中国农业大学学报 2001, 6(3): 44~ 49
Journal of China Agricultural University

① 作物品种区域试验中品种均值的 Bayes 估计①

张群远② 孔繁玲
(中国农业大学作物学院)

摘 要 本研究根据Bayes 统计原理, 提出作物品种区域试验中品种均值的Bayes 估计方法; 并利用一套包含 4 年、7 个试点和 10 个棉花品种的多年多点试验对Bayes 估值和算术平均值的预测精度进行比较。结果表明,Bayes 估值和算术平均值的平均预测差分别为 6. 88% 和 12. 77%; 而且Bayes 估值与验证值之间在数值和品种排序上都有着更高的相关, 相关系数分别达 0. 963 和 0. 976, 高于算术平均值的 0. 876 和 0. 830; Bayes 估值显示出比算术平均值更高的预测精度。

关键词 区域试验; Bayes 估计; 算术平均值; 预测精度
中图分类号 S114

Bayesian Estimation of Variety Means in Regional Crop Trials

Zhang Qunyuan Kong Fanling
(College of Crop Science, CAU)

Abstract Based on the principle of Bayesian Statistics, a Bayesian method for estimating variety means in regional crop trials was proposed. A trial involving 4 years and 7 sites and 10 varieties of cotton was carried out and analyzed to compare the predictive accuracy of the Bayesian estimates and arithmetic means. It was found that the averaged predictive differences of Bayesian estimates and arithmetic means were 6. 88% and 12. 77% respectively; the coefficients of correlation between Bayesian estimates and validation data were 0. 963 for values and 0. 976 for variety ranks, higher than 0. 876 and 0. 830 of between arithmetic means and validation data. Bayesian estimates showed higher predictive accuracy than arithmetic means.

Key words regional trial; Bayesian estimation; arithmetic mean; predictive accuracy

作物品种区域试验(简称区试)是在多环境下对新品种进行比较的规模较大的农业试验。准确估计各品种在区试中的性状总均值(简称品种均值)是评价品种平均生产性能(譬如丰产性)的重要依据, 因而也是区试中最基本和最重要的统计内容。对于品种均值, 区试中习惯于直接用算术平均值估计。尽管算术平均值满足线性无偏(linear unbiased)、最小二乘(least squares)和极大似然(maximum likelihood)等统计特性, 但对于区试品种均值的估计来说, 仍然存在一定局限。从某种意义上讲, 算术平均值实质上是对品种试验表现的事后描述, 并非对品种未来表现的预测, 而后者是我们真正感兴趣的, 具有更重要的实践意义。目前我国区试中,

收稿日期: 2000-10-19
①国家自然科学基金资助项目(30070433)
②张群远, 北京圆明园西路 2 号中国农业大学(西校区), 100094

同一批品种一般只进行 1 年的试验, 但由于年份间存在气候条件变化, 仅根据 1 年的算术平均值判断品种(在以后多年中)的应用价值, 可靠性并不高。要提高品种评价的可靠性, 一方面可以增加试验的年份数, 另一方面可以采用更有效的统计分析方法。相对而言, 前者需消耗较多的资源(占地和人工等); 而后者则几乎不增加任何试验费用。基于后一种想法, 本研究从农业试验中应用较少的贝叶斯(Bayes)统计学原理出发, 研究区试中品种均值的Bayes 估计方法; 并利用相应的多年多点试验对Bayes 估值的预测效果进行分析, 以期为区试中品种均值的估计寻求一种在预测精度上优于传统的算术平均值的方法。

1　材料与方法

1.1　品种均值的 Bayes 估计的原理和方法

Bayes 方法的基本原理是结合已有经验和认识(称为先验信息)和当前试验信息二者来进行统计推断[1,2]。品种区域试验中, 同一套品种要在多个地点(甚至多个年份)重复进行试验, 尽管时间上各地点的试验是同时进行的, 但逻辑上我们可以认为是一个地点接一个地点依次进行的。这样, 可以应用Bayes 统计的原理, 把第 1 个地点的统计结果作为已知的先验信息, 而第 2 个地点的试验资料作为试验信息, 并根据一定的准则(即Bayes 定理)[1,2], 对先验信息和试验信息进行综合, 得出新的统计结果; 然后, 再把这一结果作为先验信息, 把第 3 个地点的试验资料作为试验信息, 再得出新的结果……如此循环, 直至最后一个地点, 便得到最终的Bayes 统计结果。

针对区试中品种均值估计这一问题, 现假设我们通过第 1 个地点的 n_1 个试验观测值已认识到某品种均值 μ 服从先验的正态分布N (μ_{t1}, σ^2_{t1})。按理, μ 为品种参数, 是一个定值, 不应该有方差; 但是, 这里我们对 μ 的已有认识是不完全确定的, 这种不确定性便以方差 σ^2_{t1}来度量。这正是Bayes 方法有别于一般统计学的一个重要方面; 也正因为如此, 才有必要进一步进行试验观测来更好地认识 μ。在进一步的试验中(即第 2 个地点的试验), 获得该品种的 n_2 个重复观测值, 假定它们是来自正态分布总体N (μ, σ^2_2)的随机样本, 其样本均值为 Y_2。结合先验分布信息(即第 1 个地点的试验结果)和试验信息(即第 2 个地点的试验结果), 根据Bayes 公式, 可推导出(见附录)品种均值 μ 服从分布N (μ_{t2}, σ^2_{t2}), 其中:

$$\mu_{t2} = \sigma^2_{t2}\left(\frac{1}{\sigma^2_{t1}}\mu_{t1} + \frac{n_2}{\sigma^2_2}Y_2\right) \tag{1}$$

$$\sigma^2_{t2} = \frac{1}{1/\sigma^2_{t1} + n_2/\sigma^2_2} \tag{2}$$

此分布称为 μ 的后验分布, 即综合先验信息和试验信息所获得的关于 μ 的新认识。随后, 把 $\mu \sim$ N (μ_{t2}, σ^2_{t2}) 作为新的先验分布, 进一步试验(即第 3 个地点的试验), 假定有 n_3 个重复观测值来自总体N (μ, σ^2_3), 算术平均值为 Y_3, 再综合二者信息可再次得到 μ 的后验分布N (μ_{t3}, σ^2_{t3}), 其中:

$$\mu_{t3} = \sigma^2_{t3}\left(\frac{1}{\sigma^2_{t2}}\mu_{t2} + \frac{n_3}{\sigma^2_3}Y_3\right) = \sigma^2_{t3}\left(\frac{1}{\sigma^2_{t1}}\mu_{t1} + \frac{n_2}{\sigma^2_2}Y_2 + \frac{n_3}{\sigma^2_3}Y_3\right) \tag{3}$$

$$\sigma^2_{t3} = \frac{1}{1/\sigma^2_{t2} + n_3/\sigma^2_3} = \frac{1}{1/\sigma^2_{t1} + n_2/\sigma^2_2 + n_3/\sigma^2_3} \tag{4}$$

按此做法, 不断进行试验和估计, 形成一个迭代的过程。到最后第 s 个地点的试验, 可得到

μ 最终的后验分布为N (μ_{ts}, σ^2_{ts}), 其中:

$$\mu_{ts}= \sigma^2_{ts}\left(\frac{1}{\sigma^2_{t1}}\mu_{t1}+\sum_{i=2}^{s}\frac{n_i}{\sigma^2_i}Y_i\right) \tag{5}$$

$$\sigma^2_{ts}= \frac{1}{1/\sigma^2_{t1}+\sum_{i=2}^{s}n_i/\sigma^2_i} \tag{6}$$

式中: n_i 和 Y_i 分别为第 i 个地点的试验重复数和品种算术平均值。在 2 次损失的意义下, 依据后验风险最小的原则, 后验均值 μ_{ts}即为品种均值 μ 的Bayes 估计[2]。若以地点 i 的试验误差方差 s^2_i 估计 σ^2_i, 以第 1 个地点上试验的品种均值 Y_1 和该均值的误差方差(等于误差方差 s^2_1 除以重复数 n_1, 即 s^2_1/n_1)分别估计 μ_{t1}和 σ^2_{t1}, 则据(5)和(6)式可得到 μ_{ts}和 σ^2_{ts}的估值:

$$\hat{\mu}_{ts}= \hat{\sigma}^2_{ts}\sum_{i=1}^{s}\frac{n_i}{s^2_i}Y_i \tag{7}$$

$$\hat{\sigma}^2_{ts}= \frac{1}{\sum_{i=1}^{s}n_i/s^2_i} \tag{8}$$

式中: $\hat{\mu}_{ts}$即为品种均值的Bayes 估值, $\hat{\sigma}^2_{ts}$为品种均值方差的Bayes 估值。

更一般地, 若把 s 个试点, v 个品种, r 次重复(即所有试点上重复数 $n_i= r$)的区试中品种 j 在地点 i 上的平均值记为 Y_{ij}, 品种 j 的Bayes 估值记为B_j, 其方差记为S_{Bj}, 则:

$$B_j= \sum_{i=1}^{s}\frac{1}{s^2_i}Y_{ij}\Big/\sum_{i=1}^{s}\frac{1}{s^2_i} \tag{9}$$

$$S_{Bj}= \frac{1}{\sum_{i=1}^{s}r/s^2_i} \tag{10}$$

由此可见,Bayes 估值其实是品种在各试点上的多个算术平均值的加权平均值, 其权重是各试点试验误差方差的倒数, 而Bayes 估值的方差则是各试点品种算术平均值的方差的调和平均数。

1.2 试验资料与设计

为了对品种Bayes 估值的实际预测效果进行研究, 选择建国以来黄河流域具有代表性的 10 个棉花品种, 连续进行了 4 年的多点试验(表 1)。各点次均采用随机完全区组设计, 3～4 次重复, 3 行区, 小区面积 20 m^2, 种植管理和性状考察按目前国家区试标准进行。小区产量转换为每公顷产量, 单位 $kg\cdot hm^{-2}$。

表 1 试验资料及与设计

年 份	参 试 点	重 复 数	参 试 品 种
1996	沧州, 安阳, 西华, 临清	4	岱 15, 徐州 1818, 徐州 142
1997	沧州, 安阳, 西华, 临清, 菏泽	3	鲁棉 1 号, 鲁棉 6 号, 冀棉 8 号, 中棉所 12 号(为种质库原种)
1998	沧州, 安阳, 西华, 临清, 菏泽, 运城, 北京	3	中棉所 12 号, CK (为目前生产用种)
1999	沧州, 安阳, 西华, 临清, 菏泽, 运城, 北京	3	中棉所 19 号, 石远 321

2　结果与分析

利用以上试验的皮棉产量资料, 计算 1996 年各品种的算术平均值和Bayes 估值, 以之作为品种未来表现的 2 种预测性估值。Bayes 估值依据公式(9)进行计算, 该式中 s_i^2(即误差均方 MSe)通过单因素随机区组的方差分析[3]获得。然后, 计算各品种在后续 3 年(1997—1999)的平均值, 把它们作为品种未来表现的真值, 以对 1996 年的 2 种估值结果进行验证(所以在此称后 3 年的品种均值为验证值)。最后, 按验证值以及 2 种估值分别对品种进行排名, 并分别计算 2 种估值和验证值之间的绝对相差(称为预测差)。所有计算结果列于表 2。

表 2　Bayes 估值和算术平均值的预测差及品种排序比较

品　种	验证值	名次	算术平均			Bayes 估计		
			估值	名次	预测差	估值	名次	预测差
中棉所 19 号	1 147.62	1	1 169.66	2	22.04	1 158.64	2	11.02
石远 321	1 138.65	2	1 272.25	1	133.60	1 205.45	1	66.80
中棉所 12 号, CK	1 098.46	3	1 067.59	4	30.87	1 038.91	3	59.55
徐州 142	980.46	4	1 024.88	5	44.42	1 002.67	4	22.21
冀棉 8 号	895.95	5	930.38	7	34.43	913.16	6	17.21
鲁棉 6 号	826.07	6	1 078.41	3	252.33	952.24	5	126.17
中棉所 12 号(原种)	824.56	7	909.16	8	84.59	866.86	7	42.30
鲁棉 1 号	745.75	8	832.00	9	86.25	788.88	8	43.12
岱 15	634.17	9	935.53	6	301.36	784.85	9	150.68
徐州 1818	557.11	10	697.50	10	140.39	627.30	10	70.20
平均	884.88		991.73		113.03	933.90		60.93

从表 2 可见, Bayes 估值的预测差平均为 60.93 kg•hm^{-2}, 相当于验证总均值 884.88 kg•hm^{-2}的 6.88%; 而算术平均值的预测差平均为 113.03 kg•hm^{-2}, 相当于验证总均值的 12.77%。从这一点来看, Bayes 估值的预测精度高出算术平均值近 1 倍。也就是说, 用第 1 年试验的Bayes 估值来推断各品种在后 3 年的总表现, 其准确程度比用算术平均值提高了 1 倍左右。另外, 就品种排名来看, Bayes 估值的排名结果与验证值较为一致, 但算术平均值的排名结果则与验证值差别较大, 共 4 处不同, 说明Bayes 估值在品种排名上也比算术平均值准确。2 种估值特性的差异还可以从它们与验证值之间的相关系数上反映出来: Bayes 估值与验证值的相关系数为 0.963, 对品种排名的秩相关为 0.976; 算术平均值的这 2 种相关系数分别为 0.867 和 0.830。这些都说明, 在预测的意义上看, Bayes 估值的效果比算术平均值更好。

3　讨论

试验中, 品种均值的Bayes 估值比算术平均值具有更好的预测效果。这可能是因为Bayes 估计利用各试点试验误差的倒数进行加权计算, 更为充分地利用了试验中的变异信息。就其实际含义来看, Bayes 估值意味着, 某试点的试验误差越大, 其权重就越小, 那么该试点上的数值对品种总均值的贡献就越小。也就是说, 若某试点误差越大, 我们对其结果的"相信程度"越低,

这与我们的一般直觉和经验是吻合的。Bayes 估计把试验看作连续动态的统计推断, 从哲学意义上讲, 也似乎更符合"实践, 认识, 再实践, 再认识"的规律。当然, 该方法是否适宜在区试中广泛应用, 或者说适宜在何种作物何种区试中应用, 尚需在更多试验和研究来验证和判断。但从本文可看出, Bayes 统计学有其自身得特点, 若在农业试验中加以研究利用, 有可能使我们得到一些新的方法和启迪。

另外, 从Bayes 估值的最终计算公式(9)和(10)来看, Bayes 估值与试验地点的在分析中的先后顺序并无关系, 之所以把试验看作各试点依次进行, 只是为便于从Bayes 统计学的角度论述问题而已, 实际区试的试验方案中无须考虑试点的顺序问题。就计算过程而言, Bayes 估值的计算比算术平均值稍复杂一点, 但也很容易实现。并且, 由于根据公式(9)和(10)可同时计算出Bayes 估值及其方差, 所以, 在作实际品种分析时, 很容易对各品种的Bayes 均值作出区间估计和差异显著性测验。就此看来, 品种的Bayes 估值在实际应用中是简单和实用的。

最后要说明的是, 关于区试中品种均值的估计, 也存在着一些基于混合线性模型(mixed linear model)的方法, 如线性无偏预测(best linear unbiased prediction, BLUP)等[4~6]。这些方法通常主要用于处理非平衡数据的问题, 计算往往比较复杂, 目前作者正在对其作进一步的研究。

参 考 文 献

1 中山大学数学力学系. 概率论及数理统计. 北京: 高等教育出版社, 1984

2 王松贵. 线性模型的理论及应用. 合肥: 安徽教育出版社, 1986

3 莫惠栋. 农业试验统计. (第 2 版). 上海: 上海科学技术出版社, 1992

4 朱军. 遗传模型分析方法. 北京: 中国农业出版社, 1997

5 Hill R R J, Rosenberger J L. Methods for combination data from germplasm evaluation trials. Crop Science, 1985, 25: 467~470

6 Peipho H P. Best linear unbiased prediction (BLUP) for regional trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. Theor Appl Genet., 1994, 89: 647~654

附 录

已知Bayes 公式[2]如下:

$$p(\theta|x_i) = c \cdot p(x_i|\theta) \cdot p(\theta)$$

试中: $p(\theta)$ 为参数 θ 的先验分布; $p(x_i|\theta)$ 观测值的似然函数; c 是与 θ 无关的常量; $p(\theta|x_i)$ 为结合先验分布(已知信息)和观测值(试验信息)得出的 θ 的后验分布。

现在我们感兴趣的参数为平均值 μ。假设有 n 个观测值 $x_1, x_2, \cdots, x_n$ 为来自正态总体 $N(\mu, \sigma^2)$ 的随机样本, 其样本均值为 $\bar{x}$, 则这些观测值的似然函数为:

$$p(x_i|\mu) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\} = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{\sum x_i^2}{2\sigma^2}\} \exp\{-\frac{n(\mu^2-2\mu\bar{x})}{2\sigma^2}\}$$

若令 $c_1 = (2\pi\sigma^2)^{-\frac{n}{2}}) \exp\{-\frac{\sum x_i^2}{2\sigma^2}\}$, $c_2 = \exp\{-\frac{n\bar{x}^2}{2\sigma^2}\}$, 则有:

$$c_2 p(x_i|\mu) = c_1 \exp\{-\frac{n(x-\mu)}{2\sigma^2}\}$$

另外, 已知 μ 具有先验分布 $N(\mu_0, \sigma_0^2)$, 即:

$$p(\mu) = (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\}$$

令 $c_3 = (2\pi\sigma_0^2)^{-\frac{1}{2}}$, 则有:

$$c_2 p(\mu|x_i) p(\mu) = c_1 c_3 \exp\{-\frac{\mu^2}{2}(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}) + (\frac{nx}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})\mu + (\frac{nx^2}{2\sigma^2}+\frac{\mu_0^2}{2\sigma_0^2})\}$$

再令 $\frac{1}{\sigma_t^2} = (\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2})$, $\mu_t = \sigma_t^2(\frac{nx}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})$, $c_4 = \exp\{\frac{nx^2}{2\sigma^2}+\frac{\mu_0^2}{2\sigma_0^2}\}$, $c_5 = (2\pi\sigma_t^2)^{-\frac{1}{2}}\exp\{-\frac{\mu_t^2}{2\sigma_t^2}\}$, 则有:

$$c_2 p(x_i|\mu) p(\mu) = \frac{c_1 c_3 c_4}{c_5}(2\pi\sigma_t^2)^{-\frac{1}{2}} \exp\{-\frac{(\mu-\mu_t)^2}{2\sigma_t^2}\}$$

令 $c = (c_2 c_5)/(c_1 c_3 c_4)$, 因为 c 为常量, 且与 μ 无关, 所以最后得到 μ 的后验分布为:

$$p(\mu|x_i) = c \cdot p(\mu|x_i) \cdot p(\mu) = (2\pi\sigma_1^2)^{-\frac{1}{2}} \exp\{-\frac{(\mu-\mu_t)^2}{2\sigma_t^2}\}$$

由此可见, μ 服从后验的正态分布 $N(\mu_t, \sigma_t^2)$。在此, 只需用 μ_{t1}, σ_{t1}^2, σ_2^2, μ_{t2}, σ_{t2}^2, n_2 和 Y_2 分别代替以上的 μ_0, σ_0^2, σ^2, μ_t, σ_t^2, n 和 x, 文中(1)和(2)式即可得证。