

样本量估算过程中的 I 类和 II 类错误

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/1669.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

在样本量的估算过程中我们经常会使用到 I 类和 II 类错误，I 类错误的取值往往是 0.05，II 类错误的取值往往是 0.1 或者 0.2。那么 I 类和 II 类错误分别代表什么意思呢？又是如果出现在了样本量的计算公式中的呢？本文就此问题作一讨论。

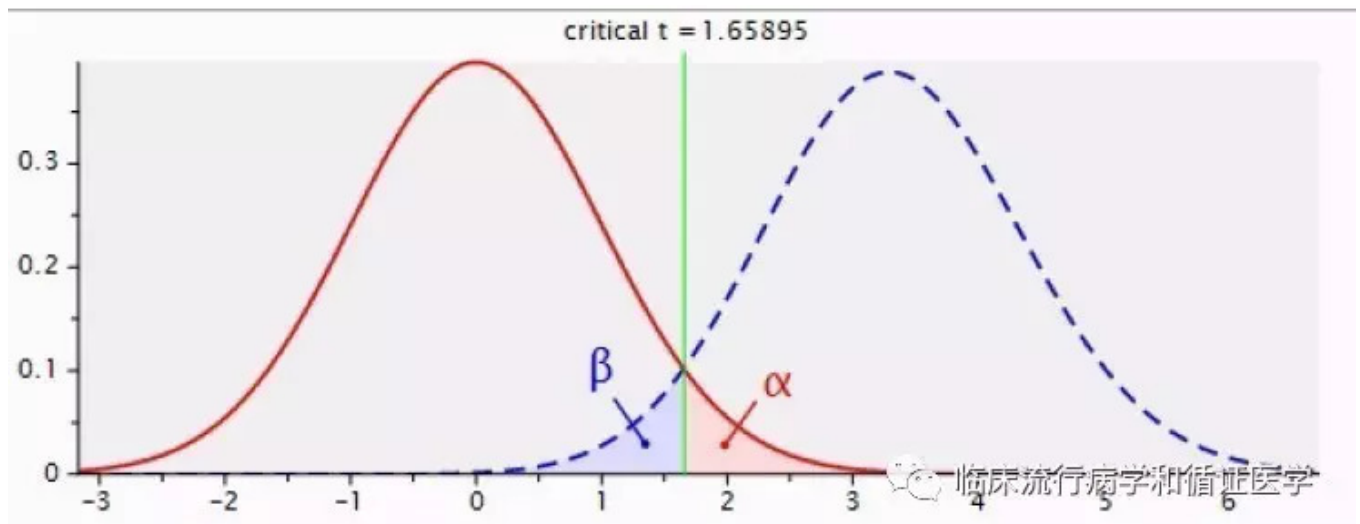
首先介绍 I 类和 II 类错误的概念。我们知道统计推断是用样本量去推论总体，那么在这个推论的过程中我们无外乎离不开两种结局，1 是对了，2 是错了。错了就是犯错误，错又有两种错法，1 种是总体是相同的，你说不同；还有 1 种是总体是不同的，你说相同。这就分别是 I 类和 II 类错误的概念。如下图：

客观实际	本次检验的结果	
	$\mu_0 \neq \mu_1$	$\mu_0 = \mu_1$
$\mu_0 = \mu_1$	I 类错误 (α)	推断正确 ($1-\alpha$)
$\mu_0 \neq \mu_1$	推断正确 ($1-\beta$)	II 类错误 (β)

临床流行病学和循证医学

举个例子说明，如果某个药物事实上是无效的，而我们的研究却说它的有效的，这就是 I 类错误。而 II 类错误是某个药物事实上是有效的，而我们却说其是无效的。在临床研究中，没有人想去犯错误。所以我们计算样本量的目的，就是在不犯 I 类错误的前提下，尽可能地少犯 II 类错误。在此条件下，我们就会有更大的机会去发现真正有疗效的药物的疗效。

在样本量固定的时候，若 I 类错误下降 II 类错误就会上升，若 I 类错误上升 II 类错误就会下降，如果我们想同时减少 I 类和 II 类错误，就只能增加样本量。如下图，随着绿色垂线的移动，两者此消彼长：



对于使用样本量计算这一过程来说，了解了上面的内容，其实就已足够。但是一定还有人想知道样本量计算的公式是怎么推导的。下面以均数检验为例进行简单的说明。样本量的估算包括参数估计和假设检验，参数估计对应着精度分析，假设检验对应着把握度分析。首先来说精度分析，精度分析就是考虑一个统计量的可信区间有多大，它也是参数估计的样本量推导过程。

假设 y_i 符合一个正态的独立同分布，那么推导过程如下：

1. y 均数的可信区间为： $\bar{y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

2. 假设 E 为最大允许宽度，则：

$$E = |\bar{y} - \mu| = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

3. 依此可以推导出 n ：

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$$

临床流行病学和循证医学

再说把握度的分析，它对应着假设检验的样本量估算过程。同样以最简单的两独立样本均数比较为例推导。由于在微信内公式编辑非常困难，以下内容以图片展示。

假设有两组，试验组 ($x_1, x_2, x_3 \dots$) 和对照组 ($y_1, y_2, y_3 \dots$)，其均数分别为 $\mu_1 \pm \sigma_1^2$ 和 $\mu_2 \pm \sigma_2^2$ 。

原假设为 $H_0: \mu_1 = \mu_2$ ，备择假设为 $H_1: \mu_1 \neq \mu_2$ 。假设两组标准差相同，样本量相同，则Z检验如下：

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}}$$

在原假设下，Z服从 $N(0, 1)$ 。当 $|Z| > z_{\alpha/2}$ 时，我们拒绝原假设。

 临床流行病学和循证医学
I.

当拒绝 H_0 ，接受 H_1 时， $\mu_1 = \mu_2 + \delta$ ，此时Z服从 $N(\mu^*, 1)$ 分布。其中

$$\mu^* = \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}} > 0$$

此时把握度则为：

$$P\{|N(\mu^*, 1)| > Z_{\alpha/2}\} = P\{|N(0, 1)| > Z_{\alpha/2} - \mu^*\}$$

 临床流行病学和循证医学
I.

则有： $Z_{\alpha/2} - \mu^* = -Z_{\beta}$

$$\text{即： } Z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}} = -Z_{\beta}$$

如此便可推导出样本量计算公式为：

$$n = \frac{(\sigma_1^2 + \sigma_2^2)(Z_{\alpha/2} + Z_{\beta})^2}{\delta^2}$$

 临床流行病学和循证医学

■

是不是有点儿蒙?这个推导的确不好理解。但是还是最简单的推断过程，除此之外还有率的推导，还有多组的推导和多因素的推导……，呃，想想就头大啊。

更多统计方法 请访问 <https://www.iikx.com/news/statistics/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发