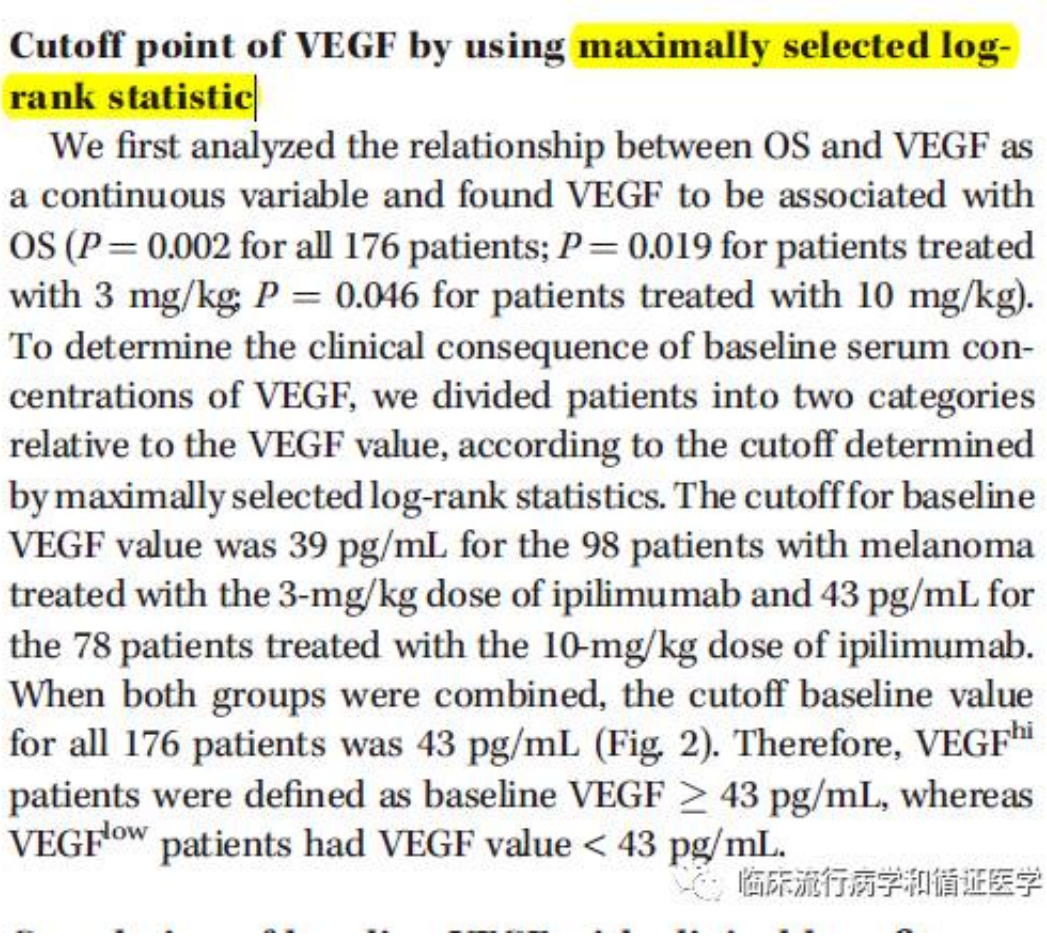
对连续性变量进行分类转换的一种方法----最大选择检验

作者：陶立元 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/1747.html

在临床研究中我们会遇到如下问题：如果x指标的测量值是连续的，对于结局指标y来说，如何将x指标进行分类(分为两组)，才能够获得y指标在两组间差异最大。

这个问题其实是一个对连续性指标x找切点的问题。很多人会想到ROC曲线，的确如果y是分类的，可以考虑用ROC来找x的切点，同时还可以考虑用分类树的方法等。但是如果y是连续的，又或者是生存数据该怎么呢?下面举个例子。



Cutoff point of VEGF by using maximally selected log-rank statistic

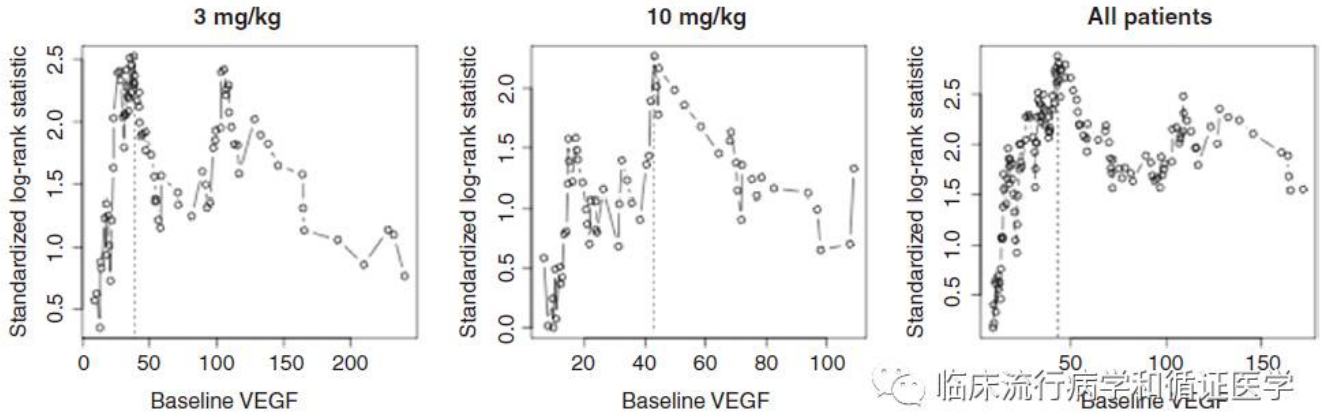
We first analyzed the relationship between OS and VEGF as a continuous variable and found VEGF to be associated with OS ($P = 0.002$ for all 176 patients; $P = 0.019$ for patients treated with 3 mg/kg; $P = 0.046$ for patients treated with 10 mg/kg). To determine the clinical consequence of baseline serum concentrations of VEGF, we divided patients into two categories relative to the VEGF value, according to the cutoff determined by maximally selected log-rank statistics. The cutoff for baseline VEGF value was 39 pg/mL for the 98 patients with melanoma treated with the 3-mg/kg dose of ipilimumab and 43 pg/mL for the 78 patients treated with the 10-mg/kg dose of ipilimumab. When both groups were combined, the cutoff baseline value for all 176 patients was 43 pg/mL (Fig. 2). Therefore, $VEGF^{hi}$ patients were defined as baseline VEGF $\geq$ 43 pg/mL, whereas $VEGF^{low}$ patients had VEGF value < 43 pg/mL.

有研究者用Ipilimumab单抗治疗晚期黑色素瘤患者，测量患者血清的VEGF水平与治疗的疗效，

研究者发现患者的OS与治疗前的VEGF水平有关。此时研究者想探索VEGF的切点在哪儿?才能够使得两组患者的OS差异最大。他们在文中便使用了Maximally Selected Log-rank Statistic(最大选择log-rank检验)。

MaximallySelected Log-Rank Statistic是最大选择检验(Maximally Selected Test Statistics)的一种，除了他以外还有Maximally SelectedChi-Square Statistics和Maximally Selected Rank Statistics等，分别应用于变量y的不同类型下。何谓最大选择检验呢?简单来说，就是对x进行若干次分类，只到找到一个切点值使得分类后的两组y值有着最大的统计量。

拿上面的例子来说，就是对基线的血管内皮生长因子找不同的切点，只到找到一个切点使得两组患者的总体生存率差异最大。上图也指出，研究者最后选择的VEGF的切点是43 pg/ml。作者还分别在不同的剂量组中，利用最大选择检验寻找了切点，如下图：



上面我们介绍了最大选择检验的一种应用场景，下面我们来看看如何实现。目前比较简单的实现方法是利用R的maxstat包，这个包中的例子是利用平均基因表达量(MGE)去区别两种弥漫性大B细胞淋巴瘤，区分的依据是患者的OS资料。其语法和计算结果如下：

```r
library("maxstat")
library("survival")
data("DLBCL", package="maxstat")
mtHL <- maxstat.test(Surv(time, cens) ~ MGE, data=DLBCL,
                     smethod="LogRank", pmethod="HL")
plot(mtHL)
mtHL

DLBCL$group<-ifelse(DLBCL$MGE>=0.1860526,1,0)
surcur = survdiff(Surv(time, cens) ~ group, data=DLBCL)
abc=survfit(Surv(time, cens) ~ group, data=DLBCL)
plot(abc, col = c("black","red"))

```

16:22 (Top Level)

Console ~/

```
> mtHL

Maximally selected LogRank statistics using HL

data:  Surv(time, cens) by MGE
M = 3.171, p-value = 0.02218
sample estimates:
estimated cutpoint
         0.1860526
```

结果显示MGE的切点是0.186。用此切点分开两组，做单因素分析其生存曲线如上图，哈哈，随意做了一个曲线，比较丑。另外需要说明的一点是最大选择检验不仅能够用来一个x指标，还可以用来同时处理几个x指标。