【继续教育园地】

统 计 学 系 列 讲 座

第 1 讲 常用统计学基本概念及统计描述(1)

安胜利

(南方医科大学 生物统计学系,广东 广州 510515)

1 基本概念

1.1 总体与样本 观察单位(observed unit)是统计研究中的基本单位,一般指一个人,一只动物,也可以是指特定的一个家庭,一个自然村,还可以是一个器官。根据研究目的所确定的同质研究对象的全体, 其某种观察值的集合构成总体(population),如调查某地 2004 年 7 岁正常女童的身高,则该地 2004 年所有 7 岁正常女童的身高就构成一个总体。总体是一个相对的概念,是根据研究目的确定的。如果想调查全国 2004 年 7 岁正常女童的身高,则该地 2004 年 7 岁正常女童的身高只是总体中的一份样本。总体有有限总体和无限总体之分,但一般都比较大,所以总体指标(即参数 parameter)——例如上例中的总体平均身高 μ一般是未知的,但却正是我们想知道的。为节省人力、物力和财力,实际工作一般都需要从总体中抽取样本。从同质总体中随机抽取有代表性的观察单位,其实测值构成样本(sample)。如上例,可从某地 2004 年 7 岁正常女童中随机抽取 110 名女童,所得身高测量值就是样本,样本平均身高就是统计量(statistic)。统计量一般不会恰好等于参数,即抽样误差(sampling error)是不可避免的。我们的最终目的就是设法用统计量来推断参数,这也就是统计推断(分析)的过程。

1.2 同质与变异 性质相同的事物具有同质性(homogeneity),但同质的事物就同一观察指标来看,各观察个体之间也有变异(variation)。变异或曰个体差异是生物医学领域普遍存在的现象。如相同年龄、性别儿童的身高有高有低;同种属、性别和年龄的小白鼠喂同样的饲料,所增体质量各不相同。统计分析就是在同质分组的基础上,通过对各组内一定样本含量(sample size)的观察单位个体变异的研究,透过相对偶然的现象反映同质事物的本质特征和规律。

1.3 资料类型 研究结果可分为 3 种不同的资料类型,即计量资料(measurement data),计数资料(enumeration data)和等级资料(ranked data)。

计量资料又称定量资料、数值变量资料,指用某种方法或仪器, 观测每个观察单位某项指标的大小所获得的资料,一般有度量衡单位。如每个人的身高、体质量、红细胞计数、各医院某年死亡人数等。

计数资料又称定性资料、无序分类变量资料或名义变量资料,是将观察单位按某种属性或类别分组计数,然后汇总各组观察单位数后所得资料。如某病患者治愈和未治愈人数,某人群 4 种血型人数等。

等级资料又称半定量资料或有序分类变量资料,是将观察单位按某种属性的不同程度分成等级后分别计数,然后汇总各等级内观察单位数后所得资料,具有半定量性质,如某病患者治疗结果分为治愈、显效、好转、无效;某血清反应结果分为-、+、++、+++等。

上述资料类型在有关专业理论指导下,可互相转化。如脉搏数(次/min)为计量资料,若定义 60~100 次/min 为正常,其他为异常,则可按正常与异常分别清点人数,汇总后转化为计数资料;又如可将具体年龄(计量资料)转化为儿童、青少年、中年、老年(等级资料)。因此建议进行研究时,尽可能获得定量指标。少数情况下,为满足某些统计分析的要求,也可设法将计数或等级资料近似转化为计量资料。

2 统计描述

统计描述是统计分析的必经之路,通过计算相应的统计指标,必要时结合统计图表就可较全面地刻画出研究结果的特征。

2.1 计量资料的统计描述

2.1.1 平均水平指标 通常使用平均指标来描述一组计量资料的集中趋势,它反映了一组观察值的平均水平。常用的有算术均数、几何均数和中位数。由于目前计算机的普及,我们一般可用软件精确计算结果,所以下述公式中没有列出样本含量较大时相应的近似计算公式,若需手工计算,请参阅有关文献。

算术均数(mean, $\bar{X}$):n 个性质相同的定量数据之和除以 n 所得结果,见式(1)。均数适用于描述对称分布,尤其是正态分布计量资料的平均水平。

$$\bar{X}=\frac{X_1+X_2+\cdots+X_n}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad (1)$$

几何均数(geometric mean, $\overline{X_G}$):对 n 个性质相同的定量数据分别取对数变换后,按算术均数计算,然后再求其反对数所得结果,见式(2)。当数据呈倍数关系或服从对数正态分布(一种正偏态分布)时,适于用几何均数描述其平均水平。

$$\overline{X_G}=\lg^{-1}\frac{\lg X_1+\lg X_2+\cdots+\lg X_n}{n}=\lg^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) \qquad (2)$$

中位数 (median, M):n 个性质相同的定量数据按从小到大排序,居中的数据就是中位数,见式(3)和(4)。中位数其实就是第 50 百分位数,它恰好将一组数据等分为 2 部分,多用于描述偏态分布资料的平均水平,尤其是当资料中有无法准确测量的数据时。

$$M=X_{(n+1)/2} \qquad (n\text{为奇数}) \qquad (3)$$

$$M=(X_{n/2}+X_{n/2+1})/2 \qquad (n\text{为偶数}) \qquad (4)$$

2.1.2 变异(离散)水平指标 对于下面两组数据,尽管其均数都是 10,但显然 A 组数据较集中,而 B 组数据较分散,即其变异度不同。因此,要全面反映一组资料的分布特征,除了用平均指标,还应结合变异水平指标。常用变异指标有极差、四分位数间距、方差、标准差和变异系数。

A 组: 8 9 10 11 12

B 组: 3 7 10 13 17

极差(range, R):一组性质相同的定量数据中最大值与最小值之差。极差不稳定,较少使用,多用于说明传染病、食物

中毒的最长、最短潜伏期等。

四分位数间距(quartile range, Q): 一组性质相同的定量数据中, 第 75 百分位数与第 25 百分位数之差, 可看作中间一半观测值的极差, 多用于描述偏态分布资料的变异水平。

方差(variance)**和标准差**(standard deviation, S): 反映一组数据的平均变异水平, 前者是后者的平方。标准差最常应用, 常结合均数描述正态分布的特征。变异系数 (coefficient of variation, CV): 标准差与算术均数的比值。当比较两组或多组计量资料的变异度大小时, 如果各组资料的度量衡单位不同, 或者各组算术均数相差悬殊, 则必须用 CV 进行比较。标准差、变异系数计算分别见式(5)、(6)。

$$S=\sqrt{\frac{\sum (X_i-\bar{X})^2}{n}} \qquad (5) \qquad CV=\frac{S}{\bar{X}}\times 100\% \qquad (6)$$

接下来看几个案例, 其他统计描述内容在下次讲座中接着讨论。

案例 1: 某论文中, 给出如下资料, 见表 1。论文中写道 "高倍镜下每例病人肿瘤区域计数 500 个细胞, 计算阳性细胞百分率。统计分析用卡方检验"。

表 1 血管瘤、淋巴瘤中 ER、PR 检测结果($\bar{x}\pm s$)

类别	例数	ER 百分率(%)	PR 百分率(%)
毛细血管瘤	45	74.18±1.77	77.92±0.54
混合型血管瘤	44	64.55±2.34	68.12±5.38
淋巴管瘤	23	26.93±5.62	30.00±8.87

分析: 作者用计量资料的表达法, 这是正确的, 但却用计数资料的分析方法卡方检验。作者以为凡是用 "率"、"百分比"表达的资料都是计数资料, 事实上本研究所收集的资料是计量资料, 因为每一个观察单位身上都测得一个 "百分率", 就如同都测得一个身高值一样。当碰到 "率"或 "百分比" 这样的资料时, 关键要看它是由 "观察单位个数"计算得到的, 还是由每一个观察单位自身测量得到的。前者属于计数资料, 后者属于计量资料。

案例 2: 某作者在论文中, 给出如下资料, 见表 2。

表 2 大鼠头部受伤后 1 d 迷宫实验出错结果

实验分组	n	出错次数($\bar{x}\pm s$)
不给予撞击	6	1.89±2.28
50 g 砝码撞击	10	3.22±3.79
100 g 砝码撞击	10	2.88±4.30

分析: 这 3 组数据标准差都大于相应的均数, 对于实际的计量资料, 基本可判断各组资料呈明显偏态分布, 而 "$\bar{x}\pm s$" 一般是用于描述服从正态分布资料的平均水平和离散水平的, 并不是适用于任何计量资料。本例宜选用中位数和四分位数间距进行描述, 其形式为 "M(Q)"。例如: 若 "不给予撞击" 组的中位数为 2.90, 四分位数为 3.68 的话, 则写作 2.90(3.68)。

案例 4: 某地 1995 年不同年龄组男童身高资料见表 3, 并据此认为 6 岁以下男童身高的均数和变异度都随年龄增长而增加。

表 3 某地 1995 年不同年龄组男童身高资料

年龄	人数	均数(cm)	标准差(cm)
1~2 月	100	56.3	2.1
5~6 月	120	66.5	2.2
3~3.5 岁	300	96.1	3.1
5~5.5 岁	400	107.8	3.3

分析: 身高的确有增加趋势, 但由于不同组的均数相差悬殊, 应该用变异系数来比较各组间的变异度。本例各组变异系数分别为 3.73%、3.31%、3.22%和 3.06%, 可见变异度随年龄的增加有减小的趋势。

[本文编辑: 方玉桂]

练 习 题

1.比较身高和体质量的变异度, 应采用的指标是

A 标准差 B 方差 C 变异系数
D 四分位数间距 E 全距

2.数列 8, -3, 5, 0, 1, 4, -1 的中位数是

A 2 B 1 C 2.5 D 0.5 E 3

3.计量资料、计数资料、等级资料的关系是

A 计量资料同时具有计数和等级资料的一些性质
B 计数资料同时具有计量和等级资料的一些性质
C 等级资料同时具有计数和计量资料的一些性质
D 三种资料不可以相互转化
E 等级资料较计量和计数资料精确

4.描述一组偏态分布资料的变异度, 以 _____ 指标较好。

A 全距(R) B 标准差(S) C 变异系数(CV)
D 四分位数间距($Q_U \sim Q_L$) E 方差(S^2)

5.统计学所说的总体是指

A 根据研究目的所确定的同质研究对象的全体
B 根据人群划分的同质研究对象的全体
C 根据地区划分的同质研究对象的全体
D 根据时间划分的同质研究对象的全体
E 根据不同单位划分的同质研究对象的全体

温 馨 提 醒

亲爱的读者朋友:

新年伊始, 首先祝大家在新的一年里身体健康, 心情愉快, 生活幸福, 事业有成!

本刊 2005 年第 10 期进行的有关统计学知识需求调查结果显示: 广大护理工作者统计学相关知识较为欠缺, 但工作科研中又极其需要, 就此本刊 2006 年 《继续教育园地》隆重推出统计学系列讲座, 希望大家认真学习, 积极参与, 工作科研中遇到统计学问题可来信来函反映, 本栏目将就大家共性问题请讲座老师进行答疑。

2006 年第 1- 第 10 期有 《继续教育园地》, 每期附有练习题, 大家每期学习后做好练习并保存好。同时提醒大家留意每期 《继续教育园地》相关信息, 严格按要求参与学习和答题, 以保证学习效果和继续教育项目的顺利进行。