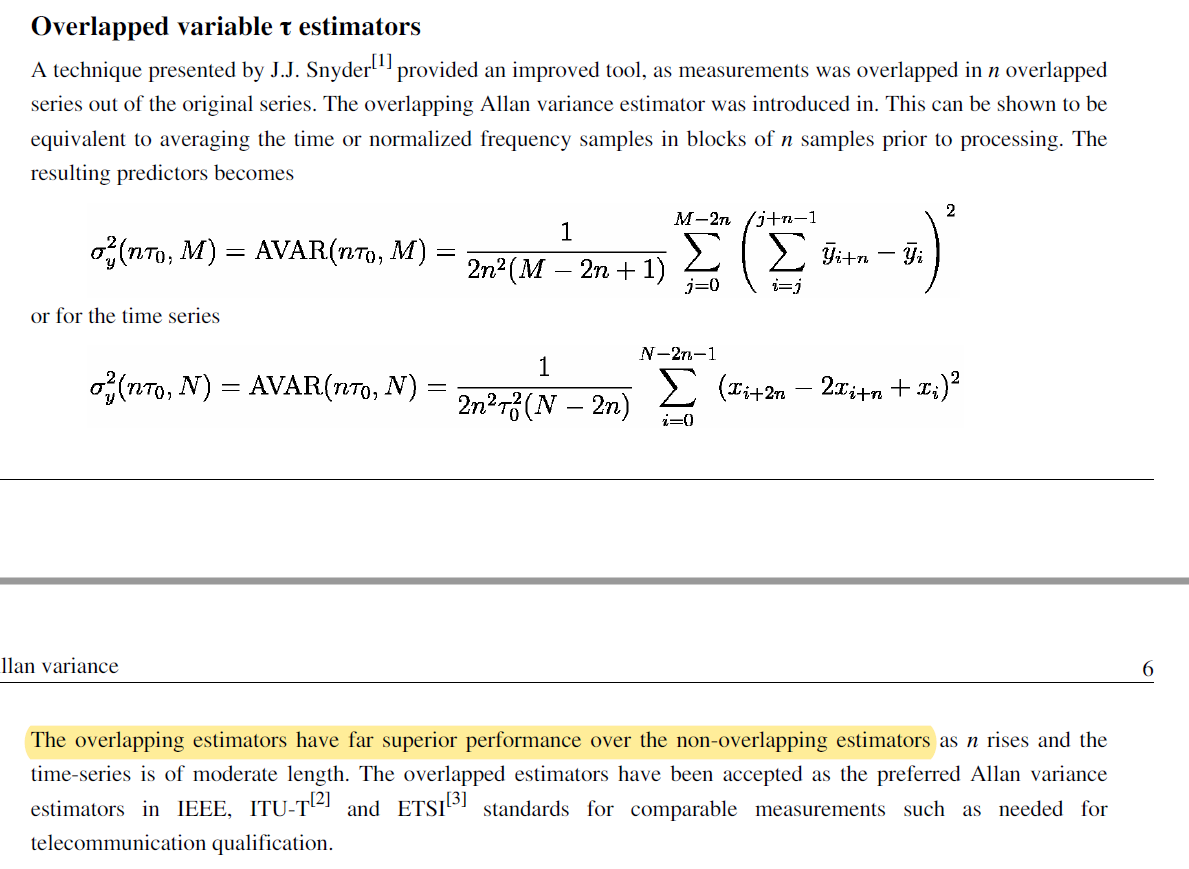
这些精细的区别，只能具体问题，具体分析，也就是我们通常讲的实事求是，不能从方法论上一概而论。科学研究的艺术，也就在于实事求是。

Q：running average的时候，感觉您好像提到了相当于把多次测量合为一次提高单次测量的信噪比；如果是这样，把相邻n个点做平均，一个点可能被用了好几次，平滑后数据不同点之间似乎有了关联，这样合理么？我以为应该把这几个点作为一组，只在组内平均，组和组之间当成独立的。

A：关于running average，或者说smoothing。如果只关心数据的最佳值满足的规律，暂时不care不确定度，那么running average是个很趁手的工具；当然，你平滑之后，不同的点的确有某种关联，这种关联无所谓合不合理，这是一种数据处理方式，关键看你如果去“弥补”或者“计入”这种关联带来的影响。有专门的统计方法来计入这种关联的影响的。

Q：您可以举一两个“专门的统计方法”的例子吗？

A：这里有一个，关键词是“Overlapped Allan Deviation”（Ref007）。



Q：如果x 和 y的测量都有误差，拟合应该怎么考虑权重？

A：我在杨氏模量的实验课上提到过，我们要分辨两个物理量中，哪个是不确定度大的那一个，哪个是不确定度小的那一个，后者当x，前者当y，因为教科书上的统计公式只计入y的误差。如果x也有误差的话，我们在计算chi square时，就需要将该函数的局部斜率考虑进去。比如你的数据大致是一条直线，那么你可以将x的不确定度通过初步拟合的斜率，转化成为y的一个独立的、等效不确定度并加以合成；这种方法在你的数据为非线性时也可以按局部斜率各自处理，问题只在于信噪比是否允许。

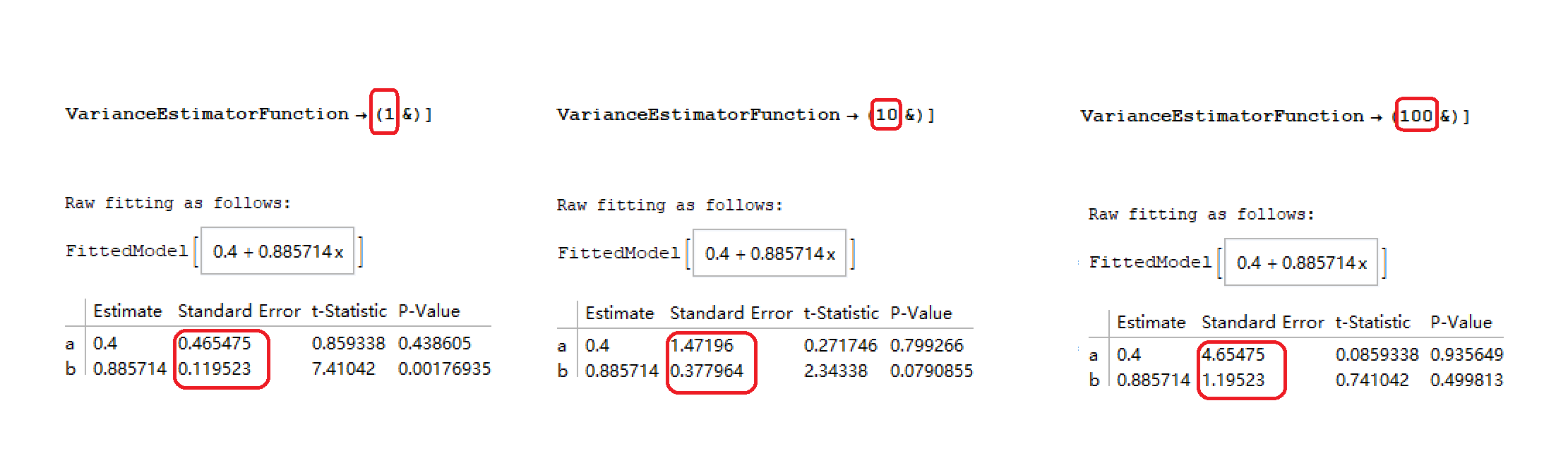
Q：请问下Hadamard diviation有实际应用的例子吗？

A：Hadamard deviation这个名词，我自己没有接触过，也没有在名义上使用过。如果让我猜测的话，这些不同的Variance Estimator是为了处理不同的noise类型而发展出来的。Allan Deviation只是其中用得最广泛的一种。这里的关键词是“noise power spectrum density (PSD)”，对于不同的noise PSD，比如白噪声（constant PSD），1/f noise （PSD 正比于Fourier 频率的倒数），等等，人们可以比较各种Estimator，选用最合适的，对精密测量的结果之不确定度进行评估。

Q：老师，我想问问录音1:12:00板书写下来的是什么。就是Mathematica里面那个要设成1的参数，不是特别懂。



A：“1 &”。 这个参数是一个比例系数，在Mathematica里面，作为练习可以随意变化的，你可以设成 2&，3&，……，5&看看结果会怎么变。设成1，其实是类似于令B=A，锁死了B = kA中的k这个自由度。



这是对于拟合结果不确定度的一个手动调节工具，我没有细究Mathematica的文档，但是可以做黑盒测试，从上图看到它的功能。既然Mathematica设置了这么一个参数，我们就需要确定它的值。现在我们不需要它，于是令它为1；如果不进行赋值，可能Mathematica会按照某些它自己的（不恰当的）判断去赋值，使我们的程序得出不正确的结果。

上面的例子，你用我提供的程序，改一改就可以自己玩出来。程序的设定，是为了方便我们的各种应用。我的程序里面，数据是有error bar的，这个程序的目的就是从有error bar的数据，去估计参数的不确定度。当测量error bar小的时候，参数不确定度由离散程度决定；当测量error bar大的时候，参数不确定度由测量error bar决定。正确的普适程序，必须在两种极限下都给出正确的结果。最后我想“说点反话”，这一讲“从数据到意义”，给大家展示了比平时教科书上更普适的一些统计方法，让大家可以更方便地去“得出一个数来”。但是任何工具，只是玩具，我们切忌玩物丧志。

我在北大时考GRE，参加了一个作文班，当时的价格是二百多元，老师是孙远。他列出了“GRE analytical writing的常见错误”，其中第一条就是：

CONFUSING CAUSE WITH CORRELATION。

我们做拟合，看模型的有效性，算相关系数，这些都是correlation，而correlation（关联关系）不等于因果关系。最少，当A和B强相关时，你无法从相关性本身看出A和B有没有因果，或者如果有的话，谁是因，谁是果。比如，虽然大学100%都有大楼，但是大楼却不是大学之为大学的原因。

我们的一些逻辑玩具，比如对照组实验方法，微分测量方法（differential measurement），等等，也都是很努力地去从关联关系中分辨、提取出深层次的根本性的因果关系，这是一个高度非平庸的过程，科学研究的美妙，也在于此。

大家多学统计学的工具，为的是用它做有意义的事情，“用之为善”，而不是膜拜它，或者成为统计教的盲目信徒。我最后再举一个例子，关于程序正确的：在一间黑屋子里面做光学实验，你要不要戴上激光护目镜？程序正确说，戴，一定要戴，戴了就没有问题了。但是真正有经验的研究人员说，最关键的不是戴护目镜，而是熟知你搭起来的每一个高功率的光路，它的光都打在哪里，这种对实际激光光路的了解才是保护眼睛的根本要义，不了解光路走向，只凭护目镜，可能让你更加麻痹大意，看不清楚，反而更危险。当然，在我的实验室里，是不允许不理解光路的新手贸然进来的，但是，如果换成一个探索性的，没有老师看着的场景呢？

不要戴上护目镜就乱闯，要先理解你的工作环境。

内容正确，永远比程序正确更重要，这就是“实践是检验真理的唯一标准”。

与大家共勉。

作业：

 3月16日发送至邮箱：[han.z@pku.edu.cn](mailto:han.z@pku.edu.cn)