技术与方法

# 常用的生物统计方法及其 R 语言实现*

刘学聪[1**] 李欣海[1, 2***]

（1. 中国科学院大学生命科学学院，北京 100049；2. 中国科学院动物研究所，北京 100101）

**摘 要** 统计分析是科学研究中一个极其重要的环节。本文以昆虫学研究为实例，利用模拟数据，总结了 14 种常用的生物统计方法及其 R 语言实现，重点强调了如何根据科学问题和样本数据的具体情形选取合适的统计方法。这些统计方法包括可用于均值比较分析的符号检验、Wilcoxon 符号秩检验、*t*-检验、Wilcoxon 秩和检验、Kruskal-Wallis 检验、Nemenyi 检验、Tukey 检验、Friedman 检验、单因素方差分析、重复测量方差分析和可用于相关性分析的卡方检验、Fisher 精确检验、Spearman 秩相关分析、Pearson 相关分析，可为生物统计或 R 语言基础薄弱的昆虫学工作者提供参考。

**关键词** 参数统计；非参数统计；均值比较；相关性分析；R 语言

## Basic biostatistical tests and their R codes

LIU Xue-Cong[1**] LI Xin-Hai[1,2***]

(1. College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China;
2. Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China)

**Abstract** Statistical analysis is important for many kinds of entomological research. Based on simulated data, this article reviews 14 basic statistical tests frequently used in entomological studies and their corresponding R codes, emphasizing how to choose the appropriate test for a given investigation and data type. These statistical tests include the sign test, Wilcoxon signed rank test, *t*-test, Wilcoxon rank sum test, Kruskal-Wallis test, Nemenyi test, Tukey test, Friedman test, one-way analysis of variance, and repeated measures analysis of variance, to compare means or medians, and the Chi-square test, Fisher's exact test, Spearman rank correlation test and Pearson correlation test, to analyze associations or correlations between variables. This article provides a useful reference for entomologists with limited knowledge of biostatistics or the R language.

**Key words** parametric test; nonparametric test; means (or medians) comparison; association (or correlation) analysis; R language

统计分析是科学研究中一个极其重要的环节。在统计学的意义上，通过室内实验或野外调查所获得的原始数据只是所研究变量的样本，需要对样本数据进行统计分析，来推断样本所来自总体（也就是所研究的变量）的情况，才能得出科学的结论。统计分析可分为参数统计和非参数统计两类（Wilcox，2003；Zar，2010）。参数统计的有效性依赖于一定的前提条件，读者比较熟悉的一个条件是研究的变量应该符合正态分布，且研究变量的分布状况是可以通过样本数据来推断的（Sokal and Rohlf，1995；Wilcox，2003；Zar，2010）。非参数统计则对变量的分布无要求。但是，非参数统计并不是基于具体的数值，而是基于数值的大小及其顺序（即“秩：Rank”），并

*资助项目 Supported projects：国家自然科学基金（31872235，31970432，31772479）；中央高校基本科研业务费专项资金（Y95401VXX2）；神农架金丝猴保育生物学湖北省重点实验室开放基金（SNJKL2017001）

**第一作者 First author，E-mail：xuecongliu@ucas.ac.cn

***通讯作者 Corresponding author，E-mail：lixh@ioz.ac.cn

收稿日期 Received：2020-12-14；接受日期 Accepted：2021-01-04

没有利用样本数据的全部信息，因而非参数统计往往比参数统计的检验功效（Testing power）要低（Wilcox，2003）。显然，在进行统计分析时，应首选参数统计，必要时进行数据转换，包括对数转换、平方根转换等。

在昆虫学研究中，经常会遇到与以下研究实例相类似的问题。广翅蜡蝉科 Ricaniidae 昆虫的翅膀形态在种间、性别间是否存在差异，这就涉及均值比较分析（潘鹏亮等，2020）；胡萝卜微管蚜 *Semiaphis heraclei* 发育速率随温度梯度的变化趋势，这就需要进行相关性分析（王堇秀等，2016）；通过气象因子预测昆虫种群动态，这就涉及回归分析（欧阳芳和戈峰，2013；费海泽等，2014）。统计分析通常是用专业软件来实现的，常用的统计软件有 R、SPSS、SAS 等，其中，R 具有免费开源、语言简单易学、功能灵活机动等优点，受到越来越多的关注（Crawley，2013）。

本文基于作者长期对动物生态学数据的分析经验，以昆虫学研究为实例，利用模拟数据，总结了 14 种常用的统计方法及其 R 语言实现。重点强调了如何根据科学问题和样本数据的具体情形选取合适的统计方法，以期为生物统计或 R 语言基础薄弱的昆虫学工作者开展科学研究提供参考。与普通软件一样，下载 R 软件后（建议采用最新版本 4.0.3，下载网址：https://r-project.org），在电脑上安装，打开后就是编写程序的窗口“R Console”，每条程序结束后，按回车键即可运行，运行结果随即也会在这个窗口显示，如果运行结果是图，则会在自动打开的“R Graphics”窗口中显示。涉及的统计方法可用于均值比较分析和相关性分析这两类问题。

# 1　均值比较分析

对单样本、两个独立样本、两个相关样本、多个（三个或三个以上）独立样本、多个相关样本 5 种情形进行均值比较分析，每种情形都有参数统计和非参数统计可选择。对于前 3 种情形而言，参数统计首先要求至少要有一定的样本量（用 $n$ 表示），如果 $n$ 太小，就很难从样本数据去推断总体的分布状况，即样本数据的分布缺乏代表性。在这种情况下，建议采用非参数统计。那么，$n$ 至少要多大，通过样本数据去推断总体的分布才能算比较有效呢？据作者所知，应该没有明确的、统一的答案，按作者经验，$n$ 至少应该>10。如果 $n$ 足够大，根据中心极限定理（Central limit theorem），用样本去推断的总体不论服从何种分布，其样本平均数（Mean）的分布都接近于正态分布（Sokal and Rohlf，1995；Zar，2010），从而用参数统计进行均值比较分析被认为是有效的。那么，$n$ 多大才算足够大呢？也没有明确的、统一的答案，有的资料推荐为 $n$≥30（李春喜等，2013）。当 $n$ 较大但不足够大时，如果符合正态分布，建议采用参数统计，否则，建议采用非参数统计。对于后两种情形（多个独立样本、多个相关样本），参数统计都有各自的特殊要求，在后面单独介绍。

## 1.1　单样本

现有某种昆虫的前翅周长数据“parameter1”（图 1），试问该昆虫的前翅周长是否与 60 mm 有明显差异？因 $n$=14，需要检验分布的正态性。正态 Q-Q 图显示，数据与正态分布偏离较大（图 2），正态性检验也不接受正态分布（Shapiro-Wilk test：$P$=0.048）（图 1），因此，采用非参数统计。可选用符号检验（Sign test），符号检验是二项式检验的一种特殊形式（即检验概率为 0.5），图 1 结果显示，该昆虫的前翅周长与 60 mm 有明显差异（$P$=0.039）。也可选用 Wilcoxon 符号秩检验（Wilcoxon signed rank test），图 1 结果显示，该昆虫的前翅周长与 60 mm 有明显差异（$P$=0.005）。比较 2 种检验的结果会发现，Wilcoxon 符号秩检验比符号检验具有更高的检验功效，即产生更低的 $P$ 值。

如果把样本量扩充到 $n$=66，得数据“parameter2”（图 3）。因样本量大，根据中心极限定理，即使与正态分布有所偏离，也可直接采用参数统计：单样本 $t$-检验（One sample $t$-test）。图 3 结果显示，该昆虫的前翅周长与 60 mm 有明显差异（$t$=10.09，$P$<0.001）。值得特别注意

```
R Console
> #建议首先为 R 指定一个工作路径，把数据文件都放在这里，
> #下面再读取数据就可以省略文件路径了，
> #指定路径的命令和格式：setwd("D:/My Documents/...")
>
> #数值少，可直接输入数据
> parameter1 <- c(59, 60, 70, 75, 78, 79, 80, 59, 60, 62, 64, 72, 74, 75)
> parameter1  #数据显示
 [1] 59 60 70 75 78 79 80 59 60 62 64 72 74 75
> qqnorm(parameter1)  #正态 Q-Q 图
> qqline(parameter1, lty=2)  #添加正态分布标准线
> shapiro.test(parameter1)  #Shapiro-Wilk 正态性检验

        Shapiro-Wilk normality test

data:  parameter1
W = 0.87445, p-value = 0.04848

> diff <- parameter1 - 60  #计算与 60 的差值
> diff  #差值显示
 [1] -1  0 10 15 18 19 20 -1  0  2  4 12 14 15
> binom.test(2, 12)  #符号检验：忽略=0 的 2 个数，<0 的数有 2 个

        Exact binomial test

data:  2 and 12
number of successes = 2, number of trials = 12, p-value = 0.03857
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.02086253 0.48413775
sample estimates:
probability of success
             0.1666667
> binom.test(10, 12)  #符号检验：>0 的数有 10 个，与上一个命令的结果相同
> wilcox.test(parameter1, mu=60)  #Wilcoxon 符号秩检验(单样本)

        Wilcoxon signed rank test with continuity correction

data:  parameter1
V = 75, p-value = 0.00532
alternative hypothesis: true location is not equal to 60
> wilcox.test(diff)  #Wilcoxon 符号秩检验(单样本)：与上一个命令的结果相同
> wilcox.test(diff ~ 1)  #Wilcoxon 符号秩检验(单样本)：与上一个命令的结果相同
```

**图 1　正态 Q-Q 图、Shapiro-Wilk 正态性检验、符号检验和 Wilcoxon 符号秩检验的运行命令**

**Fig. 1　R codes for normal Q-Q plot, Shapiro-Wilk normality test, sign test, and Wilcoxon signed rank test**

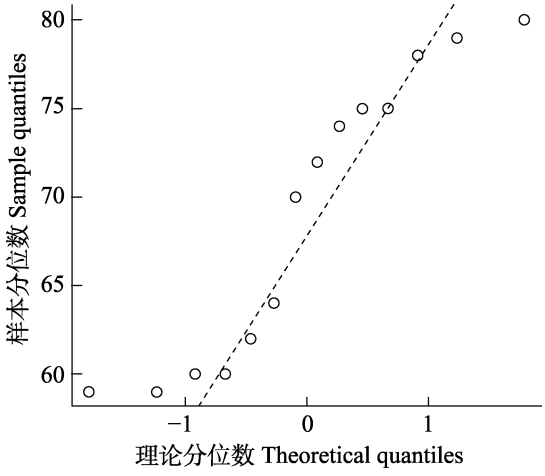



**图 2　正态 Q-Q 图**

**Fig. 2　Normal Q-Q plot**

的是，虽然在样本量大的情况下，可直接采用参数统计，但还是强烈建议在进行单样本 *t*-检验之前，用正态 Q-Q 图和正态性检验查看一下数据的分布情况，以保证统计结果确实可靠有效。

### 1.2　两个独立样本

现有某种昆虫雌性和雄性的前翅周长数据“sex1”（图 4），试问该昆虫的前翅周长是否有性别差异？因样本量太小（$n_1$=7，$n_2$=8），检验分布的正态性无意义，直接采用非参数统计：Wilcoxon 秩和检验（Wilcoxon rank sum test）。计算结果图 4 显示，该昆虫的前翅周长无性别差异（$W$=36，$P$=0.397）。

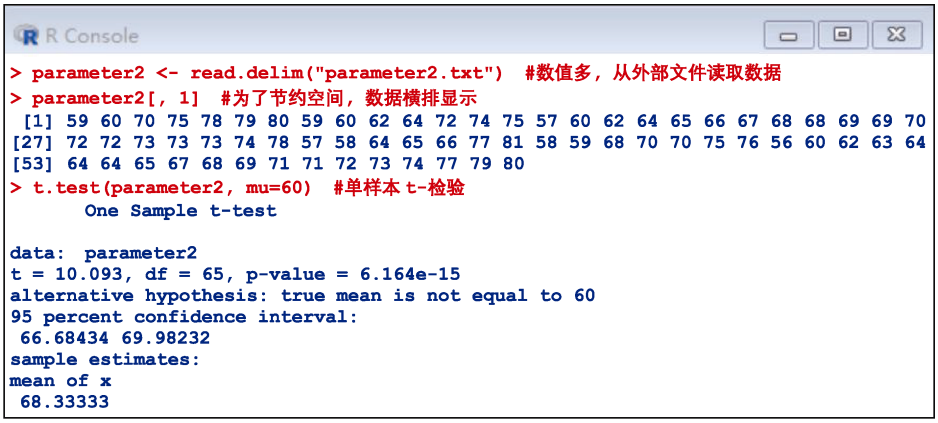



**图 3　单样本 *t*-检验的运行命令**

**Fig. 3　R codes for one sample *t*-test**

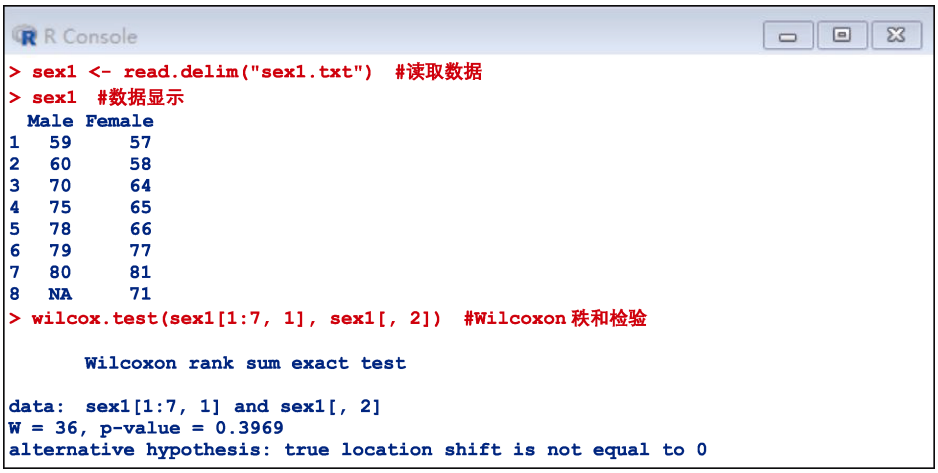



**图 4　Wilcoxon 秩和检验的运行命令**

**Fig. 4　R codes for Wilcoxon rank sum test**

如果将样本量扩充到 $n_1$=33、$n_2$=34，得数据“sex2”（图 5）。由于两个变量的样本量都够大，可直接采用参数统计：双样本 *t*-检验（Two samples *t*-test）。图 5 结果显示，该昆虫的前翅周长在性别之间没有差异（$t$=0.22，$P$=0.824）。强烈建议在进行双样本 *t*-检验之前，用正态 Q-Q 图和正态性检验查看一下数据的分布情况，以保证统计结果确实可靠有效。

### 1.3　两个相关样本

现有某种昆虫一些个体的左、右前翅周长数据“forewing1”（图 6），试问该昆虫的左、右前翅周长是否有差异？因样本量太小（$n_1$=$n_2$=7），检验分布的正态性无意义，直接采用非参数统计。可选用符号检验，此时，符号检验的计算过程是，先计算两个样本数值的差值，然后把差值组成的变量与 0 作比较。计算结果图 6 显示，该昆虫的左、右前翅周长无差异（$P$=0.453）。也可采用 Wilcoxon 符号秩检验，结果图 6 显示，该昆虫的左、右前翅周长无差异（$P$=0.306）。

如果把样本量扩充到 $n_1$=$n_2$=33，得数据“forewing2”（图 7）。由于样本量大，可直接采用参数统计：配对样本 *t*-检验（Paired samples *t*-test）。图 7 的计算结果显示，该昆虫的左、右前翅周长没有差异（$t$=1.14，$P$=0.263）。与上述符号检验的计算过程相似，配对样本 *t*-检验是先计算两个样本数值的差值，然后把差值组成的变量与 0 作比较（即单样本 *t*-检验）。强烈建议在进行配对样本 *t*-检验之前，用正态 Q-Q 图和正态性检验查看一下数据的分布情况，以保证统计结

```
R Console
> sex2 <- read.delim("sex2.txt")  #读取数据
> sex2 <- t(sex2)  #数据由竖排转为横排，仅为了节约显示空间
> sex2  #数据显示
       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
Male     59   60   70   75   78   79   80   59   60    62    64    72    74    75
Female   57   58   64   65   66   77   81   58   59    68    70    70    75    76
       [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26]
Male      57    60    62    64    65    66    67    68    68    69    69    70
Female    56    60    62    63    64    64    64    65    67    68    69    71
       [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34]
Male      72    72    73    73    73    74    78    NA
Female    71    72    73    74    77    79    80    80
> #由于在两个变量的方差相等和不相等时，双样本 t-检验的计算方式不一样，
> #需先检验方差是否相等
> var.test(sex2[1, 1:33], sex2[2, ])  #用 F 检验查看方差是否相等

        F test to compare two variances

data:  sex2[1, 1:33] and sex2[2, ]
F = 0.81043, num df = 32, denom df = 33, p-value = 0.554
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4032551 1.6349303
sample estimates:
ratio of variances
         0.8104281

> #结果显示方差相等(F=0.81, P=0.554)
>
> #双样本 t-检验：因 R 默认方差不等，需参数 var.equal=TRUE
> t.test(sex2[1, 1:33], sex2[2, ], var.equal=TRUE)

        Two Sample t-test

data:  sex2[1, 1:33] and sex2[2, ]
t = 0.22286, df = 65, p-value = 0.8243
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.973117  3.719997
sample estimates:
mean of x mean of y
 68.69697  68.32353
```

图 5　双样本 *t*-检验的运行命令

**Fig. 5　R codes for two samples *t*-test**

```
R Console
> forewing1 <- read.delim("forewing1.txt")  #读取数据
> forewing1  #数据显示
  Individuals Left Right
1           1   59    57
2           2   60    58
3           3   70    69
4           4   75    70
5           5   78    81
6           6   79    77
7           7   80    81
> diff <- (forewing1[, 2] - forewing1[, 3])  #计算每对数的差值
> diff  #差值显示
[1]  2  2  1  5 -3  2 -1
> binom.test(2, 7)  #符号检验：<0 的数有 2 个

        Exact binomial test

data:  2 and 7
number of successes = 2, number of trials = 7, p-value = 0.4531
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.03669257 0.70957914
sample estimates:
probability of success
             0.2857143
> #Wilcoxon 符号秩检验(双样本)
> wilcox.test(forewing1[, 2], forewing1[, 3], paired=TRUE)

        Wilcoxon signed rank test with continuity correction

data:  forewing1[, 2] and forewing1[, 3]
V = 20.5, p-value = 0.3061
alternative hypothesis: true location shift is not equal to 0
```

图 6　符号检验和 Wilcoxon 符号秩检验的运行命令

**Fig. 6　R codes for sign test and Wilcoxon signed rank test**

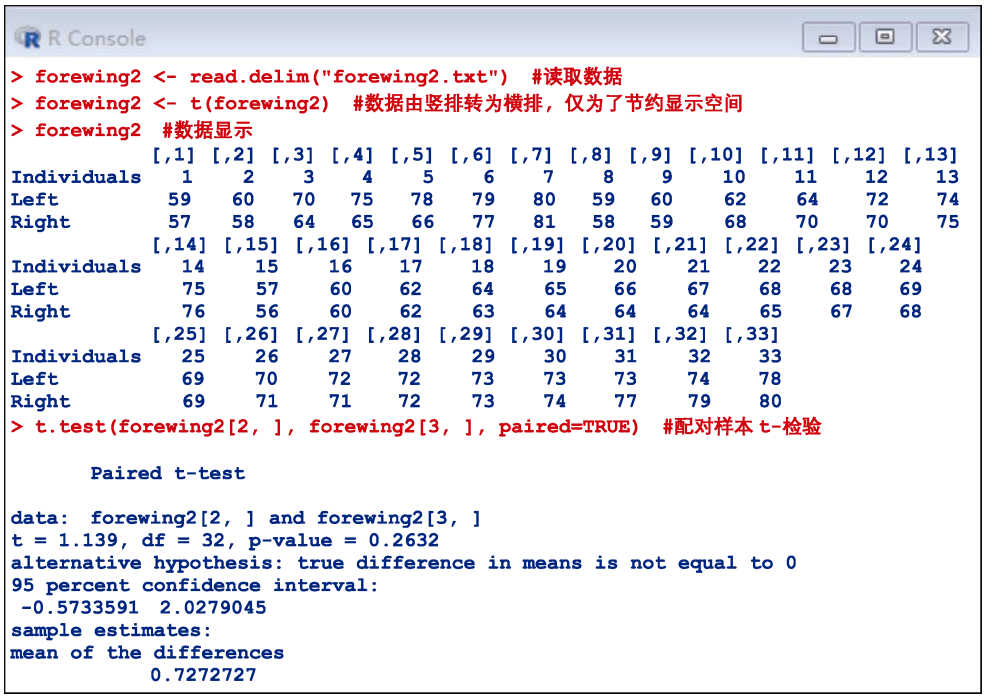



**图 7　配对样本 *t*-检验的运行命令**

**Fig. 7　R codes for paired samples *t*-test**

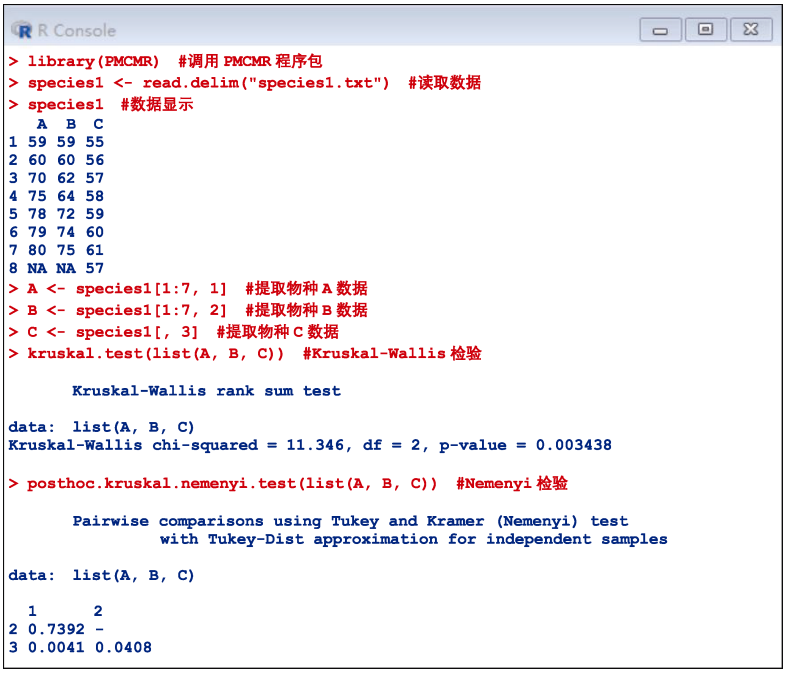
果确实可靠有效。

### 1.4　多个独立样本

如上所述，对于多个独立变量的均值比较，也可选择参数统计（单因素方差分析：One-way analysis of variance）和非参数统计（Kruskal-Wallis 检验：Kruskal-Wallis test）。单因素方差分析要求各个变量都符合正态分布，然而，随着各变量 $n_i$ 的增大，数据的非正态分布对方差分析的影响变小（Zar，2010）。同时，单因素方差分析还要求方差齐性（即各变量方差相等），但在各变量的 $n_i$ 相等或基本相等的情况下，方差不齐对方差分析的影响有限（Zar，2010）。例如，Myers 和 Well（2003）曾报道，当 $n_i \geq 5$、最大方差不大于最小方差的四倍时，I 型错误仅膨胀 0.02（显著性水平：0.05）。因此，在有些情况下，很难确定参数检验和非参数检验哪一个才是更好的选择。作者根据经验谨慎地建议，如果所有的变量 $n_i \geq 10$，可考虑采用单因素方差分析，否则，考虑采用 Kruskal-Wallis 检验。

例如，现有 3 种昆虫的前翅周长数据“species1”（图 8），试问前翅周长是否有种间差异？如果有，哪 2 种昆虫之间有差异？由于样本量小（$n_1$=7，$n_2$=7，$n_3$=8），建议直接采用 Kruskal-Wallis 检验。图 8 结果显示，3 种昆虫的前翅周长有差异（$\chi^2$=11.35，$P$=0.003）。为此，需要进一步作两两比较，可用 Nemenyi 检验（Nemenyi test），其结果显示，物种 A 和物种 B 之间无差异（$P$=0.739），物种 A（$P$=0.004）和物种 B（$P$=0.041）与物种 C 之间有差异。也可用 Wilcoxon 秩和检验作两两比较，只是显著性水平 α 需用 Bonferroni 方法进行调整，即 α/m（m：检验次数；本例中为 3 次）。当然，还有另外的两两比较方法，请参考其他资料。

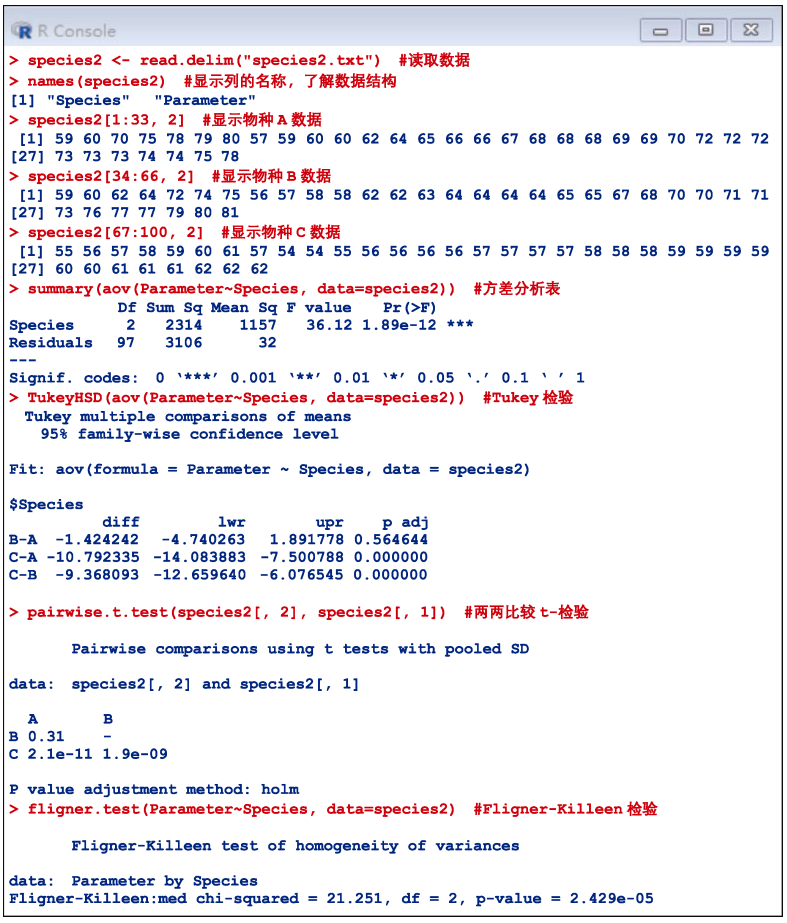
如果把样本量扩充到 $n_1$=33、$n_2$=33、$n_3$=34，得数据“species2”（图 9）。由于样本量大，可直接采用单因素方差分析。计算结果图 9 显示，3 种昆虫的前翅周长有差异（$F$=36.12，$P$<0.001）。进一步可用 Tukey 检验（Tukey test）作两两比较，其结果表明，物种 A 和物种 B 之间无差异（$P$=0.565），物种 A（$P$<0.001）和物种 B（$P$<0.001）与物种 C 之间有差异。也可用 *t*-检验作两两比较，其结论与 Tukey 检验一致（图 9）。也可选用另外的两两比较方法，请参考其他资



```
R Console
> library(PMCMR)  #调用 PMCMR 程序包
> species1 <- read.delim("species1.txt")  #读取数据
> species1  #数据显示
   A  B  C
1 59 59 55
2 60 60 56
3 70 62 57
4 75 64 58
5 78 72 59
6 79 74 60
7 80 75 61
8 NA NA 57
> A <- species1[1:7, 1]  #提取物种 A 数据
> B <- species1[1:7, 2]  #提取物种 B 数据
> C <- species1[, 3]  #提取物种 C 数据
> kruskal.test(list(A, B, C))  #Kruskal-Wallis 检验

        Kruskal-Wallis rank sum test

data:  list(A, B, C)
Kruskal-Wallis chi-squared = 11.346, df = 2, p-value = 0.003438

> posthoc.kruskal.nemenyi.test(list(A, B, C))  #Nemenyi 检验

        Pairwise comparisons using Tukey and Kramer (Nemenyi) test
                 with Tukey-Dist approximation for independent samples

data:  list(A, B, C)

  1      2
2 0.7392 -
3 0.0041 0.0408
```

图 8　Kruskal-Wallis 检验和 Nemenyi 检验的运行命令

**Fig. 8　R codes for Kruskal-Wallis test and Nemenyi test**



```
R Console
> species2 <- read.delim("species2.txt")  #读取数据
> names(species2)  #显示列的名称，了解数据结构
[1] "Species"   "Parameter"
> species2[1:33, 2]  #显示物种 A 数据
 [1] 59 60 70 75 78 79 80 57 59 60 60 62 64 65 66 66 67 68 68 68 69 69 70 72 72 72
[27] 73 73 73 74 74 75 78
> species2[34:66, 2]  #显示物种 B 数据
 [1] 59 60 62 64 72 74 75 56 57 58 58 62 62 63 64 64 64 64 65 65 67 68 70 70 71 71
[27] 73 76 77 77 79 80 81
> species2[67:100, 2]  #显示物种 C 数据
 [1] 55 56 57 58 59 60 61 57 54 54 55 56 56 56 56 57 57 57 57 58 58 58 59 59 59 59
[27] 60 60 61 61 61 62 62 62
> summary(aov(Parameter~Species, data=species2))  #方差分析表
             Df Sum Sq Mean Sq F value    Pr(>F)
Species      2   2314    1157   36.12 1.89e-12 ***
Residuals   97   3106      32
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(aov(Parameter~Species, data=species2))  #Tukey 检验
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Parameter ~ Species, data = species2)

$Species
          diff        lwr        upr     p adj
B-A  -1.424242  -4.740263   1.891778 0.564644
C-A -10.792335 -14.083883  -7.500788 0.000000
C-B  -9.368093 -12.659640  -6.076545 0.000000

> pairwise.t.test(species2[, 2], species2[, 1])  #两两比较 t-检验

        Pairwise comparisons using t tests with pooled SD

data:  species2[, 2] and species2[, 1]

  A       B
B 0.31    -
C 2.1e-11 1.9e-09

P value adjustment method: holm
> fligner.test(Parameter~Species, data=species2)  #Fligner-Killeen 检验

        Fligner-Killeen test of homogeneity of variances

data:  Parameter by Species
Fligner-Killeen:med chi-squared = 21.251, df = 2, p-value = 2.429e-05
```

图 9　单因素方差分析、Tukey 检验、两两比较 *t*-检验和 Fligner-Killeen 检验的运行命令

**Fig. 9　R codes for one-way analysis of variance, Tukey test, pairwise *t*-test, and Fligner-Killeen test**

料。强烈建议在进行单因素方差分析之前，用正态 Q-Q 图和正态性检验查看一下数据的分布情况，用 Fligner-Killeen 检验（Fligner-Killeen test）查看一下方差齐性状况（图 9），以保证统计结果确实可靠有效。

### 1.5　多个相关样本

多个相关样本数据通常是由随机区组（Randomized block design）或重复测量（Repeated measures design）实验设计获得的。对于这类数据的均值比较分析，也可选择参数统计和非参数统计，其中，重复测量方差分析（Repeated measures analysis of variance）应是最常用的参数统计方法（Zar，2010）。与单因素方差分析一样，重复测量方差分析也假定各变量呈正态分布和方差齐性，另外，重复测量方差分析还要求复合对称性（Compound symmetry），即任何两个变量之间的相关程度相等，复合对称性与方差齐性合称球形（Sphericity）。对于不符合球形条件的数据，可通过调整自由度来矫正方差分析的结果（Zar，2010），但一般不会影响最后结论。根据经验，作者谨慎地建议，如果 $n_i \geq 10$，考虑采用参数统计，否则，考虑采用非参数统计。

众所周知，绝大多数昆虫都有左、右前翅和左、右后翅 4 个翅膀。现有某昆虫一些个体的 4 个翅膀的周长数据“fourwing1”（图 10），试问该昆虫 4 个翅膀的周长有无差异？如果有，哪些翅膀之间有差异？因 $n_1=n_2=n_3=n_4=7$，样本量小，建议采用非参数统计：Friedman 检验（Friedman test）。结果图 10 显示，该昆虫 4 个翅膀的周长有差异（$\chi^2=16.89$，$P=0.001$）。可用 Nemenyi 检验进一步作两两比较，其结果显示，左、右前翅之间（$P=0.997$）和左、右后翅之间（$P=0.997$）无差异，左前翅与左后翅（$P=0.036$）和右后翅（$P=0.020$）之间、以及右前翅与左后翅（$P=0.020$）和右后翅（$P=0.010$）之间有差异。也可用符号检验或 Wilcoxon 符号秩检验作两两比较，显著性水平需用 Bonferroni 方法进行调整。还有另外的两两比较方法，请参考其他资料。

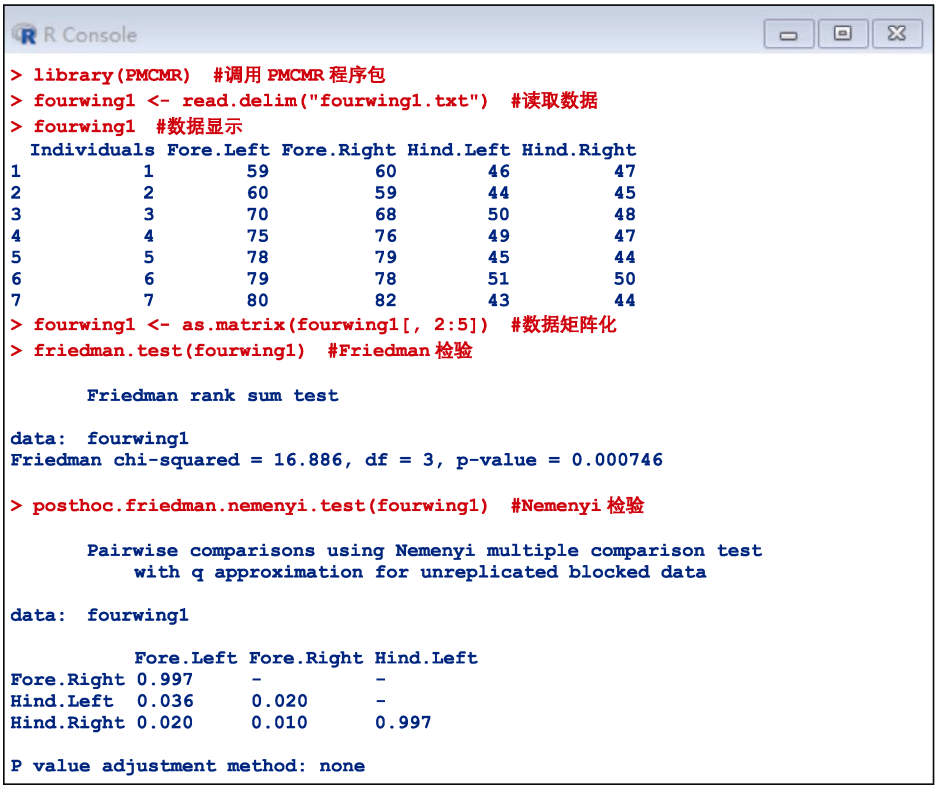



**图 10　Friedman 检验和 Nemenyi 检验的运行命令**

**Fig. 10　R codes for Friedman test and Nemenyi test**

如果将样本量扩充到 $n_1$=$n_2$=$n_3$=$n_4$=15，得数据“fourwing2”（图 11）。可考虑采用重复测量方差分析，其结果图 11 显示，该昆虫 4 个翅膀的周长有差异（$F$=67.62，$P$<0.001）。进一步可

```
R Console
> fourwing2 <- read.delim("fourwing2.txt")  #读取数据
> fourwing2  #数据显示
   Individuals Fore.Left Fore.Right Hind.Left Hind.Right
1            1        59         60        56         57
2            2        60         59        54         55
3            3        70         68        59         58
4            4        75         72        59         57
5            5        78         75        55         54
6            6        79         74        61         60
7            7        80         78        55         54
8            8        79         77        59         56
9            9        78         77        57         55
10          10        70         73        60         61
11          11        76         78        62         60
12          12        63         65        55         54
13          13        66         67        56         57
14          14        72         69        56         55
15          15        71         72        61         62
> new.fourwing2 <- stack(fourwing2[, 2:5])  #创建新数据框，包含周长和翅膀两列
> colnames(new.fourwing2) <- c("Parameter", "Wing")  #为数据框的列赋予名称
> Individuals <- rep(1:15, 4)  #创建个体变量
> Individuals  #数据显示
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15  1  2  3  4  5  6  7  8  9 10 11
[27] 12 13 14 15  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15  1  2  3  4  5  6  7
[53]  8  9 10 11 12 13 14 15
> Individuals <- as.factor(Individuals)  #设为因素
> new.fourwing2<- cbind(new.fourwing2, Individuals)  #列的合并，此时数据框包含三列
> summary(aov(Parameter~Wing+Error(Individuals), data=new.fourwing2))  #方差分析表

Error: Individuals
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 14  812.8   58.06

Error: Within
          Df Sum Sq Mean Sq F value   Pr(>F)
Wing       3 2948.1   982.7   67.62 4.01e-16 ***
Residuals 42  610.4    14.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> Wing <- as.factor(new.fourwing2[, 2])  #提取数据并设为因素
> Parameter <- (new.fourwing2[, 1])  #提取数据
> pairwise.t.test(Parameter, Wing, paired=TRUE)  #两两比较配对 t-检验

       Pairwise comparisons using paired t tests

data:  Parameter and Wing

           Fore.Left Fore.Right Hind.Left
Fore.Right 0.21      -         -
Hind.Left  3.2e-06   1.0e-06   -
Hind.Right 7.0e-06   3.2e-06   0.15

P value adjustment method: holm
> fourwing2 <- as.matrix(fourwing2)  #数据矩阵化
> mlmfit <- lm(fourwing2~1)  #多元线性模型
> mauchly.test(mlmfit, X=~1)  #Mauchly 球形检验

       Mauchly's test of sphericity
       Contrasts orthogonal to
       ~1


data:  SSD matrix from lm(formula = fourwing2 ~ 1)
W = 0.0071345, p-value = 1.083e-09

> library(nlme)  #调用 nlme 程序包
> #线性混合效应模型
> summary(model <- lme(Parameter~Wing, random=~1|Individuals, data=new.fourwing2))
```

**图 11　重复测量方差分析、两两比较配对样本 *t*-检验、Mauchly 检验和线性混合效应模型的运行命令**

**Fig. 11　R codes for repeated measures analysis of variance, pairwise paired samples *t*-test, Mauchly’s test, and linear mixed effects model**

用配对样本 *t*-检验作两两比较，结果显示，左、右前翅之间（*P*=0.21）和左、右后翅之间（*P*=0.15）无差异，左前翅与左后翅（*P*<0.001）和右后翅（*P*<0.001）之间、以及右前翅与左后翅（*P*<0.001）和右后翅（*P*<0.001）之间有差异。还有另外的两两比较方法，请参考其他资料。强烈建议在进行重复测量方差分析之前，用正态 Q-Q 图和正态性检验查看一下数据的分布情况，用 Mauchly 检验（Mauchly's test）检查一下球形条件（图 11），以保证统计结果确实可靠有效。

对于多个相关变量的均值比较分析，线性混合效应模型（Linear mixed effects model）也是一个合适的选择（图 11）。

## 2　相关性分析

只介绍两个变量之间的相关性分析，包括两个分类变量之间和两个连续变量之间的相关性分析。对于两个连续变量之间的相关性分析，只介绍线性相关关系。

### 2.1　分类变量

两个分类变量之间的相关性（或独立性）分析通常是用卡方检验（Chi-square test）来实现的。例如，用 3 种杀虫剂（X、Y、Z）分别喷杀 50 只害虫后的存活数据“count1”（图 12），试问 3 种杀虫剂的杀虫效果是否一致，或者杀虫剂种类和害虫死亡率有无关系？结果显示，3 种杀虫剂的杀虫效果不一致（$\chi^2$=16.67，*P*<0.001）。

需要注意的是，当两个变量都是二分变量时，如果在它们组成的 2×2 表中四个单元格的总数<20，或者有任何一个单元格的期望值<5，因样本量太小，通过卡方检验计算得出的统计量就

```
R Console
> count1 <- matrix(c(40, 10, 30, 20, 20, 30), nrow=2)  #输入数据
> rownames(count1) <- c("Died", "Survived")  #为行赋予名称
> colnames(count1) <- c("X", "Y", "Z")  #为列赋予名称
> count1  #数据显示
          X  Y  Z
Died     40 30 20
Survived 10 20 30
> chisq.test(count1)  #卡方检验

        Pearson's Chi-squared test

data:  count1
X-squared = 16.667, df = 2, p-value = 0.0002404

> count2 <- matrix(c(8, 42, 1, 49), nrow=2)  #输入数据
> rownames(count2) <- c("Died", "Survived")  #为行赋予名称
> colnames(count2) <- c("X", "Control")  #为列赋予名称
> count2  #数据显示
          X Control
Died      8       1
Survived 42      49
> (8+42)*(8+1)/(8+42+1+49)  #单元格 X*Died 的期望值
[1] 4.5
> (8+42)*(42+49)/(8+42+1+49)  #单元格 X*Survived 的期望值
[1] 45.5
> (1+49)*(8+1)/(8+42+1+49)  #单元格 Control*Died 的期望值
[1] 4.5
> (1+49)*(42+49)/(8+42+1+49)  #单元格 Control*Survived 的期望值
[1] 45.5
> fisher.test(count2)  #Fisher 精确检验

        Fisher's Exact Test for Count Data

data:  count2
p-value = 0.03086
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.150323 421.521088
sample estimates:
odds ratio
  9.165817
```

图 12　卡方检验和 Fisher 精确检验的运行命令

**Fig. 12　R codes for Chi-square test and Fisher's exact test**

不一定服从卡方分布，因此卡方检验不再合适，这时应采用 Fisher 精确检验( Fisher's exact test )。例如，用杀虫剂 X 和清水对照（Control）分别喷杀 50 只害虫后的存活数据“count2”（图 12），试问该杀虫剂的杀虫效果是否显著？在该例中，有两个单元格的期望值<5( 图 12 )，应采用 Fisher 精确检验，其结果显示，该杀虫剂有明显的杀虫效果（$P$=0.031）。

## 2.2 连续变量

两个连续变量之间的线性相关性分析可采用参数统计（即 Pearson 相关分析：Pearson's correlation test）和非参数统计（最常见的是 Spearman 秩相关或 Spearman 等级相关分析：Spearman's rank correlation test）。与均值比较分析相似，Pearson 相关分析要求数据呈双变量正态分布（Bivariate normal distribution）。然而，在实际研究中，不符合双变量正态分布的情形是普遍的。当两个变量相关性不强时，数据非正态分布的影响比较小，但是，当相关性较强时，数据非正态分布的影响就会较大（Zar，2010）。根据经验，作者谨慎地建议，当 $n$<10 时，可认为很难从样本数据去推断总体的分布状况，建议采用 Spearman 秩相关分析；当 $n$>10 时，如果没有可严重影响线性相关系数的数值，即使与双变量正态分布有所偏离，也可考虑采用 Pearson 相关分析，如果存在这样的数值，考虑采用 Spearman 秩相关分析。

现有某种昆虫在 5 种不同温度梯度的发育速率数据“rate”（图 13），试问该昆虫的发育速率与温度梯度是否相关？因样本量小（$n$=5），故采用 Spearman 秩相关分析。强烈建议在分析前，先用散点图查看一下两个变量的相关情况（图 14）。图 13 结果显示，该昆

```
R Console
> rate <- read.delim("rate.txt")  #读取数据
> rate  #数据显示
  Temperature Development.Rate
1          19              0.2
2          22              0.3
3          25              0.4
4          28              0.5
5          31              0.7
> Temperature <- rate[, 1]  #提取温度数据
> Rate <- rate[, 2]  #提取发育速率数据
> plot(Temperature, Rate)  #散点图
> cor.test(Temperature, Rate, method="spearm")  #Spearman 秩相关

        Spearman's rank correlation rho

data:  Temperature and Rate
S = 4.4409e-15, p-value = 0.01667
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
  1
```

**图 13 散点图和 Spearman 秩相关分析的运行命令**

**Fig. 13 R codes for scatter plot and Spearman's rank correlation test**

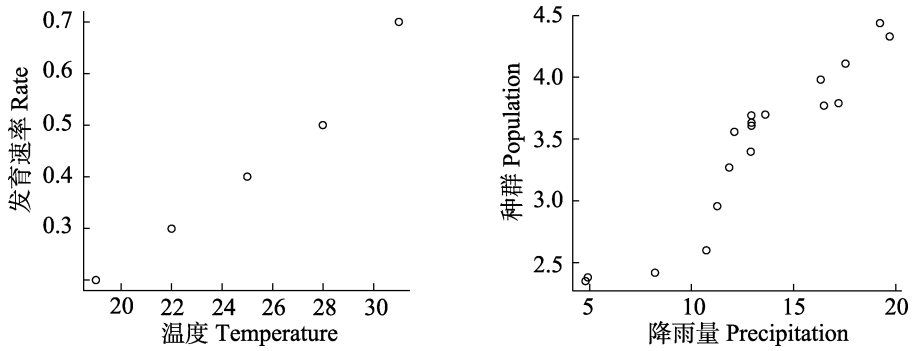



**图 14 散点图**

**Fig. 14 Scatter plots**

虫的发育速率与温度梯度呈正相关（$r$>0.99，$P$=0.017）。

现有某地某种昆虫种群数量和降雨量共 18 年的历史数据“population”（图 15），试问该昆虫种群数量与降雨量是否相关？散点图显示没有数值可以严重影响线性相关系数（图 14），况且也不拒绝双变量正态分布（Shapiro-Wilk test：$P$=0.097）（图 15），故采用 Pearson 相关分析。图 15 结果显示，昆虫种群数量与降雨量正相关（$r$=0.94，$P$<0.001）。

```
R Console
> population <- read.delim("population.txt")  #读取数据
> #把数据由竖排转为横排，因多变量正态分布检验要求矩阵至少包括三列
> population <- t(population)
> population  #数据显示
                    [,1]  [,2]  [,3]  [,4]  [,5]  [,6] [,7]  [,8]  [,9] [,10]
Sqrt.Precipitation 19.68 19.21 17.52 17.18 16.46 16.31 13.6 12.92 12.92 12.92
Log.Population      4.33  4.44  4.11  3.79  3.77  3.98  3.7  3.63  3.69  3.61
                   [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18]
Sqrt.Precipitation 12.88 12.08 11.83 11.24 10.72  8.19  4.90  4.80
Log.Population      3.40  3.56  3.27  2.96  2.60  2.42  2.38  2.35
> Precipitation <- population[1, ]  #提取降雨量数据
> Population <- population[2, ]  #提取种群数据
> plot(Precipitation, Population)  #散点图
> library(mvnormtest)  #调用多变量正态分布检验程序包
> mshapiro.test(population)  #Shapiro-Wilk 多变量正态分布检验

        Shapiro-Wilk normality test

data:  Z
W = 0.91296, p-value = 0.09702

> cor.test(Precipitation, Population)  #Pearson 相关

        Pearson's product-moment correlation

data:  Precipitation and Population
t = 11.494, df = 16, p-value = 3.831e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8542301 0.9794444
sample estimates:
      cor
0.9444427
```

**图 15　Shapiro-Wilk 多变量正态分布检验和 Pearson 相关分析的运行命令**

**Fig. 15　R codes for Shapiro-Wilk multivariate normality test and Pearson's correlation test**

## 3　讨论

本文总结了 14 种常用的生物统计方法，具体包括两方面内容：根据研究的科学问题和样本数据的具体情形选取合适的统计方法、这些统计方法的 R 运行命令，期望为生物统计或 R 语言基础薄弱的昆虫学工作者提供参考。这些统计方法可用于均值比较分析和相关性分析，虽然回归分析在昆虫研究中也是常用的（欧阳芳和戈峰，2013；费海泽等，2014），但是这类问题一般不会涉及参数统计或非参数统计的选择，这也是本文没有介绍回归分析的一个主要原因。各统计检验的 R 运行命令都可以添加多个参数（Optional arguments），包括双尾还是单尾检验、显著性水平调整方法等内容，读者可以通过在某个统计检验函数前加“?”或者“help()”，来搜索该统计检验可以设置哪些参数，比如“?cor.test”、“help(t.test)”。因 R 软件具有众多优点，近年来受到了越来越多的关注，如果想进一步学习 R 语言，《The R Book》应是一个很好的参考资料（Crawley，2013）。

其实，针对同样的数据，参数统计和非参数统计都能计算得出结果，只是因为非参数统计没有利用数据的全部信息，从而检验功效往往较低，因此在条件允许的情况下，尽可能采用参数统计，使宝贵的原始数据得到充分利用（Wilcox，

2003；Zar，2010）。统计学家在发展参数统计的算法时，预先设定了一些前提条件，比如数据的正态分布和方差齐性等，但在实际应用过程中，即使不完全满足这些前提条件，在很多时候统计效果也是可以接受的，只有在严重违背这些前提条件，导致统计结果不可信的时候，才采用非参数统计（Zar，2010）。鉴于此，有时很难确定参数统计和非参数统计到底哪一个才是更好的选择，本文中的建议只是作者根据工作经验谨慎提出的，并不一定都完全正确，还请读者根据具体的科学问题和数据情形做出合适的决定。

在科学研究中，应该尽可能地扩大样本量。虽然本文没有介绍如何获得有效可靠的原始数据，但是样本量大是采用参数统计的重要前提之一。另外，在统计分析前，应该特别注意是否有异常值的情况。异常值是指那些与绝大多数数值比起来显得特别大或特别小的数值，或者看起来严重不符合规律的数值。异常值是导致数据严重偏离参数统计前提条件的主要原因之一，对统计结果的影响也非常大（Zar，2010）。例如，一个异常值就可能会较大地改变 Pearson 相关系数，而对 Spearman 秩相关系数的影响较小。本文中提到的，即使数据不满足参数统计的前提条件，有时也可采用参数统计的建议，其实主要是基于没有异常值的情况下提出的。制图是检查数据是否有异常值的最直观途径，这也是文中多次强调，在统计分析前先查看正态 Q-Q 图和散点图的原因。异常值可能是实验错误造成的，这就需要纠正或舍弃，异常值也可能是正确的，这时应考虑数据转换或采用非参数统计。

## 参考文献 (References)


Crawley MJ, 2013. The R Book (2nd Edition). Chichester: John Wiley & Sons, Ltd. 1–975.

Fei HZ, Wang HB, Kong XB, Zhang Z, Zhang SF, Song XG, 2014. Selection and prediction of meteorological factors correlated with *Dendrolimus punctatus* outbreak. *Journal of Northeast Forestry University*, 42(1): 136–140. [费海泽, 王鸿斌, 孔祥波, 张真, 张苏芳, 宋雄刚, 2014. 马尾松毛虫发生相关气象因子筛选及预测. 东北林业大学学报, 42(1): 136–140.]

Li CX, Jiang LN, Shao Y, Zhang DJ, 2013. Biostatistics (5th Edition). Beijing: Science Press. 47–48. [李春喜, 姜丽娜, 邵云, 张戴静, 2013. 生物统计学(第五版). 北京: 科学出版社. 47–48.]

Myers JL, Well AD, 2003. Research Design and Statistical Analysis (2nd Edition). Mahwah, New Jersey: Lawrence Earlbaum Associates. 221.

Ouyang F, Ge F, 2013. Nonlinear analysis of insect population dynamics based on generalized additive models and statistical computing using R. *Chinese Journal of Applied Entomology*, 50(4): 1170–1177. [欧阳芳, 戈峰, 2013. 基于广义可加模型的昆虫种群动态非线性分析及 R 语言实现. 应用昆虫学报, 50(4): 1170–1177.]

Pan PL, Hong F, Chen JH, Liu HM, Yin XM, Xiong JW, 2020. Extraction and analysis of numerical characteristics from forewings of three plant hopper species (Homoptera: Ricaniidae). *Chinese Journal of Applied Entomology*, 57(4): 980–987. [潘鹏亮, 洪枫, 陈俊华, 刘红梅, 尹新明, 熊建伟, 2020. 三种广翅蜡蝉前翅形态数值特征提取与分析. 应用昆虫学报, 57(4): 980–987.]

Sokal RR, Rohlf FJ, 1995. Biometry (3rd Edition). New York: W. H. Freeman and Company. 109–111, 128–136.

Wang JX, Li XJ, Wang N, 2016. Influence of temperature on the development, survival and reproduction of *Semiaphis heraclei* (Takahashi). *Chinese Journal of Applied Entomology*, 53(3): 564–573. [王堇秀, 李学军, 王宁, 2016. 温度对胡萝卜微管蚜生长发育繁殖的影响. 应用昆虫学报, 53(3): 564–573.]

Wilcox RR, 2003. Applying Contemporary Statistical Techniques. San Diego, California: Academic Press. 39–43, 557–608,

Zar JH, 2010. Biostatistical Analysis (5th Edition). Upper Saddle River, New Jersey: Prentice-Hall, Inc. 66–74, 189–225, 270–277.