· 专题讲座 ·

论中药实验研究数据分析

赫　炎　沈　欣(中国中医研究院中药研究所　北京　100700)

中药实验研究中,无论评价某物质是否为药、其质量是优或劣,还是评价某工艺是否合理,都必须以药效及毒性为基准。作为药品作用在受试体上的生物表达,这是其他研究不可替代的部分。由于生命体内的复杂性,加之实验中诸多不可控制因素的存在,药效及毒副反应表现出不同程度的随机性(例如,同一批制品对同一批相同种属小鼠的毒性反应的不同),因此对这部分研究的分析必须包括一个去除不定性,寻找必然性的信息提炼过程。一般讲,一个合格的实验设计,其观测指标的随机性应尽量受到控制,使得其随机性仅由那些对观测指标作用微小的随机因素综合造成,因此这些指标(或经过适当的数量化后)应能较好地满足一种被称为正态分布的随机变量分布。基于这种分布规律,古典统计学在处理类似药理及毒理实验研究的观测指标(随机变量)中,为我们提供了方差分析(Anova)、t-检验(t test)及卡方-检验(Chi-square test)分析方法。应该指出,虽然近代数理统计及其相关的信息处理技术有了长足的进展,但这些方法仍是数据处理中不可替代的基础。下面结合一些实例说明这方面应用中较易出现的问题,帮助正确体会其基本思想。以助正确提炼中药实验研究所获的数据信息。

1　对统计显著性检验的理解

例 1. 在胆南星混合法与发酵法工艺比较研究中,采用对戊巴比妥钠催眠的增效作用指标,通过药物对中枢神经系统的抑制作用,评价其息风定惊作用。在对不同制品的醇提物的药效比较研究的预试验中,获得了如下实验数据(表 1):

表 1　不同制品醇提取物药效作用(腹腔给药)

组别	睡眠数(只)
混合法	2
发酵法	3

$n=5$

表 2　不同制品醇提取物药效作用(腹腔给药)

组别	睡眠数(只)
混合法	8
发酵法	4

$n=10$

很显然,就表 1 这批数据,我们找不到混合法与发酵法制品的睡眠率有差异的证据。那么能否就此得出两种工艺制品对戊巴比妥钠催眠的增效作用相同的结论吗?显然就此数据作肯定结论还为时过早。事实上,同样的实验,通过增加各组例数从 5 到 10 后,我们获得了表 2 的数据。对其进行统计分析,我们可以推翻两种制品作用相同这一假设($P<0.01$)。通过上述对比,不难看出在其它条件一致的情况下,增加实验例数,有时会获得不同结果。那么究竟哪一种结果更可靠呢?从肯定问题的角度(即肯定两种工艺制品药效作用相同),根据实验例数多的表 2 数据所得结果的应该更可靠些。但是从否定问题的角度,仍是实验例数越多,所得结果应该更可靠些吗?这就提出了结果可靠性与实验例数的关系问题。为了说明此问题,我们不妨作如下假设:

表 3　不同制品醇提物药效作用(腹腔给药)

组别	睡眠数(只)
混合法	4
发酵法	1

$n=5$

表 4　不同制品醇提物药效作用(腹腔给药)

组别	睡眠数(只)
混合法	8
发酵法	7

$n=10$

当每组例数为 5 时(表 3),两种工艺制品药效比较 $P<0.01$,统计检验获得了否定具有相同药效作用的证据;而每组例数为 10 时(表 4),$P>0.05$,统计检验未获得到否定具有相同作用的证据(这种情况发生的可能性虽小,但不能排除)。一般认为,表 4 的结论更可靠些(例数多),但事实上表 3、4 两种结果并不矛盾,仅是意义不同。那么那种结论更可靠呢"答案应是表 3 中的结论($P<0.01$)。因为:根据表 3 实验结果,在 $P>(1-0.01)$的把握下,我们可以推翻两种制品作用相同的假设。而对表 4 的结果,即便考察了比表 3 更多的例数,但由于没有获得统计检验上的证据,不能否定作用相同的假设;同时在没有对其样本的总体作出充分的考察时,我们也不能得出两种工艺制品作用相同的结论。那么为什么样本例数少的结果此时反倒可取呢?这里涉及到对通常所说的实验阴性与阳性结果的理解问题。统计显著性检验是以小概率事件为反例,推翻没有差异的假设

的检验。正如我们所知,从逻辑上,肯定一事物无论从时间上,还是从空间上均有待于广泛的验证,反之若想否定一事物,只要找到其反例就足以对其否定。这就是在现实生活中人们常体会的被承认难,而被否定易的道理之所在。例如在生物实验中,阴性结果作为对某一结论的肯定(例如某药的药效与另一味药的药效相同)的证据往往是不充分的,甚至是没有意义的。同样在统计检验上,在没有获得 $P<0.05$ 的小概率事件的反例时,即便在例数足够多的情况下,我们也很难接受没有差异的结论。反之,即便例数不多,一经找到了 $P<0.05$ 的小概率事件,就已经足以以此为反例推翻没有差异的假设,只是这种反例的获得,无论从经验还是从理论上都证明样本例数越多越容易获得。因此,对上述问题如不能正确理解,那种认为例数越多越可靠的经验就会干扰我们作出正确决策。对事实存在差异,通过实验分析又能检测出这种差异的能力,统计学用检验功效来表达,即(1−β),其中 β 代表事实上有差异而一次性实验又得不到 $P<0.05$ 反例的概率(犯第二类错误的概率)。由于实验设计的目的就是要找出反例以推翻没有差异的假设,其具体安排也就应该围绕提高检验功效来完成。检验功效除了与样本例数在一定的范围内有近似正比例的关系外,还与观测指标的离散程度、样本的可比性有关。即观测指标的标准差越大检验功效越小;对比的两组例数越接近及标准差越接近,可比性越大,检验功效也就越大;但更重要的是取决于对比的两组观测指标本身的内在差异。对这些因素的考虑,以及对实验对象选取的代表性的考虑基本上就构成了所谓实验设计的内容,其目标是要尽量提高检验功效。

2 正确使用统计分析,避免人为造成假阳性结果

表 5 不同中药药效作用比较研究

组别	药效指标 ($\bar{x}\pm s$)	P 值	
		与对照比	两药比
对照	83.2±14.4		
甲药	70.6±12.2	<0.05	
乙药	58.8±13.4	<0.01	<0.05

$n=10$

这是一个可以经常看到的实验研究报告形式,也是在国内一般杂志上较为常见的错误,即用 t 检验直接进行组间的两两比较。上例经如此分析后,三种比较 P 值均小于 0.05,进而得出甲药和乙药组药效均显著低于对照组,乙药组药效显著低于甲药组。作为两组药物的有效性,同时乙药优于甲药的实验证据,上述的结论的第一类错误概率实际是否就低于 0.05 呢?或者说我们能保证上述的结论中至少一个是错误的概率低于 0.05 吗?我们知道,显著性检验是以在一次性实验(或比较)中,出现小概率事件为反例推翻没有差异假设的检验。然而这种 $P<0.05$ 的小概率事件在多次重复或者在多次性实验(或事件)中,其发生的可能性会大大增加,因而降低了显著性检验结论的可靠性。在上面的实验中,每一次检验所得结论的第一类错误概率的确是不高于 0.05,但是,根据概率加法定律,三个结论至少其中一个是错误的概率为不高于 0.11,而并非 $P<0.05$。上面这个例子说明了把重复 3 次事件中所出现的小概率事件,错误地作为一次性实验结果,会得出不恰当的结论。再例如图 1 所述数据,采用 t 检验进行了 3 次各组之间的比较,结果:对照与老法比,$P<0.05$;对照与新法比,$P>0.05$;老法与新法比,$P>0.05$。显然以此统计结果得出的结论是矛盾的。出现这种矛盾的原因是由于将 3 次重复的试验事件作为了一次随机事件处理,造成了对照与老法比较结果的假阳性错误概率增大(实际 P 值增大)。因此,当 t 检验用于多于两组的均值比较时,我们应相应采用一些对上述问题有所考虑的统计方法如 Bonferroni t 检验和 q 检验(Student-Neuman-Keuls test)等。下面仅介绍较为简便的 Bonferroni t 检验。根据 Bonferroni 不等式 $P<\alpha/K$(其中 α 为设定的假阳性错误出现的事前概率限定百分值,通常取 5 或 1;K 为比较次数;P 为实际查找 $t_{0.05}$ 时所采用的 P 值),在上面图 1 的例子中,$\alpha=5$,$K=3$,$P<1.66$;应使用 $t_{0.01,6,18}=2.47$ 为 t 界值,代替 $t_{0.05,18}=2.1$。在老法与对照对比中,计算得 $t=2.39<2.47$;$P>0.05$,因此就避免了上面的矛盾结果。这一点通过方差分析,$F=2.74<F_{0.05,27,2}$ 也可证实。

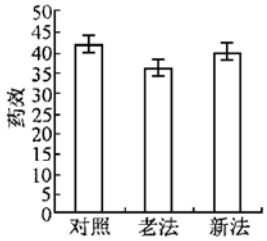

图 1 两种工艺的药效比较研究

3 灵活选用对照、降低个体差异提高统计功效

在药效实验中,有时所测得药效指标离散程度较大,受试因素可能引起的药效差异会被指标本身

的变异所遮盖，因此，在有限的实验例数下，还无法检测出药效差异是由受试因素引起，还是由受试对象个体差异及实验误差所致。或者说由于过高地估计了受试因素可能引起的药效反应的变异，而降低了统计检验功效。如果此时能够通过一定的实验设计，将施加实验因素前就存在的个体差异从药效指标的总标准差中分离除去，从而压缩其药效反应的变异(标准差)，就可以大大提高统计检验功效，使得在同样或更少的样本例数下，减少假阴性结果出现。

例如在研究某中药对血小板凝集作用的临床实验中，研究者对正常人通过对照和给药组的受试对象的测定，运用 t 检验分析，没有获得给药可明显降低人体最大血小板凝集率的证据($P>0.05$)。考虑到血小板凝集率本身个体差异较大，可将原设计的对照组改为自身对照，观察同一组受试对象的给药前后血小板凝集率的差异，这样就排除了受试者在给药前就存在的个体差异。如图 2 所示：设给药前后差异为 d，则均值 $\overline{d}=10.3$，$S_d=8.0$，$t=\dfrac{\overline{d}}{S_d/\sqrt{n}}=\dfrac{10.3}{2.41}=4.27$，$P<0.01$，说明该药对血小板凝集率有显著影响。那么，为什么两种不同设计会造成如此差异呢？原因一方面是给药组与正常对照组实验对象之间的血小板凝集率在施加实验因素前就存在差异(抽样误差)；另一方面在组间比较过程中，估计标准差既包含血小板凝集率本身的个体差异之和，还包含有个体对药物反应的差异，由于组间设计不能将二者分开，以致估计标准差很大。而自身对照设计能够分离和除去本身个体差异，使得估计标准差更接近于药物对血小板凝集率实际影响的变异，因而可大大提高检验功效。这一点从图 2 中可以看到：虽然给药前后的均值相差不大，但多数病例均显示了下降的趋势。为了说明道理，我们假设在开始的实验中，给药组与正常对照组之间的抽样误差已控制在很小范围，加之给药因素对其的影响，两组同时获得的数据从组体的角度(均数和标准差)应接近自身对照实验的给药前后数据(此时我们不妨假定一样)，即给药前后血小板凝集率的均值及标准差分别为：$\overline{X}=53.5$，$S_X=18.7$；$\overline{X}=43.1$，$S_X=15.9$；计算得 $t=1.38$，$P>0.05$。而采用自身对照的 t 检验(dependent t test)，$t=\dfrac{\overline{d}}{S_d/\sqrt{n}}=\dfrac{10.3}{2.41}=4.27$，$P<0.01$，获得了该药可明显降低人体最大血小板凝集率的实验证据。因此在检测指标本身个体差异较大时，自身对照的优点往往是突出的。在实际应用中，还可将图 2 中的前后比较扩展到多次测定比较(repeated measurements in the same subject，见图 3)，这种统计分析中所用到的方差分析，由于其设计特点，已形成一类专门讨论的统计分析方面，限于篇幅，在此不多赘言。本文在此仅给出一些常见例子，如在降血糖和降血脂药效研究中，由于血糖和血脂的个体差异很大，加之药效往往需要长时间的动态观察，实验周期长，采用组间比较，即便在样本例数较大时，检验功效仍较小，很难获得阳性结果。此时若用自身对照，利用动态数据，往往可获得事半功倍的效果。

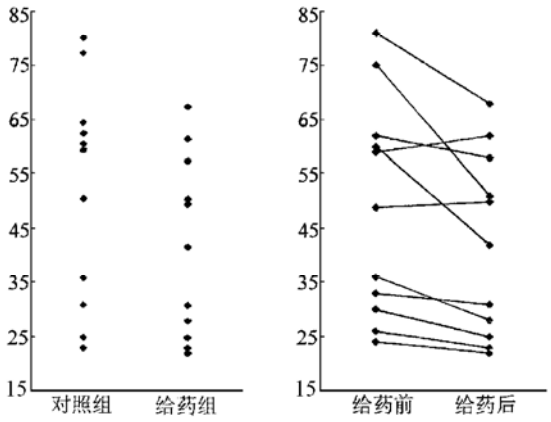

图 2　某中药对人体最大血小板凝集率的影响

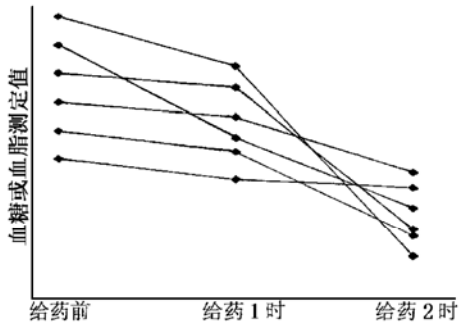

图 3　对每个受试者多次测定数据

4　小结

本文对中药生物实验的数据处理的基本思想以及相关的常见错误作了介绍。在具体应用中，实验研究人员所采用的统计量(F，t，χ^2)根据具体情况可能不同。但其显著性检验的实质是一致的。应当指出，近年来随着计算机应用的发展，各种数据处理的软件及相应实验设计已经很完备，因此，具体应用技术相对已不重要。但对实验设计及结论可靠性的评价却越显重要。作者基于多年实验工作的经历及遇到和见到的统计误用，对其略谈认识。

(收稿：1998-09-21)