样本及其统计量性质对临床医学研究统计方法恰当选择的影响

吴俊林[1]，陈 霞[1]，戴浩然[1]，唐雪莉[1]，黄艳君[1]，黄国平[1,2*]

（1. 四川省精神卫生中心·绵阳市第三人民医院，四川 绵阳 621000；
2. 川北医学院精神卫生学院，四川 南充 637000
*通信作者：黄国平，E-mail：achuanggp@163.com）

【摘要】 医学研究中，影响统计方法恰当选择的因素较多，样本及其统计量可能是两个不可忽略而又容易被忽略的因素。本文详细介绍了样本随机性来源、样本获取方法、独立同分布样本、统计量及抽样分布等内容，探讨统计方法的应用应当满足哪些条件以及为什么需要满足这些条件，以期提高医学论文中统计学内容的科学性。

【关键词】 医学论文；统计方法；样本；统计量；抽样分布；选择

中图分类号：R195.1 **文献标识码**：A doi：10.11886/scjsws20201204001

Impact of sample and its statistical properties on the appropriate choice of statistical methods in medical research

Wu Junlin[1], *Chen Xia*[1], *Dai Haoran*[1], *Tang Xueli*[1], *Huang Yanjun*[1], *Huang Guoping*[1,2*]

（1. *Sichuan Mental Health Center·The Third Hospital of Mianyang*, *Mianyang* 621000, *China*;
2. *Mental Health School of North Sichuan Medical College*, *Nanchong* 637000, *China*
**Corresponding author*: *Huang Guoping*, *E-mail*: *achuanggp*@163.*com*）

【**Abstract**】 In medical research, there are many factors that affect the proper selection of statistical methods. Samples and their statistics may be two factors that cannot be ignored but are easily ignored. This article details the source of sample randomness, sample acquisition methods, independent identically distributed samples, statistics, sampling distribution and so on, in order to explore which conditions should be met for the application of statistical methods and why these conditions should be met, so as to improve the scientific nature of medical papers.

【**Keywords**】 Medical paper; Statistical analysis; Sample; Statistic; Sampling distribution; Choice

在医学论文中，误用甚至滥用统计方法的现象仍较普遍[1-2]，过度追求具体方法的使用而对其基本概念和原理理解的不充分与统计方法的误用密切相关。本文将重点讨论两个可能会严重影响统计方法恰当选择而又容易被忽略的概念——样本及其统计量，为医学期刊编辑和医学科研人员提供参考。

1 样本及随机性来源

众所周知，必须带有随机性的数据才能成为统计学研究的对象，所处理的数据是否具有随机性，是区别统计方法与其他数据处理方法的根本所在[3]。根据数据收集的方法或数据随机性的来源，可将数据分为两类，即观察数据和实验（试验）数据[4-5]。

观察数据常来源于抽样调查，当某研究所涉及的研究对象数量很大时，一般不可能也没必要对全部对象进行研究，而只需抽取其部分对象加以考察，如拟调查某地区特定时点某疾病的患病率，假设该地区特定时点共有 500 000 人，其中有拟调查疾病患者 m 人，m 未知，所以患病率 $P=m/500000$ 也未知，要确切知道 P，则须对 500 000 人逐一进行诊断，获得符合诊断标准的患者人数。如前所述，操作可能不易实现，研究者打算采取另一种方法获得 P，从 500 000 人中随机抽取 2 000 人进行调查，500 000 人称为总体，2 000 人称为样本，根据样本对象的诊断结果估计 P。随机性影响就表现在：抽样实施前研究者并不知道具体哪 2 000 人会被抽中，经过有放回地重复抽样，可得到大量同一容量的不同样本，随机性就来自于抽样，对这类样本数据做分析必然会用到统计方法。

实验数据来源于在实验中操纵一个或多个变量后测量的观察指标结果，它的随机性来源于实验的随机误差。假如，某项研究拟探讨维生素 C 对感

基金项目：四川省卫健委课题（项目编号：NO.18PJ314）

冒的影响，研究者将符合纳入标准且不符合排除标准的300名受试对象随机地分配到研究组和对照组，每组150人，研究组每日口服维生素C 1 500毫克，对照组口服安慰剂1 500毫克，考察研究期间两组对象平均患感冒的次数。分组过程中，研究者相当于从300名受试者总体中随机抽取了一部分对象作为A样本，接受A处理，再随机抽取另一部分对象作为B样本，接受B处理，不过，B样本实际上是A样本抽取后剩下的那部分研究对象。研究完成，假设得到两组平均患感冒的次数分别为1.8次和3.2次，是否据此可以认为维生素C对预防感冒有效。然而，两组患感冒次数的差异可能仅仅是由于抽样的随机性所致，让易患感冒的研究对象过多地被分配到样本B中，而免疫力较强的对象更多地抽取到样本A中，因此，这个差异很可能是由机会变异所致，机会变异往往由一些无法或不能完全加以控制且对实验结果产生随机性影响的因素构成，它给结果带来不确定性，称为随机误差，需用统计方法加以分析两组的差异是由随机误差造成的还是存在系统误差（即处理之间存在差别）。

2 样本特征

假设$X_1,X_2,X_3,\cdots,X_n$是从总体X中随机抽取的容量为n的样本，在观察或试验之前，研究者其实并不知道$X_1,X_2,X_3,\cdots,X_n$各取哪个具体的值，但每个分量的可能取值与总体X的可能取值是一致的，它们有相同的值域，且每个分量是一个与总体X分布完全一样的随机变量，因此，样本是一组随机变量，具有随机性特征。如果抽样是有放回地抽取，即抽出一个对象记录后放回总体，将总体摇均匀再进行第二次抽取，以此类推，每次抽样时总体的抽样环境相同，样本各分量之间独立，称这样的样本为独立同分布样本或简单随机样本。在观察或试验之后，样本中每个分量都能获得一个具体的取值，共得到n个已知的数，即$x_1,x_2,x_3,\cdots,x_n$，统计中常以英文大写字母代表随机变量，小写字母代表实数，没有随机性，作为随机变量的取值，显然样本具有数的特征。

样本既可以被看成一组数据也可被看成一组随机变量，这就是样本的二重性。样本二重性虽然简单，但很重要，尤其在实际运用过程中，研究者更容易把样本看成一组具体的数字，但“样本是一组随机变量”这一特征不可忽视，否则容易导致统计学的误用和滥用。如调查研究中，某研究者调查某日去过商场的前100人，这显然不是一个随机样本，称为方便样本。在医学论文中，常可见研究者已描述为所获样本为方便样本，但仍然采用了仅随机样本才能使用的统计方法来处理数据。又如临床研究中的病历资料回顾性分析，这类样本往往并不具有随机性，仍采用分析随机数据的统计方法处理相应资料，显然其结果的可信度存疑。

3 统计量及其抽样分布

由于分布完整地描述了随机变量，所以研究样本的概率分布成为必然，样本的概率分布也称为样本分布，由于样本的独立同分布特征，所以其分布可由总体分布获得。

3.1 统计量

样本获得后，需对其进行加工、整理，从中提取有用信息，统计量就是样本某一方面信息的集中体现，所以选择恰当的样本统计量是处理数据的关键。如用X表示100个学生的身高，从这100人中随机抽取10人测量其身高，用X_1代表准备抽取的第一个人的身高，这个人可以是100人中的任何一个人，其身高取值记为x_1，要求将抽取的第一个人放回总体后，再进行第二次抽取，准备抽取的第二个人的身高用X_2表示，它也可以是100人中的任何一个人，记身高取值为x_2。以此类推，重复抽取10次，得到样本容量为10的一个随机样本，即$X_1,X_2,X_3,\cdots,X_{10}$。现选择样本均值这个统计量对样本$X_1,X_2,X_3,\cdots,X_{10}$进行加工，其计算表达式为：$\overline{X}=\frac{X_1+X_2+\cdots+X_{10}}{10}$，$\overline{X}$是样本均值，是对样本$X_1,X_2,X_3,\cdots,X_{10}$进行加工处理后得到的量。因为样本是一组随机变量，所以$\overline{X}$也是一个随机变量。如果按设定的抽样方案把这10人从100人总体中取出，并测量了这10人的身高，此时就获得了10个具体的数，将其代入样本均数表达式中，获得随机变量$\overline{X}$的一个具体值。可见抽样实施前样本统计量是一个随机变量，实施后，即可得到统计量一个具体值。

通常，统计量是样本的已知函数，它只依赖于样本而不包含任何未知参数，用于总体参数的估计和检验。统计量既然是样本的已知函数，如上所述它也是一个随机变量，也有其概率分布，且这个分布理论上可由样本分布给出，称为抽样分布。然而在实际工作中，统计量的抽样分布的

计算是困难的。如果总体服从正态分布，则样本均值、样本方差等常见统计量的精确分布比较容易算出，抽样分布定理就是它们的分布描述，但对于非正态性分布的更一般总体，其样本统计量的精确分布难以获得，幸运的是可以借助中心极限定理和大数定理获得一些统计量的近似分布，不过，只有当样本量较大时近似才有效，大样本要求在应用过程中是不可忽略的问题[6]。

由于统计学家根据不同的目的构造了许多不同的统计量，下面将以常用的样本均值统计量为例分析相应的抽样分布。

3.2 渐进抽样分布

设$X_1,X_2,X_3,\cdots,X_n$是来自总体均值为μ、方差为σ^2的独立同分布样本，根据中心极限定理，当n充分大时，样本均值$\overline{X}$近似地服从均值为μ，方差为$\frac{\sigma^2}{n}$的正态分布，即$\overline{X}\sim N(\mu,\frac{\sigma^2}{n})$。也就是说，不管总体分布的具体形式如何，只要它的均值为μ，方差为σ^2，中心极限定理保证了从这个总体抽取的简单随机样本，其均值就近似地服从均值为μ、方差为$\frac{\sigma^2}{n}$的正态分布。这为对来自非正态总体的样本数据进行统计学处理提供了理论支持。

3.3 精确分布

设$X_1,X_2,X_3,\cdots,X_n$是抽自均值为μ、方差为σ^2的正态总体的简单随机样本，则样本中各分量均服从正态分布，且是绝对服从，而非近似服从。因为服从正态分布的多个随机变量的和仍为正态分布，一个正态分布乘以一个常数也为正态分布，样本平均数$\overline{X}=\frac{X_1+X_2+\cdots+X_n}{n}$是由服从同一正态分布的多个随机变量$X_i$的和乘以一个常数$\frac{1}{n}$构成，所以$\overline{X}$服从正态分布，并且这个正态分布的均值为$\mu$、方差为$\frac{\sigma^2}{n}$，即$\overline{X}\sim N(\mu,\frac{\sigma^2}{n})$，等价于$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$，也就是说样本均值减去总体均值再除以总体标准差与样本量平方根的商所得量服从标准正态分布，$\frac{\sigma}{\sqrt{n}}$称为标准误差。根据t分布的定义可推知，$\frac{\overline{X}-\mu}{S/\sqrt{n}}$服从自由度为$n-1$的$t$分布，即$\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t_{n-1}$。可见，用样本标准差$s$代替总体标准差$\sigma$后的统计量的分布也是已知的，它解决了实际工作中在总体标准差σ常常不可知情况下使用统计方法处理样本数据的问题。

两个总体均值比较是医学科研中常见的问题。设$X_1,X_2,X_3,\cdots,X_n$是抽自均值为μ_1、方差为σ_1^2的正态总体的简单随机样本，$Y_1,Y_2,Y_3,\cdots,Y_m$是抽自均值为μ_2、方差为σ_2^2的正态总体的简单随机样本。因为来自两个不同总体的样本总是相互独立的，所以两样本均值的差服从均值为$\mu_1-\mu_2$、方差为$\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}$的正态分布，即$\overline{X}-\overline{Y}\sim N(\mu_1-\mu_2,\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m})$，等价于$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}\sim N(0,1)$，也就是说，经标准化处理后得到的变量$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}$服从标准正态分布。如果$\sigma_1^2=\sigma_2^2=\sigma^2$，即两样本方差相等或方差齐，$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}$，根据$t$分布的定义，如果用样本标准差$s$代替总体标准差$\sigma$，则$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n}+\frac{1}{m}}}$服从自由度为$n+m-2$的$t$分布，即$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim t_{n+m-2}$。这解决了在两总体方差未知的情况下，只要方差相等，即可比较两总体均值是否相同的问题。但方差齐的假定须重视，否则公式不适用。

从上面公式导出过程可知，无论是渐进抽样分布公式还是精确抽样分布公式，都是建立在简单随机样本基础上，也称为独立同分布样本，指在相同条件下对总体X进行n次重复且独立的随机抽样，所获得的是由n个独立且与总体X具有相同分布的分量组成的随机样本。严格地讲，为保证抽样条件相同，须采取有放回地重复抽样方法随机抽取。由此可见，公式的使用首先必须强调样本分量间的独

立性，独立性的判断主要根据专业知识和样本获取过程以及研究者采用的抽样方法。其次，总体的正态性考察，当总体为非正态总体，所获得的统计结果多为近似结果，而且要在大样本量的情况下这种近似才有效，如果是正态总体，在其他条件都严格满足的情况下，所得结果精度更高。再次，需要对方差齐性考察，公式推导过程显示，只有当两总体方差相等时，才可得到两样本均值差的抽样分布，否则，相应抽样分布定理不可能存在。

然而，在实际应用中，上述诸多条件往往并不能严格地满足统计理论的要求，为处理问题方便，如常把一些非正态总体近似地看成正态总体，用正态分布来逼近总体的分布，这种逼近所带来的误差从应用的角度看可忽略不计。又如在试验分组时，往往获得的各分组样本并不满足独立同分布要求，因为抽样过程无法真正做到有放回地抽样，而采取的是不放回抽样方法，改变了抽样条件，正如前例，某研究对象被抽到维生素 C 组，则他就没有机会被抽到安慰剂组，这种分组会稍许影响两组平均数，不过在实际应用过程中，仍然采用了独立同分布样本的抽样分布相关定理来处理数据，原因在于非独立性可致使标准误差变小，不放回抽样使其增大，相应统计结果误差是可以被接受的[7]。尽管如此，研究者也应当明白，当理论要求与现实条件相差甚远时，这种逼近不是一个好的选择。

4 抽样分布在参数估计中的应用

医学统计学的任务主要包括统计描述和统计推断，前者主要针对样本，采用图、表和统计量等工具对样本进行加工处理，后者是在样本加工的基础上对相应总体的特征进行推断，其工作范围已超越样本指向了总体，主要包括参数估计和假设检验，参数估计又可分为点估计和区间估计，下面以总体均值的区间估计为例，介绍抽样分布的应用。

以前例 100 人总体为例，估计他们的平均身高 μ，拟从总体中随机抽取 10 人，分别用 $X_1, X_2, X_3, \cdots, X_{10}$ 代表其身高，实施抽取后获得 10 人样本，并分别测量其身高，计算平均值 $\overline{X}$，假设总体服从正态分布，样本为独立同分布样本，由抽样分布定理可知，样本均值 $\overline{X}$ 服从均值为 μ、方差为 $\frac{\sigma^2}{n}$ 的正态分布，即 $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$，根据正态分布性质，有约 68% 的样本平均数落在以总体均值 μ 为中心的 1 个标准误差 $\frac{\sigma}{\sqrt{n}}$ 范围内，有约 95% 的样本平均数落在 2 个标准误差范围内，有约 99.7% 的样本平均数落在 3 个标准误差范围内，如图 1。

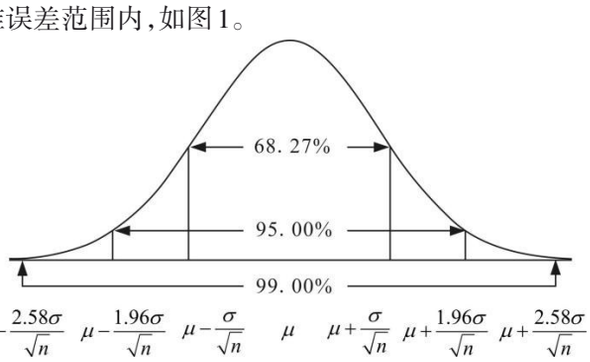

图 1 正态抽样分布

由图可见，如果以落在 2 个标准误差范围内的样本均值加减 2 个标准误差构造区间，即 $(\overline{X}-2\frac{\sigma}{\sqrt{n}}, \overline{X}+2\frac{\sigma}{\sqrt{n}})$，则这个区间就完全可能覆盖总体均值 μ，然而，在所有的样本均值中，仅有约 95% 的数落在以总体均值 μ 为中心的 2 个标准误差范围内，还有约 5% 的样本均值落在这个范围之外，若以范围外的样本平均数加减 2 个标准误差所得区间是不可能覆盖总体均值 μ，如果用这个区间作为总体平均数的估计区间，就会犯错误，不过犯错的可能性仅为 5%，而所获区间覆盖总体均值 μ 的把握度则为 95%，统计学上称为 95% 的置信区间。这个过程更形象地讲，它就像从一只装满区间的盒子中随机抽取一个区间，因盒子中所装的区间有 95% 是覆盖了总体均数 μ，仅有 5% 的没有覆盖，抽取一次所得到能够包括总体均值的区间的概率为 95%，犯错的概率为 5%。值得注意的是，实际工作中总体标准差 σ 往往未知，可用样本标准差 S 代替，根据抽样分布定理，相应统计量则服从 t 分布，可用 t 分布求置信区间。

5 小 结

本文详细介绍了样本及其产生过程，样本的随机性及其来源，样本统计量特征及其抽样分布定理等内容，回答了在统计方法应用中应当满足哪些条件以及为什么需要满足这些条件，尽管实际应用不一定完全能满足要求，但在做近似处理时应当把握“度”的问题[8]，最终不让统计结果偏离较远。

参考文献

[1] 韩宏志，刘仲祥，官鑫，等. 医学期刊中常见的几种统计学分析方法错用误用案例辨析[J]. 编辑学报，2019，31(1)：37-40.

[2] 郭瑞，宋国营，张媛，等. 医学期刊编辑应重点审核的统计学问题分析及建议[J]. 编辑学报，2019，31(6)：623-625.

[3] 李金昌. 论统计数据的随机性[J]. 中国统计，2018(7)：17-19.

[4] 李金昌. 统计学三要素：问题、数据和方法[J]. 中国统计，2018(3)：40-42.

[5] 陈希孺，倪国熙. 数理统计学教程[M]. 合肥：中国科学技术大学出版社，2009：2-4.

[6] 王松桂，张忠占，程维虎，等. 概率论与数理统计[M]. 北京：科学出版社，2011：120-127.

[7] Freedman D，Pisani R，Purves R，等. 统计学[M]. 魏宗舒，施锡铨，林举干，等译. 北京：中国统计出版社，1997：555-557.

[8] 吴俊玲，周晴，周英智. 医学论文统计学的“度”问题分析及建议[J]. 编辑学报，2017，29(4)：348-350.

（收稿日期：2020-12-04）

（本文编辑：戴浩然）

（上接第 509 页）

[9] 李卫红. 高中生心理健康状况调查及分析[J]. 内蒙古民族大学学报，2010，16(2)：133-134.

[10] 唐蕾，应斌. 新冠肺炎疫情时期中学生心理健康状况及影响因素调查分析[J]. 中小学心理健康教育，2020(10)：57-61.

[11] 余清香，曾艺敏，卢文洁. 新冠肺炎疫情期中学生心理健康状况调查分析[J]. 江苏教育，2020(32)：44-47.

[12] Hou TY，Mao XF，Dong W，et al. Prevalence of and factors associated with mental health problems and suicidality among senior high school students in rural China during the COVID-19 outbreak[J]. Asian J Psychiatr，2020，54：102305.

[13] 王慧，黄琦岚，尹红新，等. 新冠肺炎期间研究生心理健康状况[J/OL]. 中国健康心理学杂志，http://kns.cnki.net/kcms/detail/11.5257.R.20200604.1105.012.html，2020-12-04.

[14] 邢晓辉，常军武，王华峰. 医学研究生心理健康状况和人格特征分析[J]. 学位与研究生教育，2005(3)：21-24.

[15] 孙洪礼. 大学生心理健康与人格特质的相关[J]. 中国健康心理学杂志，2017，25(10)：1567-1571.

[16] 王欢，彭亚玲，严嗣刚，等. 高中生对新型冠状病毒肺炎认知、态度及其心理健康状况[J]. 中国健康心理学杂志，2020，28(12)：1829-1833.

[17] 张艳，庄凌云，杨伟. 新冠疫情期间中学生创伤后应激障碍症状调查——以成都市树德中学为例[J]. 教育科学论坛，2020(17)：45-48.

[18] 梁剑玲，张均华，陈晓新，等. 疫情背景下中学毕业年级学生心理健康状况分析——基于与非毕业年级学生的比较[J]. 中小学心理健康教育，2020(15)：15-18.

[19] Zhou SJ，Zhang LG，Wang LL，et al. Prevalence and sociodemographic correlates of psychological health problems in Chinese adolescents during the outbreak of COVID-19[J]. Eur Child Adolesc Psychiatry，2020，29(6)：749-758.

（收稿日期：2020-09-29）

（本文编辑：吴俊林）