关于“二项”的分布类型的研究

李远景,李方东*　(安徽农业大学理学院,安徽合肥 230036)

摘要　参考有关专著,结合多年的教学研究,认为“二项”的基本分布类型有二项总体分布、二项分布(二项次数分布)和二项成数分布3种。前者为总体分布,后二者都是抽样分布。

关键词　二项分布;总体分布;抽样分布
中图分类号　S11+4　　**文献标识码**　A　　**文章编号**　0517 - 6611(2014)36 - 12793 - 02

Study on the Binomial Distribution Types

LI Yuan-jing, LI Fang-dong* (College of Science, Anhui Agricultural University, Hefei, Anhui 230036)

Abstract　According to relevant monographs and several years' teaching research, it was proposed that binary has three kinds of distribution types: the binary population distribution, the binomial distribution (binary frequency distribution) and the binary into number distribution, the former is population distribution, the latter two are sampling distribution .

Key words　Binomial distribution; Binary population distribution; Sampling distribution

“二项”的分布类型有几个?它们的确切定义或概念是什么?如何正确理解且区别不同类型的分布?对于这些问题,目前的教科书特别是生物统计教科书中阐述得不太明确,导致在教学过程中学生难以深入理解。针对上述问题,笔者在参考有关专著,结合多年的教学研究,对“二项”的基本分布类型进行了定义和区分。结合实际问题的分析,帮助学生理解这些概念,从而取得良好的教学效果。

1　有关基本概念的定义

1.1　二项试验　设在每次试验中只可能出现两种互不相容的结果(此事件A和彼事件$\overline{A}$)且每种结果在每次试验中出现的可能性都相同,称为二项试验(也称为贝努里试验)[1-3]。例如,在产品抽样检查中,每抽出一件产品即为一次试验,我们所注意的是该产品是正品还是次品;有些试验结果不只2种,如在育种的杂交后代中会有多种类型,但是我们所关心的只是需要部分,其余的为不需要,即把试验结果根据需要与否(或是否合格)划分成非此即彼的两个互不相容的事件,同样为二项试验。

1.2　二项总体　由非此即彼事件构成的总体称为二项总体[4]。二项试验研究的总体都是二项总体。为了方便研究,一般给此事件变量1,给彼事件变量0,因此二项总体又可称为0、1总体[4]。例如,观察施用某种农药后小麦蚜虫的死亡情况,记“死”为0,记“活”为1,该试验研究的总体即为0、1总体。

2　三种分布

首先,应明确什么是概率分布。所谓概率分布,是指随机变量的取值(某一定值或区间)与其对应概率的关系。这种对应概率的关系既可以用分布列(离散型随机变量的概率分布)表示,又可以用分布函数或分布图(离散型随机变量的概率分布和连续型随机变量的概率分布都行)表达。

作者简介　李远景(1957 -),男,安徽合肥人,副教授,从事生物统计的教学、生物方面的试验设计和生物统计分析方面的研究。*通讯作者,中级实验师,硕士,从事计算机在生物统计教学方面的研究。

收稿日期　2014-11-27

2.1　二项总体分布(0、1总体分布或贝努里分布)[2,5]　由二项试验的定义可知,无论此事件或彼事件,它们各自在每次试验中出现的概率相同,这样若给此事件以变量1,具有概率为p,给彼事件以变量0,其概率则为$q(q=1-p)$,也就是说,在二项总体中,随机变量X仅取2个值1和0,其对应的概率分别为p和q。因此,二项总体分布就可以定义为:二项总体随机变量X取值1和0时所对应的概率分布。其分布列为:

$$\begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}$$

其分布图见图1。

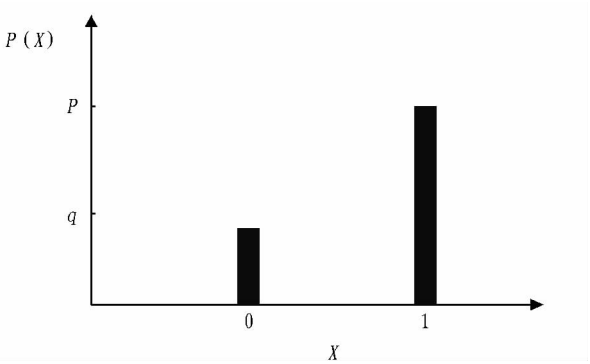

图1　二项总体分布示意图

由概率论[2],离散型随机变量X的总体平均数为各变量值x_i与其对应的概率p_i的乘积和,其总体方差为各变量的离均差平方与其对应的概率p_i的乘积和,即$\mu=\sum_i^N p_iX_i$,$\sigma^2=\sum_i^N p_i(X_i-\mu)^2$。因此,对于二项总体的均值和方差,有

$$\mu=\sum_i^N P_iX_i=p(1)+q(0)=p$$

$$\sigma^2=\sum_i^N P_i(X_i-\mu)^2=p(1-p)^2+q(0-p)^2=pq(q+p)=pq$$

为了说明上述公式的实际意义,现假定观察一小麦种子的发芽情况,以10粒($N=10$)为一个总体,$x=1$时指发芽粒,$x=0$时指未发芽粒,总体内10个个体观测值的结果分别

为 1、0、1、1、0、1、1、0、1、1。其总体平均值和方差按基本公式计算为:

$$\mu = \frac{\Sigma x}{N} = \frac{1+0+1+1+0+1+1+0+1+1}{10} = \frac{7}{10} = 0.7 = p$$

$$\sigma^2 = \frac{\Sigma (x-\mu)^2}{N} = \frac{7\times(1-0.7)^2 + 3\times(0-0.7)^2}{10} = \frac{2.1}{10}$$

$$= 0.21 = 0.7\times 0.3 = pq$$

这里的 0.7 实际上就是发芽率或 $x=1$ 的概率 p,0.3 就是未发芽率或 $x=0$ 的概率 q。

2.2 二项分布(二项次数分布) 在二项总体中,如果以 n 为样本容量进行随机抽样,某一事件(如此事件)出现的次数 x(取值 0、1…n)所对应的概率分布称为二项分布[5]或二项次数分布,记为 $X \sim B(n,p)$。也就是说,在样本容量 n 内,此事件出现 0、1…n 次所对应的概率分布为二项分布。在 n 次观察中,计算此事件出现 x 次的概率可以利用二项展开式 $(p+q)^n$ 或二项概率函数 $f(x)$ 计算。

$$f(x) = C_n^x P^x q^{n-x}, \text{其中 } C_n^x = \frac{n!}{x!\ (n-x)!} (x = 0,1\cdots n)$$

例如,以 $n=5$ 在一个 $p=0.5$ 的二项总体中进行随机抽样,此事件出现 2 次的概率为 $f(x=2) = C_5^2 0.5^2 0.5^3 = 0.3125$。其含义为:若以 $n=5$ 在一个 $p=0.5$ 的二项总体中进行随机抽样,在抽出的 100 个样本中,此事件出现 2 次的样本约为 31.25 个或其比例约为 31.25%,其二项分布图形如图 2 所示。

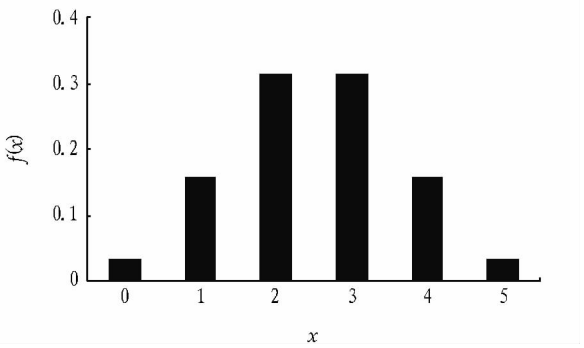

图 2 二项分布示意图($P=0.5, n=5$)

二项分布随机变量 X 的总体平均值(数学期望)和方差[2]为:

$$\mu = EX = \sum_{x=0}^{n} XP(x) = \sum_{x=1}^{n} XC_n^x p^x q^{n-x} = np\sum_{x=1}^{n} C_{n-1}^{x-1} p^{x-1} q^{n-x} = np(p+q)^{n-1} = np$$

$$\sigma^2 = DX = EX^2 - (EX)^2 = \sum_{x=0}^{n} X^2 C_n^x p^x q^{n-x} - (EX)^2 = npq + n^2p^2 - n^2p^2 = npq$$

例如,以 $n=8$ 在 $p=0.5$ 的二项总体中进行随机抽样,此事件出现次数的平均值和方差分别为:$\mu = np = 8\times 0.5 = 4$ 次,$\sigma^2 = npq = 8\times 0.5\times 0.5 = 2$ 次^2。

2.3 二项成数分布 在二项总体中如果以 n 为样本容量进行随机抽样,某一事件(如此事件)出现的成数 $\hat{p}$ 所对应的概率分布称为二项成数分布。显然,$\hat{p}$ 成数是一个随机变量,它的取值是在 n 内此事件出现的次数 x 与 n 之比。由于次数 x 的取值为 0、1、2…n,所以成数 $\hat{p}$ 的取值应该为 0/n、1/n、2/n…n/n。由此可以推知,成数取某一值的概率应该等于对应次数的概率,因此若要计算某一成数的概率,则只要把成数化为对应次数求其概率即可。很明显,在二项分布图中,只要把横坐标次数 x 除以 n 即可得到二项成数分布图。把次数分布的平均数和方差分别除以 n 和 n^2,就可以得到成数的平均数和方差。

$$\mu_{\hat{p}} = \frac{np}{n} = p$$

$$\sigma_{\hat{p}}^2 = \frac{npq}{n^2} = \frac{pq}{n}$$

例如,以 $n=8$ 在一个 $P=0.5$ 的二项总体进行随机抽样,此事件出现的成数的平均值为 $\mu_{\hat{p}} = p = 0.5$,成数的方差为 $\mu_{\hat{p}}^2 = \frac{pq}{n} = \frac{0.5\times 0.5}{8} = 0.031\ 25$。

从上面的分析可以看出,二项的分布类型有 3 种。为方便区别 3 种分布,现将 3 种分布的有关特征进行列表(表 1)。

表 1 3 种分布的有关特征列表

分布类型	分布性质	变量取值个数	均值(μ)	标准差(σ)
二项总体分布	总体分布	0 和 1 两个	p	$\sqrt{pq}$
二项(次数)分布	抽样分布	($n+1$)个	np	$\sqrt{npq}$
二项成数分布	抽样分布	($n+1$)个	p	$\sqrt{pq/n}$

3 讨论

(1)“二项”的分布共有 3 种基本分布:二项总体分布是变量只取 2 个数值(1 和 0)的总体分布,而二项分布和二项成数分布是变量可以取($n+1$)数值的抽样分布。明确这一点,就不难利用次数或成数进行显著性的假设测验。

(2)二项分布是一种次数分布,即在二项总体中以 n 为样本容量进行随机抽样,某事件出现的次数是一随机变量。所以,二项分布也可以称为二项次数分布,有的解释为样本总和数分布[6],其实质虽然一样,但理解为次数分布更易接受。

(3)将二项分布的次数变量除以样本容量(n),就可以把二项分布转变成二项成数分布。

参考文献

[1] 扬纪珂,齐翔林,陈霖. 生物数学概论[M]. 北京:科学出版社,1982:399.
[2] 复旦大学.《概率论》第一册《概率论基础》[M]. 北京:高等教育出版社,1984:76-174.
[3] 浙江大学数学系高等数学教研组. 工程数学(概率论与数理统计)[M]. 北京:高等教育出版社,1983:42-116.
[4] 莫惠栋. 农业试验统计[M]. 上海:上海科学技术出版社,1984:34-110.
[5] SUEEOR G W,COHRAN W G. Statistical methods[M]. 7th ed. Iow,Ames:The Iowa State University Press,1980:30-1189.
[6] 盖钧镒. 试验统计方法[M]. 北京:中国农业出版社,2006:72.