应用昆虫学报 *Chinese Journal of Applied Entomology* 2013,50(4):1163 - 1169. DOI:10.7679/j.issn.2095 - 1353.2013.160

生态数据统计分析方法专题

昆虫种群数据分析及在 SPSS 软件上的实现*

董兆克[1,2] 戈 峰[1**]

(1.中国科学院动物研究所农业虫害鼠害综合治理研究国家重点实验室 北京 100101;2.北京农学院 北京 102206)

摘 要 选择合适的统计分析方法对昆虫种群分析至关重要。本文以昆虫种群数据常用的分析方法为基础,介绍了单因素方差分析、多因素方差分析、重复测量方差分析和回归分析等多种分析方法的基本原理,强调了各种分析方法的应用前提,避免误用方法导致结果判读产生偏差,并结合 SPSS 软件的使用,实现相应的分析,旨在为昆虫种群数据分析提供方法论的参考。

关键词 统计分析,单因素方差分析,多因素方差分析,重复测量方差分析,线性回归

Statistical analysis of insect population data and the use of SPSS

DONG Zhao-Ke[1,2] GE Feng[1**]

(1. State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;2. Beijing University of Agriculture, Beijing 102206, China)

Abstract It is crucial to choose the proper statistical methods when analyzing insect population data. This paper introduces the more frequently used statistical methods in insect population data analysis, including one-way ANOVA, multivariate ANOVA, repeated-measures ANOVA, and linear regression, and introduces the basic principles and premises of these methods. We emphasize the assumptions associated with various methods and provide advice to avoid biasing the results through misusing methods. Combined with the use of SPSS software, the implementation of these analyses are shown step by step. This paper provides a useful methodological reference for analyzing insect population data.

Key words statistic analysis, one-way ANOVA, multivariable ANOVA, repeated-measures ANOVA, linear regression

对田间昆虫种群数据的统计分析,是获得科学、可靠结论的基础;也是研究论文撰写中"材料与方法"所必须要有的描述。显然,选择合适的统计分析方法对昆虫种群分析至关重要。

田间试验设计是数据分析的前提。在进行野外调查、田间试验和室内试验所获取的昆虫种群数据统计分析之前,应进行科学合理的试验设计。关于试验设计方法有许多重要的参考书,如由盖钧镒(2000)主编的《试验统计方法》和明道绪(2008)主编的《田间试验与统计分析(第 2 版)》等均有很好的介绍。本文是在合理的试验设计研究结果的基础上,重点解决数据的分析问题。

数据分析通常是利用统计软件来实现,常用的统计软件有 SPSS 和 SAS 等,它们的应用领域很广泛。本文根据作者长期对昆虫种群数据分析的经验,归纳分析了昆虫种群研究中常用的方法,介绍了如何利用统计软件 SPSS 对昆虫种群数据进行分析,重点指出田间试验数据分析过程中的注意事项,避免因误用方法给结果的判读带来偏差,为开展昆虫种群数据分析提供方法论的参考。

1 数据的转换

直接由田间取样调查、收集或室内测定分析的数据为原始数据。原始数据在参与分析前需要

* 资助项目:国家自然科学基金(31200321)和国家科技支撑计划项目(2012BAD19B05)。

**通讯作者,E-mail:gef@ioz.ac.cn

收稿日期:2013-06-10,接受日期:2013-06-28

确保单位的统一，因为在试验执行中有可能受到干扰因素影响，导致处理之间取样点数量不一致。如调查苗蚜时，通常全株调查；而调查伏蚜时，通常全株上、中、下三叶调查，显然苗蚜与伏蚜的单位不一致。所以原始数据必须换算成统一的单位，如为每 n 株（同时单株或百株）的个体数量或每平方米的个体数量。消除误差的最好方法是用平均值，比如多个样点取平均值参与统计分析。另外一个需要注意的问题是避免试验过程中的伪重复，确保每个分析单元是独立的。常见的伪重复问题是把取样的重复当成了处理的重复（牛海山等，2009）。

在数据分析前最先遇到的问题是数据转换，要知道该数据是不是需要转换，怎么转换？这就需要了解数据转换的目的。参数分析（以方差分析为代表）对数据有前提要求，即数据符合正态或近似的正态分布。昆虫由于个体小、数量丰富但可能分布不均。如蚜虫发生高峰时，三叶虫量可达到上万头；而刚开始发生时，全株只有几头。因此，当数据不符合参数分析的前提条件时，就需要对数据进行转换，如果转换后的数据仍不符合参数分析要求，建议用非参数检验。当然，非参数检验的效力相比参数分析就差很多。对于数量级较大的计数数据都可以用 log 转换，通常进行 $\log_{10}(x+1)$ 的转换，公式中加 1 是为了避免出现零取对数的无效计算。数量不太大的可以用平方根转换，百分率的数据通常用反正弦平方根转换。

以蚜虫的计数数据 aphid 为例，采用 SPSS 软件进行数据转换。选择 SPSS 软件菜单 transform → compute variable，出现对话框（图 1），选数学计算 Arithmetic 内容中的 Lg10 进入计算框，点选变量 aphid 进入计算框内，然后使用计算器完成公式。转换后的数据将会自动存储为新的变量，在这里我们将其命名为 log_aphid，点击 OK 运行即可。新的变量会出现在 SPSS 数据窗口中。

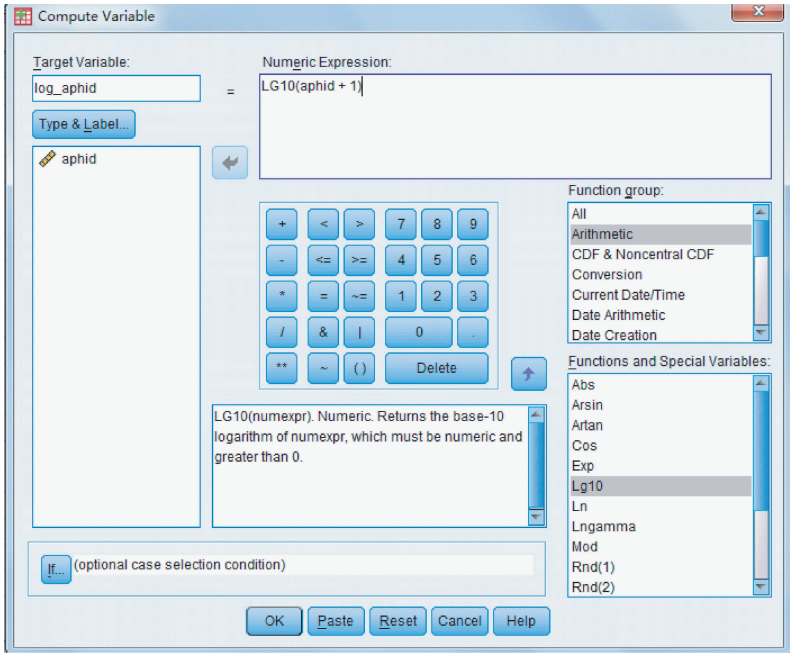

图 1　SPSS 软件中的变量计算框

Fig. 1　The compute variable frame of SPSS

2　方差分析

方差分析是根据多个组间样本均数的差别推断总体均数是否存在差别。方差分析包括单因素方法分析、多因素方差分析和多元方差分析。适应于单因素方差分析的试验设计有完全随机设计、随机区组设计和拉丁方设计等。适用于多因素方差分析有析因设计、裂区设计、正交设计等。

根据试验设计来选择合适的分析方法。使用前必须满足该方法的假定条件,即各样本随机、独立,正态分布和方差齐性。

2.1　单因素方差分析(one-way ANOVA)

单因素方差分析用于进行两组或多组间样本均数的比较。在解决实际问题中,我们还需要回答究竟哪两个之间是有差异的,这就需要进行随后的两两比较。有一系列的两两比较方法可以用于方差分析后的进一步检验。最小显著性差异(least significant difference,LSD)法是比较常用的多重比较方法,其检验敏感性高,即水平间的均值只要存在一定程度的微小差异就可能被检验出来,但它并没有对犯一类错误的概率问题加以有效控制。S-N-K 方法是一种有效划分相似性子集的方法。该方法适合于各水平观测值个数相等的情况。

方差分析是基于变异分解的原理,在单因素方差分析中,整个样本的变异可以看成由随机变异和处理因素导致的变异构成。其中,随机变异是不可避免的存在的,处理因素导致的变异是否存在是研究目的。各组内部的变异即组内变异,反映随机变异的大小;组间变异反映随机变异与处理因素的影响之和。采用一定方法来比较组内变异和组间变异的大小,如果后者远远大于前者,则说明处理因素的影响的确存在,如果两者相似,则说明该影响不存在。

以玉米蚜虫数据分析为例　表 1 列出了来自一个随机区组试验的数据,在该试验中考察玉米间作不同植物下玉米蚜虫的发生量,有 4 种不同间作方式,通过在每个区组内随机安排 4 个处理的田块次序,可以使试验简单可行并消除田块不同带来的误差。

表 1　不同间作处理的玉米蚜虫数量(头/株)

Table 1　Numbers of corn aphid in different intercropping treatment plots (per plant)

区组 Group	处理 Treatment			
	A	B	C	D
区组 1 Group 1	136.2	39.5	50.5	36.7
区组 2 Group 2	94.5	37.2	22.5	52.6
区组 3 Group 3	104.9	42.1	32.6	26.1

表 1 中每个处理田块的数据为多个取样点的平均数值。数据分析方法采用单因素方差分析。首先需要将数据按照 SPSS 软件要求录入,每列为一个变量,表 1 的数据包含 2 个变量:处理变量和响应变量(区组是重复,不作为变量)。在变量窗口 variable view 定义变量名,处理变量命名为 treatment,取值分别为 1、2、3 和 4,它们分别代表处理 A、B、C 和 D;响应变量为蚜虫数量,命名为 aphid(图 2)。蚜虫数据先进行 log 转换,新变量命名为 lg_aphid,接下来进行方差分析。

	treatment	aphid	var
1	1	136.20	
2	1	94.50	
3	1	104.90	
4	2	39.50	
5	2	37.20	
6	2	42.10	
7	3	50.50	
8	3	22.50	
9	3	32.60	
10	4	36.70	
11	4	52.60	
12	4	26.10	
13			

图 2　数据在 SPSS 中的显示

Fig. 2　Display of the data in SPSS variable window

用 SPSS 软件实现单因素方差分析的步骤:

(1)菜单 Analyze → Compare means → One-way ANOVA。

(2)Dependent list 框选入 log_aphid。

(3)Factor 框选入 treatment。

(4)窗口 Options:选择 Descriptive,Homogeneity-of-variance 和 Means plot 等。

(5)窗口 Post Hoc:选 LSD。

(6)点击运行。

SPSS 软件的一个特点是把所有结果都显示出来,初学者可能对这么多的结果感到不知所措。实际上我们只需要关注几个关键的表格。表 2 是方差齐性检验结果,$P > 0.05$,认为样本所在的各总体方差齐性。

表 2 方差齐性检验

Table 2 Test of homogeneity of variances log_aphid

Levene statistic	*df*1	*df*2	Sig.
1.384	3	8	0.316

最重要的表格就是方差分析表,如表 3 所示。

方差分析表结果显示,$F = 11.681$,$P = 0.003$。因此可认为 4 种处理下玉米蚜虫数量有极显著差异。仅分析到这一步是不够的,我们还想知道两两之间的比较,因此就需要表 4 输出的多重比较结果。

表 3 方差分析表

Table 3 ANOVA log_aphid

	Sum of squares	*df*	Mean square	*F*	Sig.
Between groups	0.511	3	0.170	11.681	0.003
Within groups	0.117	8	0.015		
Total	0.627	11			

表 4 多重比较结果

Table 4 Multiple Comparisons log_aphid LSD

(I) Treatment		(J) Treatment		Mean difference (I-J)	Std. error	Sig.	95% Confidence interval	
							Lower bound	Upper bound
Dimension2	处理 A	Dimension3	处理 B	0.43942*	0.09857	0.002	0.2121	0.6667
			处理 C	0.51101*	0.09857	0.001	0.2837	0.7383
			处理 D	0.46793*	0.09857	0.001	0.2406	0.6952
	处理 B	Dimension3	处理 A	-0.43942*	0.09857	0.002	-0.6667	-0.2121
			处理 C	0.07159	0.09857	0.488	-0.1557	0.2989
			处理 D	0.02851	0.09857	0.780	-0.1988	0.2558
	处理 C	Dimension3	处理 A	-0.51101*	0.09857	0.001	-0.7383	-0.2837
			处理 B	-0.07159	0.09857	0.488	-0.2989	0.1557
			处理 D	-0.04309	0.09857	0.674	-0.2704	0.1842
	处理 D	Dimension3	处理 A	-0.46793*	0.09857	0.001	-0.6952	-0.2406
			处理 B	-0.02851	0.09857	0.780	-0.2558	0.1988
			处理 C	0.04309	0.09857	0.674	-0.1842	0.2704

注:* 表示在 0.05 水平上差异显著。

* mean difference is significant at the 0.05 level.

两两分析的结果显示,处理 A 与其他处理之间有极明显差异,查看数据均值可以发现,处理 A 的蚜虫数量明显高于其他处理。

2.2 多因素方差分析(two-way ANOVA or three-way ANOVA)

该方法分析两个或两个以上的因素对一个变量的影响。多因素方差分析不仅能分析多个因素对观测变量的影响,还能分析多个控制因素的交互作用对观测变量是否产生影响,以及分析协方差。多因素方差分析是以方差分析的原理对分析模型进行扩展,是将总变异分解为两个或多个部分,除了一部分代表随机误差的作用外,其他部分分别代表各因素作用,通过一定方法的比较,了解某个因素对结果变量是否有明显影响。使用时必须满足多因素方差分析的假定:每个总体服从正态分布,方差齐性及观察值独立。常用的试验设计几乎都可以用多因素方差分析,如随机区组设计、裂区设计、交叉设计、析因设计等。

用 SPSS 软件实现多因素方差分析的步骤:

(1)菜单 Analyze → General liner model → Univariate。

(2)Dependent list:选入因变量,即响应变量,只能选入一个。

(3)Fixed factor:选入由试验处理因素组成的自变量。

(4)如果试验存在随机因素的影响,可以把该因素选入 Random factor 框中,另外,covariate 框中可以选入协变量,这些依据试验情况而定。

(5)根据需要在 Model 中选择模型,通常为默认;在 Post Hoc 中选择两两比较方法,点击 OK 运行。

结果中的方差分析表与上面介绍的单因素方差分析结果类似,只不过是多个因素的。在此不做详细介绍。

裂区设计(嵌套试验设计分析方法与此相同)分析时,需要特别注意分清楚主区因素和副区因素。裂区设计的特点是试验因素分两次或多次安排完成。首先安排最重要或必须最先,或材料消耗最大的因素,然后是其他因素。不能直接套用 SPSS 软件中 Univariate,需在编程窗口对程序进行进一步修改。例如,主区因素为 A 和副区因素为 B,区组为 block,响应变量为 data(图 3)。

	A	B	block	data
1	1	1	1	1.05
2	1	1	2	.82
3	1	1	3	.93
4	2	1	1	1.26
5	2	1	2	1.12
6	2	1	3	1.02
7	1	2	1	.94
8	1	2	2	.97
9	1	2	3	1.09
10	2	2	1	1.06
11	2	2	2	1.28
12	2	2	3	1.14

图 3　SPSS 数据窗口截图

Fig. 3　Screen shot of SPSS variable window

用 SPSS 软件进行裂区设计分析,步骤如下:

(1) Analyze → General liner model → Univariate。

(2)点选 data 为因变量(Dependent variable),选 A 和 B 为 Fixed factor,选 block 为 Random factor。

(3)选择合适的模型,在 Model 窗口,点 Custom,将 A 和 B 分别选入模型中,也可选它们的交互。不选 block,忽略提示。

(4)点击 Paste 进入编程窗口,在 DESIGN 行的主区因素 A 后面插入 block(A),运行即可(图 4)。

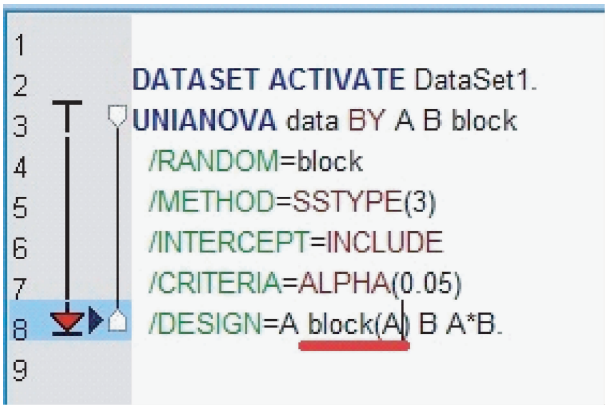

图 4　SPSS 中的编程窗口操作,下划线表示需添加的内容

Fig. 4　Programming window of SPSS, the underlining means added content

如果不区分主区因素和副区因素,软件会默认两者同等重要,这样就会给结果的判读带来偏差。

2.3　重复测量方差分析(repeated-measures ANOVA)

许多情况下研究者会对同一取样点或同一试验对象进行重复测量。比如将某种昆虫放置于不同 CO_2 水平下,测量该物种在不同水平上、不同时间的个体反应。研究者感兴趣的是反应随时间、处理水平变化的格局或形式。重复测量得到的数据可以用单因素(随机区组和裂区设计)的参数分析方法,也可以用多变量方差分析。这里侧重介绍在数据分析过程中的如何应用重复测量方差分析方法,关于重复测量试验的设计可以参考牟溥(2008)译著的《生态学实验设计与分析(第 2 版)》。

重复测量方差分析使用时必须满足的前提条件是方差-协方差矩阵的循环性。但是在某些情况下以时间为组内因子进行重复测量 ANOVA 分析

时,循环性(或球形性)的假定并不能满足。这就会导致组内因子(及其相互作用)的 *F*-统计量会膨胀,因而很可能将一个不显著的效应检测为具统计显著性。当此假定条件无法满足时,一是调整 *F* 统计的自由度使其更加保守;或者采用多因素分析。

用 SPSS 软件实现 repeated-measures ANOVA 分析的步骤:

(1)菜单 Analyze → General liner model → Repeated-measures。

(2)定义重复测量的变量名:Within-subject factor name 框输入合适的名称,Number of levels 框键入重复测量的次数,点击 Define。

(3)出现新的窗口,在 Within-subject variables 框点选重复测量的响应变量,Between subjects factor 框点选试验因素(自变量)。

(4)其他项可以根据需要选择,点击 OK 运行。

2.4 多变量方差分析(MANOVA)

多变量方差分析属于多元统计的范畴,分析一个或多个效应因子是如何影响一组反应变量的。在研究许多实际问题时,经常遇到多个响应变量的问题,如群落研究要解决多个共存物种对环境的响应。由于有些指标变量之间往往不独立,仅研究某个指标或者分别研究都不能从整体上把握问题的实质,这时候就用到多变量方差分析。MANOVA 的用途仍然是检验不同样本间是否存在显著差异,它也可以用于重复测量试验的数据分析。在这种方法中,每个水平上的响应变量被当做不同的因变量。但与重复测量 ANOVA 不同,它不要求因变量间相关性一致。它假定方差-协方差矩阵"没有结构",即对方差-协方差矩阵没有特殊要求。MANOVA 的应用有数据要求和基本假设:实验对象相互独立,所有随机影响(特别是组内或者是实验单元内的误差效应)正态分布,误差效应的方差在组内或单元内相同(即方差齐性),这些与 ANOVA 的假定相同。此外,MANOVA 要求响应变量之间有相关关系;有较大的总样本量,且每个处理有足够的重复;还假定各响应变量的联合分布为多元正态分布,且组间协方差相同。但是实际上满足每个响应变量正态分布即可,然后假定多变量联合分布为正态。

用 SPSS 软件实现 MANOVA 分析的步骤:

(1)菜单 Analyze → General liner model → Multivariate。

(2)Dependent variables 框中选入多个因变量(注:如果只选入一个变量是无法运算的)。

(3)Fixed factor 框选入试验处理因素作为自变量。

(4)其他项可以根据需要选择,点击 OK 运行。

3 回归分析

回归分析是处理两个及两个以上变量间线性依存关系的统计方法。线性回归是最常用的方法。包括只有一个自变量的一元线性回归和有多个自变量的多元线性回归。

3.1 一元线性回归

一元线性回归的方程为:$y = a + bx$

用 SPSS 进行一元线性回归的操作步骤:

(1)菜单 Analyze → Regression → Linear。

(2)Dependent 框中选入因变量 *y*(响应变量)。

(3)Independent 框中选入自变量 *x*。

(4)在方法 Method 中采用默认的选项 Enter,又称作强制进入法,表示要求系统在建立回归方程时把所选中的全部自变量都保留在方程中。点击 OK 运行。

3.2 多元线性回归

多元线性回归在 SPSS 中的操作与一元线性回归相同,Independent 框中选入自变量 *x*1、*x*2 或更多变量。可供选择的几种变量进入回归方程的方法:

Enter 选项,强行进入法,即所选择的自变量全部进入回归模型,默认选项。

Remove 选项,消去法,建立回归方程时,根据设定的条件剔除部分自变量。

Forward 选项,向前选择法,根据在 Option 对话框中所设定的判据,从无自变量开始,在拟合过程中,对被选择的自变量进行方差分析,每次加入一个 F 值最大的变量,直到所有符合判据的变量都进入模型为止。

Backward 选项,向后剔除法,根据在 Option 对话框中所设定的判据,先建立全模型,然后根据设

置的判据,每次剔除一个使方差分析中的 F 值最小的自变量,直到回归方程中不再含有不符合判据的自变量为止。

Stepwise 选项,逐步进入法,是向前选择法和向后剔除法的结合。根据在 Option 对话框中所设定的纳入和排除标准进行变量筛选。首先计算各自变量对 y 的贡献大小,选择对贡献最大的进入回归方程。重新计算各自变量对 y 的贡献;然后根据向后剔除法,将模型中 F 值最小的且符合剔除判据的变量剔除模型,重复进行直到回归方程中的自变量均符合进入模型的判据,模型外的自变量都不符合进入模型的判据为止。

参考文献(References)

盖钧镒, 2000. 试验统计方法. 北京:中国农业出版社. 1 – 407.

明道绪, 2008. 田间试验与统计分析(第 2 版). 北京:科学出版社. 1 – 284.

牛海山, 崔骁勇, 汪诗平, 王艳芬, 2009. 生态学试验设计与解释中常见问题. 生态学报, 29(7):3901 – 3910.

牟溥译, 2008. 生态学实验设计与分析(第 2 版). 北京:高等教育出版社. 1 – 330.