单因素分析筛选变量时变量应与多因素分析所用变量相同

作者：李楠，赵一鸣 来源：临床流行病学和循证医学

本文原地址：https://www.iikx.com/news/statistics/1981.html

切勿“ 前后不一 ”——单因素分析筛选变量时变量应与多因素分析所用变量相同。

在统计分析的过程中，当结局存在多个影响因素的时候，我们常常需要借助多因素分析的方法得到每个因素的独立效应。对于这类分析，临床流行病学经典的做法是先通过单因素分析，初步筛选出可能有意义的变量(不同的研究者推荐的初筛界值略有差异，通常为p<0.1，p<0.15或p<0.2)。经过初筛后，再将
①初筛p值符合上述标准，或是②p值大于上述界值但是临床或之前的研究中认为该因素确为结局影响因素的指标，将这两类指标纳入多因素分析进一步验证是否有效应，并估计效应大小。

在这一过程中，涉及了两类分析手段的过渡，就是从单因素分析到多因素分析。在多数临床研究中，如果样本量足够负担多因素分析时，单因素分析往往只用作初步变量筛选，而不作为最终的结论。不少研究者也习惯接受这样的看法，但是您的单因素分析真的用对了么?这里面有个至关重要的问题，就是
您的单因素分析、多因素分析中，自变量(影响因素)和因变量(结局变量)的尺度是否是一致的?

举个例子，假设我们想研究我国成年社区居民中血清总胆固醇(TG)过高的危险因素，初步纳入了BMI、家族史、年龄、性别等指标。此时，反应肥胖的BMI可能是TG过高的危险因素，假设我们最关心的结局是TG是否达到或超过了有临床意义的高值。毕竟TG一般意义上的增高似乎临床意义也不大，只有达到了一个水平(比如大于5.98mmol/L)才会有临床意义。这也就意味着，在最终分析的时候，**研究者很可能会将TG作为一个二分类变量，使用logistic回归进行多因素分析。**

此时问题就来了，我们手里有TG的实际水平(连续变量)，还有BMI的实际水平(连续变量)。我们在通过单因素分析筛选变量的时候，是否还能直接分析TG水平和BMI水平之间的相关性呢?此时不少研究者会觉得“ 没问题啊?不都是这两个变量么? ”其实不然，作为连续变量的TG从信息量上看与二分类变量TG已经不一样了，二分类变量的信息量少了很多;此外从临床意义上看，TG水平与BMI水平之间存在相关，与BMI在TG是否异常升高两组间是否存在差异，这已经是两个完全不同的任务了。当然，单纯从统计分析的过程考虑，连续变量间相关的假设检验与两组间均数的比较，两者的p值已经不是一回事儿了。总之，

直接点儿说，**多因素分析用什么样的变量，单因素分析就要用完全相同的指标来筛选变量。**

同理，BMI也可以从连续变量变为二分
类变量。如果我们
最终要纳入多因素分析(logistic回归)的是BMI的连续变量，则在最开始的单因素分析中，我们就应该把BMI视为连续变量来处理
，展示BMI在TG高低两组间的平均水平，并比较BMI水平在TG组间的差异(t检验或非参数检验)。但如果我们最终关心的是肥胖与否是否会影响TG水平及效应大小，也就是logistic回归的时候要将BMI超标与否
这个二分类变量代入模型，此时单因
素分析的时候我们也要
将BMI作为分类变量处理——单因素分析与多因素分析中变量的形式保持一致。

此外，上述说法仅限于基于临床流行病学传统思路的数据探索和分析。其实近年来越来越多的方法学专家都逐渐接受了一个看法，就是在建立统计学模型的过程中，建立一个每个变量都有统计学意义的模型并不一定最好，更重要的是建立一个理论上完整和正确的模型。毕竟当一个理论上对结局非常关键的因素，即便没有统计学意义，把它剔除在模型之外也会影响其他指标效应大小的估计。因此在通过模型筛选变量的同时，我们一定不要忽略模型本身的合理性。