

·科学博客·
文/王德华

不理想的数据可以剔除吗？

在与学生讨论实验的数据分析和统计问题时,涉及到实验数据的选择和利用问题。下面是我们师生的交流。

学生的问题 1: 在根据(某参数)测定的数据做的散点图中,有一个数据点出奇得高,结果导致统计分析时没有检测到差异显著性。如果删除这个数据,差异显著性就检测到了(就是说概率值 $P < 0.05$)。学生问能否删掉这个明显高的数据,并且说他已经计算了这个数据,高出平均值(Mean)的 2 倍的标准差(SD)(有些文章建议如果达到这个标准,是可以删除的)。

我的意见:要删除数据,一定要有充足的理由。下面几种情况是可以删除的:

① 如果确定是由于动物本身的问题(如健康等原因)造成的数据偏离,是可以删除的,但是,所有关于这个动物实验的其他参数也都要一起删除。这样做实际上等于我们减少了一个样本。② 如果确定是实验记录等方面出现的错误,如平均值为 50 克的一只老鼠的重量被记录成 250 克,则得出的实验数据可以删除。但是,对于不明原因的“歧异”数据,则应该特别予以关注,并应分析其高或低的原因是什么,是由哪些可能的因素造成的。这样的数据是不能随意删除的,更不能轻易就把它丢弃。在我们的实验中,尽管这个“歧异”数据不能代表整个样本量的变化趋势,但也许会有其他什么生物学的含义,说不定通过“歧异值”还会有新的发现呢。

学生的问题 2: (这只动物)除了这一个参数异常外,其他参数都在正常范围内,能否保留其他参数,而只删除这一个参数呢?因为这个数值实在是很不理想。

我的意见:我们要清楚什么样的数据是理想的数据呢?科学的严格性和客观性,不允许我们按照自己的意愿随意挑选数据。科学要的就是客观,数据展现的和说明的就是这种客观存在。万万不可根据统计分析结果的差异性显著与否,再回过头来挑选数据。为什么说差异显著的数据就是理想的数据呢?达到差异显著的结果



本文作者 王德华,中国科学院动物研究所研究员,中国科学院研究生院教授,兼任中国生态学会动物生态专业委员会主任、《兽类学报》主编。

博客链接 <http://www.sciencenet.cn/u/wangdh/>
栏目主持人 王飞跃,中国科学院自动化研究所研究员。

就是好结果吗?一般说来,数据在统计分析前,需要先检测数据的分布特征(如正态分布),然后根据实验设计的特点选定合适的分析方法,最后检测的结果是什么就是什么。如果规定概率值 $P < 0.05$ 为差异显著性,那么 $P = 0.049$,结论就是差异显著;如果 $P = 0.051$,结论是差异不显著。数学的含义就是这样。但是,在这个问题上又不能太教条,数据的数学含义和生物学含义是要重视的。有些数据由于各种原因,差异没有达到显著性水平,但变化趋势的生物学含义却是很明显的。同样,也有些数据可能已经检测到了差异显著性,但是并没有明显的生物学意义。

实际上,生物学实验中数据偏离平均值较大的现象时常产生,这就是所谓的“歧异值”。关于歧异值的处理,具体起来,生物统计学著作上有一些相关介绍。是保留还是删除,有时候会很伤脑筋。如果确定是记录或实验或动物等等客观原因造成的,当然是可以删除的。如果没有充分的理由,就应该保留这个数据。最好的方法就是重复一下实验,加大样本量,这样就可以很明确确定了。应该注意的是:生物医学研究涉及到药物、病理、药理等等领域,人命关天,必须高度重视,万万不可大意。

科研的一般原则是要严格遵守的:尊重事实,尊重数据。一定要获得准确的数据,获得真实的数据,同时也要正确运用数据,正确解释数据。数据是为结论服务的,结论是以数据为基础的。万不可(也不允许)为了(某个有倾向性的)结论而去人为决定数据的取舍。

根据自己的意愿去修改数据,更是不允许的了,那就属于学术不端了。随意编造数据更是不能容忍,必须坚决杜绝。如果可按照自己的意愿随意改变数据,那就不是科学研究了,而完全是数字游戏了。这样做就是对科学的玷污,毫无疑问,当属坚决打击的学术不端行为之列。

恰巧前几天有英国外宾来访,在研究组学术讨论会(Seminar)上,大家又聊到这个问题。外宾的意见与我的意见是完全一致的。他还补充到:如果选择了一种数据的取舍原则,那么,这个实验室或这个学者的以后的所有数据都要采取这样一致的原则进行处理,这样一个原则也许会伴随你的终生。他强调,不能同样的数据,(出于不同的目的)这次就删除,下次就保留——这样做也是不允许的。他还说,如果你要欺骗读者,要告诉他你是怎么欺骗他的。

~~~~~ 博友热议 ~~~~~

- * 我最初开始做普通物理实验的时候,上理论课的老师曾经说过,只有很明确是实验错误所导致的数据(坏值)允许删除,其他必须原封不动地记录和使用。
- * 只要是从实验中做出来的数字,只要是没有做错,任何一个数字都不应该删除,因为科学本来就有未知的存在,在现阶段可能无法解释它,但不等于这个数字不存在。
- * “剔除”和“删除”还是有点区别的。明白王老师这里说的“剔除”或“删除”是指在发表文章时。“剔除”要慎之又慎,学生切不可自作主张,要和导师认真讨论。必须明确,实验的原始记录决不可“删除”或改动。

(责任编辑 苏青)