讲座

简明生物统计方法

王广仪
（吉林医科大学）

这是为初学者写的普及性讲座。为了易懂和实用，叙述力求浅显，多用实例，并着重于介绍常用的统计推断方法，而不过多的论及数理统计原理。作者根据统计方法制作出若干图表，可供查用，以减少繁冗的计算。同时，作者还介绍了自己提出的一些简便的新办法。对于这种尝试，欢迎读者提出意见。

简明生物统计方法全文共分四讲：一、概述；二、计数资料的统计推断方法；三、计量资料的统计推断方法；四、回归与相关。从本期起依次连载。

——编　者

第一讲　概　述

一、生物科学实验与统计方法

生物科学实验所以需要应用统计方法，是由实验对象的特点决定的。

生物科学研究实验对象的特点是具有比较大的变异性，例如，选择同种属、同性别、同年龄、同体重的动物，并尽可能控制在相同的条件下进行某种药物毒性实验时，各受试动物的中毒反应并不相同：有的死亡，有的生存，中毒程度轻重不一。又如，用同样疗法或同样药物对于同类病症，在几乎同样条件下进行治疗观察，其结果也往往不尽相同：有的有效，有的效果不大，有的无效，甚或恶化。这些都说明有变异性。那末变异性现象中有没有规律性呢？有。只有通过大量对个别现象的观察才能揭示出来。例如，观察某种疗法或药物的疗效时，只就个别病人来看，就很难据以做出判断；但如治疗观察的例数达到相当大量之后，则治愈者占全部受治者的百分率（称治愈率），却是相当稳定的，足以说明其疗效之高低。对于具有变异性的现象，通过大量观察，然后计算出一些比较稳定的统计数值，用以描述研究对象固有规律性的方法，称"描述统计"，属于大样本统计方法，在生物科学实验研究中早已广为应用。

然而在实际研究工作中，往往因受各种条件限制，不允许我们进行大量观测，所能取得的数据资料有时是比较少的。例如在实验室做动物实验，不是经常一次都能观测几百只或几千只动物；而是十几只或几十只。临床治疗观察也如此，例数有限。通过少数例数的观测结果来推测全部研究对象的规律，其结果自然不如例数多那样稳定，会出现一定波动，这是个矛盾，这在生物科学研究中普遍存在。为解决这类问题，近年来广泛应用了数理统计方法，即所谓小样本统计方法（在某种意义上也可说是推测统计方法）。

关于采用小样本进行观测的原因，一方面由于受到各种条件限制，客观上不允许大量观测；另一方面是从提高观测结果的精确性本身考虑的。因为在理论上讲观测例数越多，结果越趋于稳定，越能反映研究对象的固有规律性，而实际上并非如此。因为有的生物现象观测例数越多，条件越不易控制一致，越容易出现偏差。如果一次观测的例数虽然少些，但在选择对象与控制实验条件等方面都能更严密些，再加上对观测结果用数理统计方法加以处理分析，结果会更精确些。总之，在能够应用数理统计方法加以推测的基础上，观测例数不必太多，能达到研究目的即可，盲目地进行大量观测效果不一定就好。

生物统计方法包括实验设计和统计推断两部分内容。统计推断主要研究如何根据部分资料对全部研究对象进行科学的推测，从而使研究结论由局部到全体，由特殊到一般，能更具普遍意义。当然，要做出正确的统计推断，事先必须有合理的实验设计。为便于理解统计推断的特点，首先介绍几个基本概念。

二、几个基本概念

（一）总体和样本 统计观测分全面观测与部分观测(抽样观测)。如果实际观测的范围和研究结论所要包括的范围完全（或基本上）一致，为全面观测。这时不存在总体与样本问题。只有当结论所要说明的范围远远超过实际观测的范围时，才出现总体和样本问题；把结论所要说明的全部对象称为总体，实际加以观测的部分对象称为样本。如：

例 1. 某地区为摸清本地区慢性气管炎患病情况，曾在全地区进行普查，发现全区 127,270 人口中有 6,963 名慢性气管炎患者，占 5.5%（称患病率）。结论是某地区慢性气管炎患病率为 5.5%。本例因调查结论所要说明的范围和实际调查的范围一致，故不存在总体与样本问题。

例 2. 某生物制品所为要了解所生产的某批某种活菌苗里活菌所占百分率，培养观测了 200 个杆菌，结果有 77 个活菌，占 38.5%（称活菌率）。结论是某批某种活菌苗的活菌率为 38.5%。很明显，本例结论所要说明的范围是某批菌苗里的全部杆菌，而实际培养观测的 200 个杆菌只是其中很小一部分，所以存在总体与样本问题。

关于总体和样本概括地说就是：根据研究目的确定的，符合指定条件的全部观测对象称为统计总体，简称总体。凡实际观测的范围小于总体所包括的范围的，统称样本。样本有大有小。

（二）抽样误差与随机抽样 样本与总体之间的差异，称抽样误差。抽样误差的特点有二：一是如果其它条件不变，则样本越大，抽样误差越小；如果样本大到等于或接近总体时，则抽样误差为零或极小。二是如果样本含量不变，则观测对象的变异程度越小，抽样误差也越小；如果没有变异时，抽样误差为零。当然，没有变异的现象也不必用统计方法加以研究。

在实验观测中造成数据出现差异的因素，除生物个体间固有的差异性（可通过实验设计缩小个体差异）之外，还包括一些其它因素，如实验误差等等。除抽样误差以外的所有误差统称非抽样误差。非抽样误差不在统计学讨论范围之内，但它不仅直接影响抽样误差的大小，而且如混有系统误差（偏向一方的误差），还能使统计结果造成错觉，这是值得注意的。

凡属抽样结果都带有抽样误差，而抽取样本的目的又在于推测总体，可见对抽样误差的处理是统计推断的核心问题。然而，并不是任何一种抽样所产生的抽样误差都能用数理统计方法处理。只有随机抽样的抽样误差，才有一定规律可循，才能加以处理。

随机抽样，如用数理统计的说法就是：凡在抽取样本时使构成总体的每个个体都能有同等机会被抽取到样本中去。随机抽取的样本称随机样本，简称样本。一般所说的样本实际上都指随机样本。因为只有随机样本对总体才有代表性，才能用来推测总体，否则便失去抽样研究的意义。那末究竟怎样才算随机抽样？打个比喻，就象从米袋里任意抓出一把米，假如袋内的米混合得非常均匀，则不论从袋内那一部分抓出的一把米都是随机样本，对全袋米都有代表性，然而，假如袋内的米不是混合得非常匀，而是上下质量不同，这时无论从袋内那一部分抓出的一把米，都不是随机样本，也都没有代表性。

随机抽样的抽样误差，虽然也具有上述一般抽样误差的两个特点，但不同的是它有一定规律性。统计推断方法所以能根据样本推测总体，就是建立在有规律性的基础上的。因此非随机抽样，不能应用统计推断方法，这是值得特别强调的。

（三）概率与概率分布 概率论是数理统计的基础。

所谓概率就是对可能性的一种定量表示。譬如说今天"八成"要下雨，就是说今天下雨的可能性（概率）有 80%，或说象今天这样天气十有八、九是要下雨的。当然，象这样对可能性的定量说法是极其粗略的，缺乏周密的统计和计算，只是根据过去的经验得出的结论，这必然因人而异。但毕竟是一种对可能性的定量说法，只是不如数理统计那样准确而已。

概率显然都是根据统计结果得来的，然而严格说来，统计得来的结果实际上都是频率，并非概率。只有当统计的数量相当增大之后，频率便开始稳定在一个常数周围，这个常数能反映事件出现可能性的大小，便称它为事件的概率。例如，某批菌苗里的活菌率究竟是多少？我们无从知道。但毕竟有个数。随机抽取 200 个杆菌培养，观察的结果，活菌率为 38.5%是频率，是有波动的。但假如随机抽取的样品数逐渐增大，则得到的活菌率便围绕一个常数摆动，此常数即某批菌苗的活菌率，即所说的概率。

假如从某批菌苗里每次随机抽取的样本比较少，则每个样本的活菌率（频率）的波动虽较大，但如果样本含量不变，随机抽取的样本数达到相当量之后，样本活菌率的分布（频率分布）并非杂乱无章，而是有一定规律的（见下面模拟抽样实验结果）。

假定某批菌苗的活菌率为 50%，每次随机抽取 10 个杆菌为一个样本，共抽取 1024 个样本的活菌率分布情况如表 1。

由表 1 可见，频率分布非常接近概率分布。因此概率分布也可说是理论频率分布。这就是随机抽样分布的规律性，也就是随机抽样误差的规律性。

概率分布可根据概率模型直接计算，无需进行大量重复抽样。因此，只须事先确定抽样的概率模型（数学模型），便可根据样本结果及其概率分布进行统计推

表 1　1024 个样本活菌率的分布 ($n = 10$)

活菌率 (%)	样本数	频率分布 (%)	概率分布 (%)
0	2	0.2	0.1
10	13	1.3	1.0
20	42	4.1	4.4
30	125	12.2	11.7
40	205	20.0	20.5
50	257	25.1	24.6
60	213	20.8	20.5
70	119	11.6	11.7
80	39	3.8	4.4
90	8	0.8	1.0
100	1	0.1	0.1
合计	1024	100.0	100.0

断。常见的概率分布有以下几种：

二项分布　二项分布是一种离散型分布，如上述抽样实验结果便属二项分布。假设某批菌苗的活菌率（概率）为 p，则死菌率必为 $q = 1 - p$。且 $p + q = 1$。其概率分布可用下式计算：

$$f(x) = C_n^x p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \quad (x = 0, 1, 2, \cdots, n)$$

n——样本含量；

x——样本中活菌数。

上式中 $\frac{n!}{x!(n-x)!}$ 即二项系数。如把每个样本中的活菌数一一代入上式，即得概率分布。如当 $x = 0$（即 10 个杆菌全为死菌）的概率为：

$$f(0) = \frac{10!}{0!10!}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{10} = \frac{1}{1024}$$

余可类推。

凡分成对立两组（如死、活，阳性、阴性，治愈、未治愈，等等）的计数资料，一般皆服从二项分布。

正态分布　正态分布是一种连续型分布。如果样本含量逐渐增大，则二项分布便趋近正态分布。二项分布的极限是正态分布（许多分布的极限都是正态分布）。实用上当 $n > 50$，且 p 不太偏向 0 或 1 时，把二项分布的数据按正态分布处理已足够精确了。有不少原始测量数据本身成正态分布（如身高、胸围及其它生理指标测定值等）。从成正态分布的总体里随机抽取的样本平均数也成正态分布。如果样本含量较大（$n > 30$），不管原总体是否成正态分布，则由样本平均数构成的统计量

$$Z = \frac{\bar{x} - \mu_H}{\sigma_{\bar{x}}}$$

$\bar{x}$——样本平均数；

μ_H——总体平均数；

$\sigma_{\bar{x}}$——根据总体标准差计算的标准差

仍趋近正态分布。这点很有实用意义，即当 $n > 30$ 时，按正态分布处理已足够精确了。

正态分布的概率可用下式求出：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty, \sigma > 0$$

μ——总体平均数；

σ——总体标准差；

x——可为任意数。如 x 为样本平均数（用 $\bar{x}$ 表示）时，则 σ 用标准误差（用 $\sigma_{\bar{x}}$ 表示）代替。

t 分布　t 又称“Student t”，当总体标准差 σ 属于未知，而用样本标准差 S 代替 σ 时，统计量

$$t = \frac{\bar{x} - \mu}{S_{\bar{x}}}$$

$S_{\bar{x}}$——根据样本标准差计算的标准误差

便成 t 分布。其概率分布的计算公式为：

$$f(t) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)\sqrt{\pi df}} \cdot \frac{1}{\left(1 + \frac{t^2}{df}\right)^{1/2(df+1)}}$$

df——自由度，当标准误差 $S_{\bar{x}}$ 是由单个样本得出时，则 $df = n - 1$。

t 分布的极限也是正态分布。实用上当 $n > 30$ 时，t 分布和正态分布的结果便差不多。但 $n < 30$ 时，t 分布要求原始变量（数据）成正态分布，应用时要注意。

泊松分布　当样本含量相当大，且事件出现的概率又很小（所谓稀有事件）时，二项分布就变化成泊松分布。

以上对几种常见概率分布的简介，目的在于对概率分布有个一般概念，运用时备有统计数值表，无需现去计算，只需知道所要处理的样本，基本上服从那种分布即可。以下将要讨论的关于统计推断的概率，无论是区间估计的置信度或者显著性测验的 P 值，都是按某种概率模型的概率分布公式算得的，虽然比较精细，但有一点应该反复强调的就是想要生物统计结果可信，则力求使生物实验数据符合数学模型所要求的条件。否则，既使统计分析得出错误结论也无从查对。统计分析只是一种假定实验数据符合（或基本符合）某种概率模型条件下的数学加工，条件不符，统计分析反而容易造成谬误。

（四）统计推断　用样本推测总体统称统计推断。但因有抽样误差的影响，推测时难免有一定出入，所以

存在推断的精确性问题；统计推断既要保证一定的推断的可靠性(确度)，同时也要指明“出入”的大小(精度)。

根据样本推断总体大致可分为下列二种情况：一是用样本结果(称样本统计量)估计总体的结果(称总体参数)；二是根据样本之间的差异推断其所代表的总体之间是否也有一定差异。把第一种情况的推测称“参数估计”，第二种情况称“统计检验”或“差异显著性测验”。

统计推断的概念和处理手法是初学者常常感到困难的地方，从应用的角度来看虽然无须过多探讨数理统计的原理，但至少得弄清楚一些最基本的问题，否则就不能正确运用。

首先介绍一下关于参数估计问题。有两种估计法：一种是点估计法，另一种是区间估计法。这两种估计法的名称乍一听来可能感到陌生，但日常生活中确有类似的用法。譬如估计某人的年龄有多大，就有两种估计法：“某人可能有33岁”；或者“大概三、四十岁吧！”估计33岁的便属于点估计法(因33是个点值)，而估计有三、四十岁的便是用“30—40”岁这个范围(称区间)进行估计，属于“区间估计法”。点估计法固然很具体，即估计的精密性高，但对的把握不大，即估计的可靠性差。区间估计法虽不够精密，但估计比较有把握，一般不易出错。可见在“估计”这个问题上，精密性和可靠性是互相制约的。为保证估计得更有把握些，往往需要降低精密性，如估计“某人大概有三、四十岁”，或“大概在30到35岁之间”，或许说“至少有30岁”或“至多不超过35岁”等等。

用样本统计量估计总体参数也是同样的道理，如用样本活菌率38.5%直接作为总体活菌率的估计值，便是点估计法。如果根据38.5%制定一个区间而估计总体活菌率在此区间之内，便属区间估计法。问题是用多大区间进行估计。当然，区间越大，估计的可靠性越高。但区间越大精密性越小，要提高精密性就得相应地降低可靠性。但估计的可靠性在统计上规定一般不得低于百分之九十五。通常用两种估计区间：估计的可靠性为95%的估计区间(称95%的置信区间，95%为置信度)；估计的可靠性为99%的估计区间(称99%的置信区间，99%为置信度)。置信区间又称置信节或可信限。例如根据样本活菌率38.5%求出的95%与99%的置信区间为：“31.6%—45.6%”与“29.6%—47.8%”，即估计全批菌苗的活菌率在31.6%到45.6%之间的可靠性为95%；而估计在29.6%与47.8%之间的可靠性为99%。后者估计的可靠性略高些，但区间也稍宽些，即估计的精密性稍低些。所说可靠性95%或99%是指作这样估计时平均每100次能有95次或99次正确，只有5次或1次估错。并非指作某一次具体估计时，有95%或99%是正确的，有5%或1%是错误的。因为对于某次具体估计，对就对了，错就错了，不存在百分之几正确和百分之几错误的问题。这是两种截然不同的概念，切勿混淆。

一般认为对于理论方面的研究(如基础医学等)应该用99%的置信区间；对于应用方面(如临床上)只需用95%的置信区间。但也应该根据具体情况而定。例如估计某种疗法或药物的疗效时，用95%的置信区间或只指出区间的“下限”也就够了，如说“满山红酒”对老年慢性气管炎的有效率在63.4%(95%的置信区间下限)以上；而估计某种药物毒性作用时，则宜用99%的置信区间或着重指出区间的“上限”，如说“醋酸铊”内服治疗头癣时的中毒反应可高达34.9%(99%的置信区间“上限”)。

至于说用95%或99%的置信区间进行估计时，平均每100次尚有5或1次估错(即总体参数不在此区间之内)的危险性，这对于只做一次具体估计来说，可认为是小到几乎不能遇到的程度，已经是十拿九稳，很有把握的了。但不能为了提高精密度(即缩小估计区间)而用置信度低于95%的估计区间进行估计。因为置信度小于95%，便不能保证实验结果的“重现性”，即精密度虽高却不能令人置信，价值也不大。在不降低必要的置信度前提下来提高精密度，可采用适当增加样本含量或缩小个体差异的办法，这在实验设计时应予考虑。关于置信区间的具体求法详后。

以上讨论的是从单个总体里随机抽取一个样本时的统计推断，即参数估计问题。下面介绍从两个(或多个)总体里，随机抽取两个(或多个)样本时的统计推断，也就是差异显著性测验问题。

差异显著性测验，就是根据样本之间的差异程度推断其所代表的总体之间是否也有一定差异。如有，就是“差异显著”；否则就是“差异不显著”。例如：从某生物制品所先后生产的两批某种活菌苗里随机抽取的两个样本之间活菌率相差5.5%，能否认为两批菌苗的活菌率之间也有一定差异？不能。不能贸然做出这种结论的原因是因为有抽样误差影响。就是说，即或两批菌苗之间本无差异，然而从这样两批菌苗里随机抽取两个样本时，有如从同一批菌苗里抽取两个样本一样，纯粹由于抽样误差的影响，两个样本活菌率之间也会出现某些差异。当然，如果两批菌之间本来就有一定差异时，则随机抽取的两个样本活菌率就更容易出现差异了，只是后一种差异和前一种混在一起，为前者所掩盖，不易直接辨认出来。总之，样本之间的差异可能出于两种情况：一是总体之间本无差异，样本之差是抽样误差造成，纯属偶然现象；二是总体之间确有一定差异，样本之差除抽样误差影响之外，还反映了总体间之差，并非完全出于偶然。

现在的问题是，在实际工作中需要根据两个样本活菌率之差，推断两批菌苗的活菌率之间是否真有差

异。事实上，只要两批菌苗里的杆菌未被全部培养观察，这问题就无法彻底弄明白。差异显著性测验只是在一定假设前提下，利用概率分布原理设法相对地排除第一种情况，从而断定为第二种情况，以得出两批菌苗的活菌率之间亦有一定差异的结论。所谓相对地排除，是指只要用显著性测验方法测得属于第一种情况的可能性(概率——用 P 表示)小于 5%（即 $P<5\%$），便认为"差异显著"，从而排除第一种情况，推断为第二种情况，即结论为两批菌苗的活菌率之间有一定差异。但不应忘记，这种结论是在仍有 5% 的属于第一种情况的可能性条件下做出的，因而其可靠程度只有95%。为提高推断的可靠性，只有当 $P<1\%$ 时才认为"差异显著"，这时做出的结论便有 99% 的把握。

从提高推断的可靠性角度考虑，当然"差异显著"的水准订得越高(即 P 值越小)越好。但即或在 P 值小于千分之一或万分之一的情况下做出的推断，也只是一种推测，推断与总体之间并不是绝对没有一点差异了；只是这种情况下做出的推断搞错的危险性很小，小到平均每千次或万次中只有一次错误。然而在使 P 值达到这样小以前，轻率地认为推断万无一失，势必要把已经反映出总体之间有一定差异的样本之间的差异，当做抽样误差而忽略过去，这对研究本身是非常不利的。可见那种"万无一失"的想法在统计学中是不合适的。显著性测验通常(特殊情况例外)只采用二个标准：5%与 1%。如果检验标准定为5%，而显著性测验结果指出 $P>5\%$ 时，便认为"差异不显著"，即不能排除抽样误差的影响，而做出总体之间也有一定差异的结论。这里所说"差异不显著"并不意味着总体之间就无差异了；只是说如果 P 值大于预定标准，还勉强做出有差异的推断时，就达不到预期的可靠性。因此，与其做出这种根据不足的结论，不如暂不做出有差异的结论而再继续观测下去。

在数理统计方法书上都提到统计推断中的两类错误问题。一类是指在进行显著性测验时，把纯属抽样误差造成的样本之间的差异，误作总体之间真有差异的反映，做出"差异显著"的结论；这叫第一类错误(或 I 型误差)。I 型误差至多不超过 P 值的标准（如 5% 或 1%）。第二类错误是指总体之间真有差异，但显著性测验时 P 值未能达到规定的标准，致使把"真差"误作"假差"，这类错误称 II 型误差，一般不能测出。两类误差的关系是相互制约的；缩小 I 型误差必然增大 II 型误差，反之亦然。进行显著性测验时，因为很容易犯上述两类错误中的一种，为了防止发生错误，一般只把住犯第一类错误的关，使其不超过 5% 或 1%；超过时，则宁肯暂不做出结论而继续进行观测。

关于统计结论和研究结论的关系，可以说统计结论是研究结论的依据，但二者并非一回事，统计结论不能代替研究结论。譬如显著性测验指出 $P<5\%$（或 1%），认为差异显著，这是指根据实验数据得出的统计结论，即所谓有"统计学意义"；至于对研究本身意义如何，尚须根据专业知识作全面分析。

关于差异显著性测验的具体方法，根据实验数据的性质不同虽有多种多样，但其意义与处理手法不外乎如上所述，即设法在一定假设条件下按照某种概率分布求出 P 值，然后根据 P 值大小结合既定标准（5% 或 1% 等）进行统计推断。求 P 值的各种具体方法详后。

（五）计数资料与计量资料 统计资料（实验数据）基本上可分为两大类：一类属于计数资料；另一类属于计量资料(或称测量资料)。

对于构成总体的每个统计单位，凡用定性的结果表示的，如化验结果的阳性或阴性，治疗结果的有效或无效，血型分类属于那一型等等，所获得的数据必然是计"个数"的，称为计数资料。上述例 1 与例 2 皆为计数资料。

对于构成总体的每个统计单位，凡用定量的结果表示的，例如每个人的身高、体重等测量结果，每个人都有一个具体的数值，这类资料称为计量(或测量)资料。

三、小结

（一）在实验(或调查)研究中，首先必须明确实际加以观测的对象是总体还是样本，如果是样本，则必须应用统计推断方法加以处理。

（二）进行抽样观测时，事先对所抽样的总体必须有明确规定，然后再按照对总体的规定条件和选样方案严格选取样本，以保证样本对总体的代表性。

（三）随机抽样是统计推断的基础，不是随机样本便没有推断的依据，就失去抽样研究的意义。

（四）统计推断的可靠性，一般不能低于 95%，无论用样本估计总体（参数估计），还是统计检验都是如此。只是前者从正面指出推测的可靠性（置信度）为 95%(或99%)；后者从反面指出推断错误的危险性为 5%(或1%)，实际是一个意思。

（五）统计结论不能代替研究结论，必须根据专业知识进行全面分析。

参 考 资 料

[1] 中国科学院数学研究所统计组编：常用数理统计方法，科学出版社，1973。

[2] 吉林医科大学：医学统计方法讲义，1973。

[3] Snedecor, G. W.: Statistical methods applied to experiments in agricalture and biology. Jowa state College Press, Amer., Jowa, 1959.

（待续）