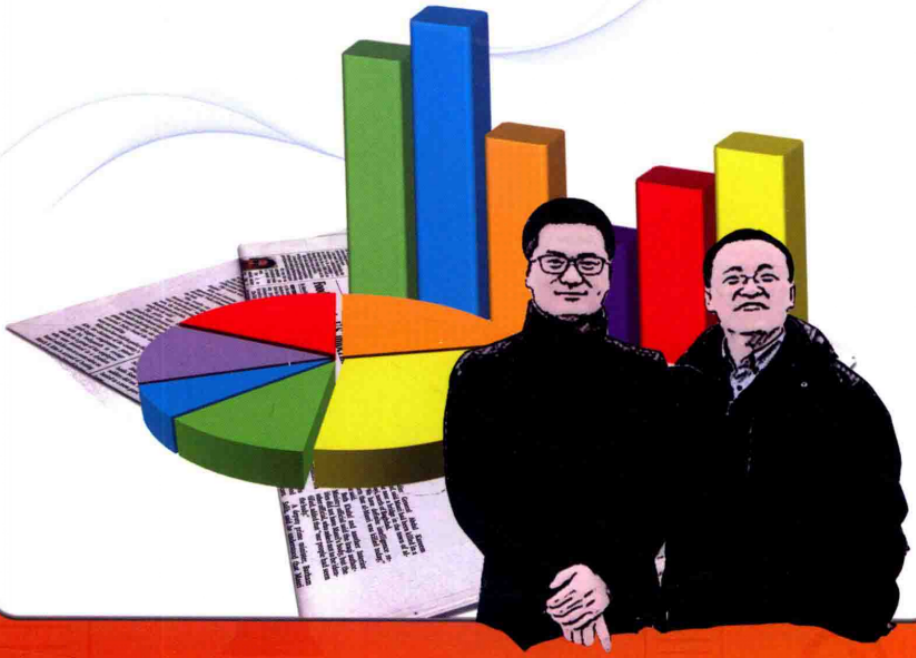
AME科研时间系列医学图书009

聪明统计学



主编　周支瑞　胡志德

南大学出版社
www.csupress.com.cn
科研时间
AME Publishing Company
丁香园 WWW.DXY.CN

AME科研时间系列医学图书009

聪明统计学

主编　周支瑞　胡志德

中南大学出版社
www.csupress.com.cn
科研时间
AME Publishing Company
丁香园 WWW.DXY.CN

图书在版编目（CIP）数据

聪明统计学/周支瑞，胡志德主编. —长沙：中南大学出版社，2016.5

ISBN 978-7-5487-2289-2

I.①聪 II.①周… ②胡… III.①统计学-通俗-读物 IV.①C8-49

中国版本图书馆CIP数据核字（2016）第115311号

AME 科研时间系列医学图书 009

聪明统计学

周支瑞 胡志德 主编

□丛书策划 汪道远 昌 兰

□责任编辑 陈海波 孙娟娟

□责任校对 石曼婷

□责任印制 易建国 潘飘飘

□版式设计 胡晓艳 林子钰

□出版发行 中南大学出版社

社址：长沙市麓山南路 邮编：410083

发行科电话：0731-88876770 传真：0731-88710482

□印 装 湖南印美彩印有限公司

□开 本 720×1000 1/16 □印张 18 □字数 350 千字 □插页

□版 次 2016 年 5 月第 1 版 □ 2016 年 5 月第 1 次印刷

□书 号 ISBN 978-7-5487-2289-2

□定 价 49.00 元

丛书介绍

很高兴，由AME出版社、中南大学出版社和丁香园网站联合策划的“AME科研时间系列医学图书”，如期与大家见面！

虽然学了四年零三个月医科，但是，仅仅做了三个月实习医生，就选择弃医了，不务正业，直到现在在做医学学术出版和传播这份工作。2015年，毕业十周年。想当医生的那份情结依旧有那么一点，有时候不经意间会触动到心底深处……

2011年4月，我和丁香园的创始人李天天一起去美国费城出差，参观了一家医学博物馆——马特博物馆(Mütter Museum)。该博物馆隶属于费城医学院，创建于1858年，如今这里已经成为一个展出各种疾病、伤势、畸形案例，以及古代医疗器械和生物学发展的大展厅，展品逾20 000件，其中包括战争中伤者的照片、连体人的遗体、侏儒的骸骨以及人体病变结肠等。此外还有世界上独一无二的收藏，比如一个酷似肥皂的女性尸体、一个长有两个脑袋的儿童的颅骨等。该博物馆号称“The Birth of American Medicine”。走进一个礼堂，博物馆的解说员介绍宾夕法尼亚大学医学院开学典礼都会在这个礼堂举行。当时，我忍不住问了李天天一个问题：如果当初你学医的时候，开学典礼在这样的礼堂召开的话，你会放弃做医生吗？他的回答是：不会。

2013年5月，参加BMJ的一个会议，会议之后，有一个晚宴，BMJ对英国一些优秀的医疗团队颁奖，BMJ的主编和BBC电台的著名节目主持人共同主持这个年度颁奖晚宴。令我惊讶的是，BMJ给每个获奖团队的颁奖词，从未提及该团队在过去几年在什么大牛杂志上发表过什么大牛论文，而是，关注这些团队在某个领域提高医疗服务质量，减轻病患痛苦，降低医疗费用等方面所作出的贡献。

很多朋友好奇地问我，AME是什么意思？

AME的意思就是，Academic Made Easy, Excellent and Enthusiastic。2014年9月3日，我在朋友圈贴出3张图片，请大家帮忙一起从3个版本的AME宣传彩页中选出一个喜欢的。最后，上海中山医院胸外科的沈亚星医生竟然给出一个AME的“神翻译”：欲穷千里目，快乐搞学术。

AME是一个年轻的公司，拥有自己的梦想。我们的核心价值观第一条是：Patients Come First！以“科研(Research)”为主线。于是，2014年4月24日，我们的微信公众号上线，取名为“科研时间”。“爱临床，爱科研，也爱听故事。我是科研时间，这里提供最新科研资讯，一线报道学术活动，分享科研背

后的故事。用国际化视野，共同关注临床科研，相约科研时间。”希望我们的AME平台，能够推动医学学术向前进步，哪怕是一小步！

如果说酒品如人品，那么，书品更似人品。希望我们“AME 科研时间系列医学图书”丛书能将临床、科研、人文三者有机结合到一起，像西餐一样，烹调出丰富的味道，搭配出一道精美的佳肴，一一呈现给各位。

汪道远

AME出版社社长

序（一）

写一本适合临床医生阅读的统计学

2014年，微信正火，各种微信公众号让人应接不暇，但其中鱼目混珠、良莠不齐。2014年7月AME出版社有了自己的微信公众号——“科研时间”。从创办之初，我们就希望在这个平台上推出一些对临床医生有用的信息，通俗地说，我们希望这个平台推出的都是“干货”。我的好友胡志德第一次把自己的一些统计学体会放在了AME出版社的“科研时间”微信公众号上，反响不错，后来他又陆续写了几篇，反响依然不错。彼时我正坐在复旦大学医学院的课堂里，上着无聊透顶但又经常随堂点名的博士阶段必修课程，想来我们的教育也只能堕落到如此地步了。与其浪费时间，不如写点东西吧，我心里想：我写出的小文多少能够给大家提供点有用的信息。当我决定做一件事，我会变得非常专注与高效。短短的一个多月的时间里，我为“科研时间”微信公众号写了十几篇循证医学与统计学相关的文章，而且全部放在了“循证杂谈”专栏里。文章在微信平台推出后反响还行，至少其中提到的信息对于临床医生来说是有价值的。与此同时，胡志德的“AME统计”专栏也红红火火。就这样积累到20多篇文章的时候，AME出版社的汪道远社长提议：把统计专栏与循证医学专栏的文章编纂成册出版，书名就定为《傻瓜统计学》。当时我并不十分赞同编纂出版，主要基于两个考虑：第一，文章数量有限，未成系统；第二，语言太过通俗，算不上“雅”，但我保留了我的意见。这本书后来顺利出版了，而且不到一年的时间里卖出了近万册。当然我们并不指望这本书赚什么钱，现如今出版行业不景气，出本书只要自己不掏腰包赔钱都算赚了。我和胡志德完全是凭着爱好与热情在做这件事情，而且凭着这股劲儿，我们希望把这件事情坚持下去。

自《傻瓜统计学》出版以来，我们就一直在筹划着再做一本姊妹篇，其实书名是早就定好的，名曰《聪明统计学》。总觉得名字有点“俗”，但是“俗”的另外一面就是“接地气”，是的，我们就要做一本接地气的、通俗易懂的、可以复制和模仿的、适合各层次临床医生阅读的统计学书籍。亲爱的同行，如果您有缘接触到这本《聪明统计学》，而且有足够的耐心阅读了前言，那么下面两个问题是我要重点阐述的，而且我相信下面两个问题对于临床医生学好统计学大有裨益：

为什么要写这本书？

过去的一年中，我们AME College（亦塾）主办了一个旨在提高临床医生

SCI论文写作水平与投稿技巧的培训项目，在这个课程中有3个小时的时间专门用于讲授与临床科研暨SCI论文写作密切相关的统计学知识，这个培训课程的主要受众是在读医学研究生以及年轻的临床医生。理论上说，这些具有高学历的临床医生，至少在本科教育阶段与研究生教育阶段都系统学习过《医学统计学》，然而让我感到吃惊的是，绝大多数同行连基本的统计学知识都没有掌握，统计学知识的匮乏程度让我觉得不可思议。后来，在与他们的交流中，我总结造成这种局面的原因主要有以下几点：第一，无论大学还是研究生阶段，统计学压根就没有学好，基本是应付考试过关。这到底是学生没有学好还是老师没有教好？在我看来后者的可能性更大。教授统计学的老师，多半是统计学专业出身，他们的思维方式与临床医生或医学生的思维方式有很大不同，前者更倾向于逻辑思维，后者更倾向于形象思维，这如何能让医学生在大学就学好统计学？第二，我们使用的统计学教材过度强调知识结构与表达的严谨性，而忽略了教材的可读性。我的家里有各种统计学教材几十本，国内统计学教材无一例外是一堆的统计学公式与符号，还有繁琐的公式推导过程，而这些符号与公式让人眼花缭乱，实在提不起来阅读的兴趣。反观一些国外的教材，公式很少见，插图做得很精美，读起来让人觉得轻松。第三，我们的教育方式重基础理论，轻视实践运用。而医学统计学这门学科恰恰应该以案例教育为主，怎么强调实际运用都不为过。我们在培训过程中也发现，有些同行理论知识相对扎实，但是却很难把理论知识运用到具体的实际案例中去，这种情况就是典型的教学方式不恰当造成的。因此我们要做一本与众不同的适合临床医生阅读的统计学教材。

我们为什么写得好？

细心的读者可以发现，无论是《傻瓜统计学》还是这一本《聪明统计学》，我们的编者队伍是清一色的临床医生，这就是为什么我们能把这本书写好的根本原因。具体说来有以下几点：第一，前文也提到，国内的统计学教材的编者多数是统计专业出身，他们的思维方式与我们这些医生可能不同。也许在他们看来公式符号之类的就是美丽的音符，而事实上对于我们临床医生来说，公式与符号就是不可逾越的鸿沟。因此，我们更了解临床医生应该怎样学好统计学，概括起来说就一句话：以案例为主。事实上对于繁琐的公式我们并不需要掌握，我们的重点在于实际运用，而非理论。在我看来，临床医生只需掌握统计学的三把斧：正确地选择统计方法，统计软件得出结果，最后正确地解读结果。第二，我们更了解临床医生需要什么，我们更了解SCI论文写作中需要什么样的统计结果与统计图表。本书中介绍的统计学方法都是临床医生使用频率很高的方法，比如生存分析、三大回归分析、倾向性匹配得分、样本量估算等等。举个例子，我们在报告生存曲线的时候，经常涉及计算“历险数

（Numbers at risk）”并在生存曲线底部显示，这种小细节在正统的统计学书籍里找不到答案，但在我们这本书里有完整的解决方案。第三，我与胡志德认识多年，我们的成长轨迹就是一个普通临床医生学好统计学的励志故事。起初我们的统计学也就是在大学里学的三脚猫功夫，但是我们两个有个共同特点：热心肠。经常有同行会问到我们这样那样的棘手的统计学问题，我们凭着这股热心肠的劲儿，帮助别人一个一个解决问题。当解决的问题慢慢多了，我们也在一次一次帮助别人的过程中慢慢成熟了。所以本书中讨论的大部分问题都是基于临床医生面对的现实问题，绝大多数问题被不同的人反复咨询过，而非空谈统计学理论。综上，能把医学统计学写得这么“接地气”，舍我其谁？

最后，简单介绍本书的逻辑架构，本书共分为四个专题：统计学趣谈、统计软件实战、大数据与科研、数据纵横。“统计学趣谈”专题主要由胡志德医生操刀，其中丁香园鼎鼎大名的四叶虫——郑炜平医生，号称临床医生中统计学最好的张天嵩医生也加入其中，贡献了一把火。这一专题力图以“趣谈”的形式让艰涩深奥的统计学知识润物细无声。“统计软件实战”专题以案例的形式讲解具体统计学方法的软件实战，把第一部分的理论知识化为实际的战斗力，这一部分主要由上海静安区中心医院的张天嵩、复旦大学附属中山医院沈亚星、胡志德，以及我本人操刀。“大数据与科研”专题属于目前临床科研领域较热门的话题，我们也赶了趟时髦，希望能给读者一些启发。最后一个专题“数据纵横”，由浙江大学金华医院的章仲恒医生操刀，章医生也是我们的朋友，擅长临床数据收集与处理、临床数据库构建、数据统计分析，看起来这也是他业余时间最重要的爱好。本书的四个专题逐层递进，先易后难，但书中的案例坚持从临床实际运用出发，相信定会让您觉得开卷有益。

另外，本书实践操作部分的数据主要来自于以下统计学教材共享的案例数据：张文彤主编《SPSS统计分析高级教程》、周登远主编《临床医学研究中的统计分析和图形表达实例详解》、陈峰主编《现代医学统计方法与Stata应用(第2版)》等教材，这些书籍写的通俗易懂，很适合医学相关专业同行阅读，在此对以上书籍的编者表示感谢！

(周支瑞)

序（二）

将复杂的问题简单化

据目测，《傻瓜统计学》一书的销量应该突破一万本了，在AME出版的所有图书中，这本书的销量长期高居榜首。如此巨大的销量，确实出乎意料。分析其中原因，我认为一方面是价格偏低且广告打得较好；另一方面是内容“接地气”，因为好几个素未谋面的同行都反馈过来信息：书写得很好。不管如何，这个销量肯定是极大地满足了我的虚荣心。当然，也是在虚荣心的驱动下，我鼓起勇气，和周博士、章医生等好友一起写了这本《聪明统计学》，也就是《傻瓜统计学》的升级版。

谈到出书，虽然过去几年我作为副主编，实实在在地参编过几本专业书，但真正谈到出版一本统计学的书，我还是很没底气的。毕竟，我并非流行病学或统计学专业人士，编写统计学书籍对我来说就属于跨区作业了，随时可能出洋相。我常笑谈，流统专业人士掌握的是“科班统计学”，而我掌握的只是“山寨统计学”，或者叫“非主流统计学”。我了解的这些统计学知识只不过是通过反复研读教科书、科技论文的案例，结合自己长期从事临床科研的经验总结而来的。

我们为什么要努力传承《傻瓜统计学》的风格，编写这本《聪明统计学》？

笔者在攻读硕士和博士期间，应该算是系统、认真地学完了所有医学统计学课程。授课的老师均为专业的流行病学与统计学人士，水平肯定是不容置疑的。但我发现，很多同学听了之后还是一头雾水，到写论文时还是会把统计学方法弄错。究其原因，我认为主要是“科班统计学”的理念强调：只有对统计学原理有十分深入的了解，才可能谈得上正确应用。正因如此，授课老师十分强调每个统计学方法背后的数学原理。其实，在我看来，这个理念没错，但问题在于，多数医学生从高中毕业后几乎就不再碰数学，如何能看懂那些纷繁复杂的数学公式和数学概念？再则，统计学是一种科研工具，只有在不断应用中才能体会到掌握统计学要领，纯理论的东西，如果长期不用，势必逐渐淡忘。就像游泳，如果自己没有亲自到水中去摸爬滚打，不管老师怎么传授理论知识，最终也不可能学会游泳。

另一方面，统计学并不是孤立存在的，在考虑用什么统计学方法的时候必须先要回答两个问题：研究目的是什么？研究是如何设计的？所谓“临床需

求决定研究目的；研究目的和现有资源决定科研设计；科研设计决定统计方法”，抛开临床需求和实验设计谈统计学，无异于缘木求鱼。

正因如此，《聪明统计学》在编写过程中，我们一再强调不要讲那些晦涩的数学原理，也不要试图把内容写得像教科书那样全面，而是着眼于一些虽然很小，但是很具有实战意义的话题，结合自己做临床科研的体会展开讨论。我们所撰写的内容，虽然源自教科书，但是我们不抄袭教科书的文字，而是将教科书上对我们最有用的知识提炼出来，用最简单的语言进行描述总结。此外，我们还结合自己长期从事临床医学科研以及为杂志审稿的经验，列举了一些常见的例子用于示例。我们提倡的核心理念是：去除冗余信息，将复杂的统计学问题简单化。

(胡志德)

目　录

第一部分　统计学趣谈

第一章　标准差和标准误：两个经常被混淆的概念

在医学统计学中，有两个基本的概念：标准差(SD)和标准误(SEM)。据笔者观察，很多医学研究者，特别是刚走上医学科研道路的研究生，分不清楚标准差和标准误的区别，以至于常常得出一些令人啼笑皆非的结论。在此，笔者拟简要阐述标准差和标准误的区别，并引申出“参考范围”和“95%可信区间”这两个同样被经常混淆的概念。

1　标准差与参考范围

标准差这个概念最好理解，在初中数学教科书上就有过比较清晰的解释。简而言之，标准差反映的是数据的离散程度，或者说波动幅度。打个比方，有一项研究拟调查黑龙江省和海南省居民的个人年收入状况，结果发现黑龙江省居民个人年收入的标准差较大，而海南省的个人年收入的标准差很小。这一结果所暗含的意思就是：黑龙江省的贫富差距很大，而海南省的贫富差距很小。因为一个省内所有居民年收入的离散程度其实就是指“贫富差距”问题。假定海南省每个人的年收入都是一样的，其标准差自然就为0了，即数据一点都不存在“离散”问题，海南省不存在贫富差距问题。

在医学上，利用均数(mean)和标准差(SD)可以计算某项实验室指标的参考范围，其做法就是截取该实验室检查结果95%的分布区间作为参考范围。比

如，目前需要确认血钾浓度的参考范围，标准的做法就是：首先需要检测一部分(通常为100个样本以上)健康个体的血钾浓度。然后绘制出其分布状况(如图1所示)的直方图。根据专业知识可知，健康个体的血钾浓度不可能太高，也不可能太低，经过正态性检验后发现血钾浓度在健康个体中的分布完全是呈正态分布的(过程略)。因此，就设定数据95%的分布范围作为参考范围。众所周知，在正态分布曲线中，mean−1.96×SD与mean+1.96×SD所涵盖的区间刚好覆盖了95%的样本，因此参考范围的下限就是mean−1.96×SD，约为3.5 mmol/L；参考范围的上限就是mean+1.96×SD，约为5.5 mmol/L。换而言之，大约有5%的健康个体的血钾浓度在参考范围以外，但仍然属于健康个体。

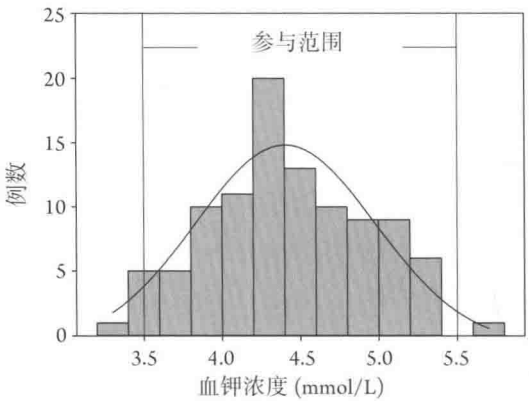


图1　100名健康个体的血钾分布状况

当然，如果实验室指标不是呈正态分布的(比如肿瘤标志物)，参考范围的制定就相对麻烦点了，由于不属于本文论述的范围，暂且不作讨论。

2　标准误与95%可信区间

在讨论标准误和95%可信区间之前，我们需要先将话题说远一点，谈谈医学研究的特点。医学研究最大的特点就是抽样调查，通过样本去推断总体。比如，某课题欲确定山西人的平均身高，理想的调查方案显然是把所有山西人的身高数据全部汇总起来，如果结果显示山西人的平均身高是178 cm，那自然可以理直气壮地得出结论：山西人的平均身高就是178 cm。但问题在于，在医学研究中，很多时候无法将某一特定的群体全部汇总起来，因为这个总体本身就是无法确定的。比如：要调查健康个体的平均甲胎蛋白(AFP)水平，总不能把所有健康个体的甲胎蛋白全部测一遍吧，即使财力上允许，这在技术上也很不现实，因为“健康个体”本身就是一个很模糊的总体。

既然研究总体不现实，那就研究样本吧。于是，人们抽取了100个健康个

体，检测了其AFP水平，假设发现AFP平均水平为100 ng/mL(标准差略)，然后就用这100个人的结果去推断所有的健康个体，于是得出结论：健康个体的AFP平均水平是100 ng/mL。这是目前最常用的研究方式，但是弊端显而易见：用样本去推断总体必然存在误差！简而言之，研究者抽取的这100个健康个体(样本)能代表所有的健康个体(总体)吗？从统计概率的角度来讲，完全可能出现以下情况：其实健康个体真实的AFP平均水平应该是200 ng/mL，但是由于研究者在抽样时出现了抽样误差(手气问题)，全部抽取到了AFP较低的健康个体，因此误认为健康个体平均AFP的真实值是100 ng/mL。

根据样本去推断总体这种研究方式有一个特点，就是样本量越大，越不容易出现抽样误差，结果也越准确。以“健康个体平均AFP水平”这个问题为例，举个极端的例子：假定抽取的样本只有两人，AFP当然也会有一个均值，但这两个健康个体(样本)的AFP平均水平可能与所有健康个体(总体)的平均AFP水平相距甚远。如果抽取的样本不是两个人，而是两万，结果就不一样了，毕竟在两万个样本时出现抽样误差的概率是很小的。用两万个样本去推断总体，结果显然更为准确。

人们发现(这段是重点)：当固定样本量(比如每次都抽100个健康个体)的时候，每次抽样后得到的平均值虽然不尽相同，但是总体而言，所有的平均值都呈正态分布(这句话虽然不太严谨，但是为便于理解，大致可以这样认为)。用通俗的话来讲就是：假定张三抽取了100个健康个体，得到了一个平均AFP(AFP1)；李四采用和张三相同的方法再去抽取100个健康个体，得到了一个平均AFP(AFP2)；王五、赵六、钱七等人如法炮制，就会得到AFP3、AFP4、AFP5直至AFPn。如果把这n次抽样得到的AFP均值汇总起来，会发现这些均值是呈正态分布的(如图2所示)。

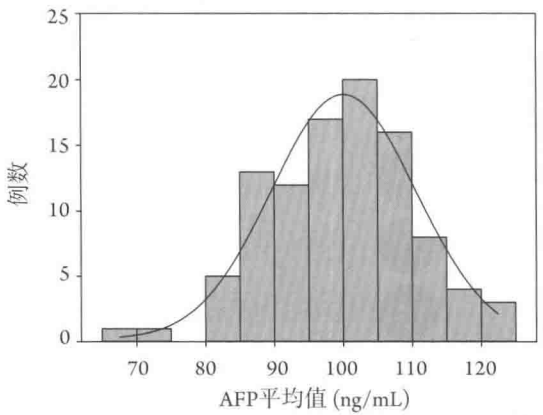


图2 100次抽样所获得的健康个体AFP均值的分布状况

需要注意的是：图2虽然与图1相似，但是绘制原理不同。图1中所有的血钾浓度结果是基于每个个体的检测结果；图2中的每个AFP检测结果则是一次抽样所得到的均值。这n个AFP均值的标准差就叫标准误(SEM)。

可能有的读者读到这里就开始犯愁了。SEM表示多次抽样获得的均数的标准差，因此，如果要获取SEM，就必须做很多次抽样。既然有那个闲工夫做很多次一模一样的抽样，那为啥不把这些研究合在一起，做一个样本量更大的研究？这确实是个问题，好在统计学家帮助我们解决了。

SD和SEM虽然是两个完全不同的概念，但是在数学上有一定的关联，如下所示：

$$SEM = SD / \sqrt{N}$$

简而言之，就是标准误(SEM)等于标准差(SD)除以样本量(N)的0.5次方。这一公式隐含的数学原理不必深究，对于非卫生统计专业人士而言，只需要记住该公式就行了。由上述公式可见，标准误考虑了样本量问题，仅需要知道一次抽样的样本量和得到的标准差，就可以计算其标准误了。样本量足够大时，标准误趋于0，说明样本量越大，结果越精确；样本量越小，标准误越大，说明小样本的研究得出的结果不稳定，不能很好地反映总体。

由标准误又衍生出了另一个概念，即95%可信区间(95%CI)。由图2可知，虽然每次抽样得到的AFP均值不尽相同，但是大致是围绕在100 ng/mL左右的，且均数本身呈正态分布。因此人们提出：均数95%的分布范围就是95%可信区间。注意，注意，注意(重要的事情说三遍)：个体95%的分布范围是参考范围；样本均数95%的分布范围才是95%可信区间！如图3所示：

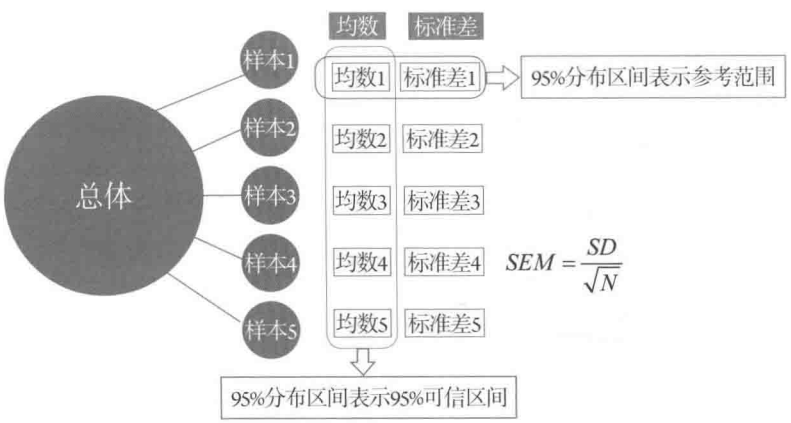


图3　标准差与标准误的联系和区别