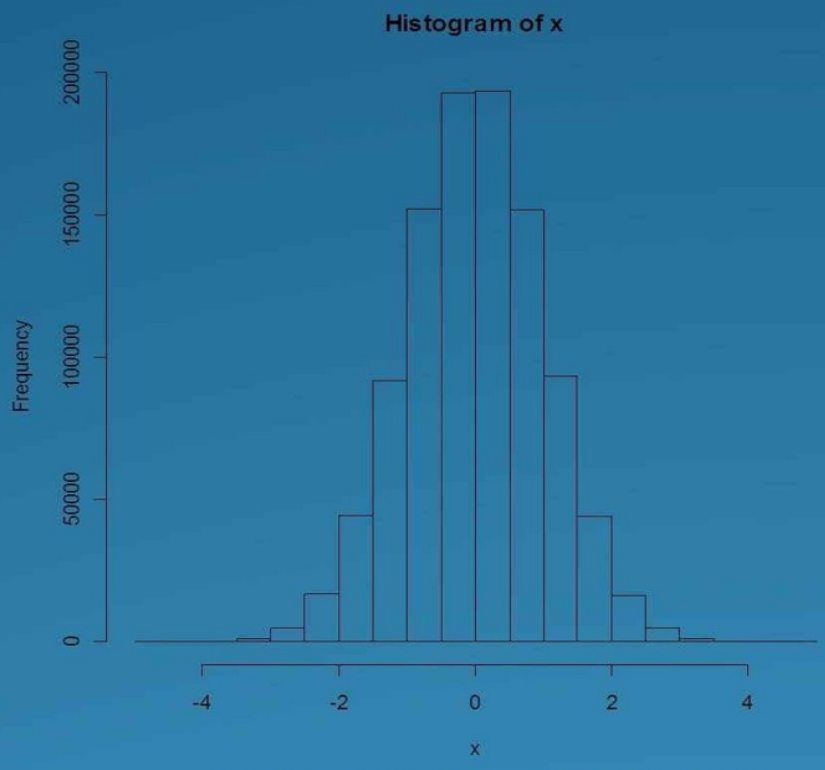
100 STATISTICAL TESTS IN R



N.D.LEWIS

Easy R Series
Heather Hills Press

100 STATISTICAL TESTS IN R

What to choose, how to easily calculate, with over 300 illustrations and examples

N.D Lewis

Heather Hills Press

http://www.AusCov.com

TABLE OF CONTENTS

Test 51 Cochran-Mantel-Haenszel test

Test 52 McNemar's test

Test 53 Equal means in a one-way layout with equal variances

Test 54 Welch-test for more than two samples

Test 55 Kruskal Wallis rank sum test

Test 56 Friedman’s test

Test 57 Quade test

Test 58 D’ Agostino test of skewness

Test 59 Anscombe-Glynn test of kurtosis

Test 60 Bonett-Seier test of kurtosis

Test 61 Shapiro-Wilk test

Test 62 Kolmogorov-Smirnov test of normality

Test 63 Jarque-Bera test

Test 64 D’ Agostino test

Test 65 Anderson-Darling test of normality

Test 66 Cramer-von Mises test

Test 67 Lilliefors test

Test 68 Shapiro-Francia test

Test 69 Mardia's test of multivariate normality

Test 70 Kolomogorov – Smirnov test for goodness of fit

Test 71 Anderson-Darling goodness of fit test

Test 72 Two-sample Kolmogorov-Smirnov test

Test 73 Anderson-Darling multiple sample goodness of fit test

Test 74 Brunner-Munzel generalized Wilcoxon Test

Test 75 Dixon’s Q test

Test 76 Chi-squared test for outliers

Test 77 Bonferroni outlier test

Test 78 Grubbs test

Test 79 Goldfeld-Quandt test for heteroscedasticity

Test 80 Breusch-Pagan test for heteroscedasticity

Test 81 Harrison-McCabe test for heteroskedasticity

Test 82 Harvey-Collier test for linearity

Test 83 Ramsey Reset test

Test 84 White neural network test

Test 85 Augmented Dickey-Fuller test

Test 86 Phillips-Perron test

Test 87 Phillips-Ouliaris test

Test 88 Kwiatkowski-Phillips-Schmidt-Shin test

Test 89 Elliott, Rothenberg & Stock test

Test 90 Schmidt - Phillips test

Test 91 Zivot and Andrews test

Test 92 Grambsch-Therneau test of proportionality

Test 93 Mantel-Haenszel log-rank test

Test 94 Peto and Peto test

Test 95 Kuiper's test of uniformity

Test 96 Rao's spacing test of uniformity

Test 97 Rayleigh test of uniformity

Test 98 Watson's goodness of fit test

Test 99 Watson's two-sample test of homogeneity

Test 100 Rao's test for homogeneity

Test 101 Pearson Chi square test

FORWARD

On numerous occasions, researchers in a wide variety of subject areas, have asked how do I carry out a particular statistical test? The answer often involved programming complicated formulas into spreadsheets and looking up test statistics in tabulations of probability distributions. With the rise of R, statistical testing is now easier than ever. 100 Statistical Tests in R is designed to give you rapid access to one hundred of the most popular statistical tests. It shows you, step by step, how to carry out these tests in the free and popular R statistical package. Compared to other books, it has:

•Breadth rather than depth. It is a guidebook, not a cookbook.

•Words rather than math. It has few equations.

•Illustrations and examples rather than recipes and formulas.

Who is it for?

100 Statistical Tests in R, as with all books in the Easy R series, came out of the desire to put statistical tools in the hands of the practitioner. The material is therefore designed to be used by the applied researcher whose primary focus is on their subject matter rather than mathematical lemmas or statistical theory. Examples of each test are clearly described and can be typed directly into R as printed on the page. To accelerate your research ideas, over three hundred published applications of statistical tests across engineering, science, and the social sciences are contained in these pages. These illustrative applications cover a vast range of disciplines incorporating numerous diverse topics such as the angular analysis of tree roots, Angelman syndrome, breastfeeding at baby friendly hospitals, comparing cardiovascular interventions, computing in those over 50, dog-human communication, effects of t'ai chi on balance, environmental forensics, the randomness of the universe, emotional speech, the solar orientation of sandhoppers, horses concept of people, hematopoietic stem cell transplantation, idiopathic clubfoot in Sweden, prudent sperm use, men's artistic gymnastics, software defects, stalagmite lamina chronologies, sexual conflict in insects, South London house prices, Texas hold'em poker, vampire calls, and more! Comprehensive references are given at the end of each test. In keeping with the zeitgeist of R, copies of all of the papers discussed in this text are available for are free.

New to R?

New users to R can use this book easily and without any prior knowledge. This is best achieved by typing in the examples as they are given and

reading the comments which follow the result of a test. Copies of R and free tutorial guides for beginners can be downloaded at http://www.r-project.org/

N.D Lewis

P.S. If you have any questions about this text or statistical testing in general you can email me directly at nd@AusCov.com . I'd be delighted to hear from you. To obtain additional resources on R and announcements of other products in the Easy R series please visit us at http://www.AusCov.com

HOW TO GET THE MOST FROM THIS BOOK

There are at least five ways to use this book to boost your productivity. First, you can dip into it as an efficient reference tool. Flip to the test you need and quickly see how to calculate it in R. For best results type in the example given in the text, examine the results, and then adjust the example to your own data. Second, browse through the three hundred applications and illustrations to help stimulate your own research ideas. Third, you may have already collected data and have a question in mind such as "is this timeseries useful in forecasting another timeseries?" Look up a suitable statistical test given your research question. Forth, by typing the numerous examples, you will strengthen you knowledge and understanding of both statistical testing and R. Finally, use the classification of tests given below to determine which types of test are most suitable for your data.

Correlation and causality test numbers 1,2,3,4,5,6,7,8,9,10,11

One sample tests for the mean and median test numbers 12,13,14

Two sample tests for the mean and median test numbers 15,16,17,18,19,20,21,22,23,24

Randomness and independence test numbers 25,26,27,28,29,30,31

Difference in scale parameters test numbers 32,35

Homogeneity of variances test numbers 33,34,35,36,37,38,39,40,41

Rates and proportions test numbers 2,43,44,45,46,47,48

Count data test numbers 49,50,51,52

Central tendency for three or more samples test numbers 53,54,55,56,57,57,59,60

Normality of sample test numbers 61,62,63,64,65,66,67,68,69

Differences in distribution test numbers 70,71,72,73

Stochastic Equality test numbers 74

Outliers in sample test numbers 39,69,75,76,77,78

Heteroscedasticity test numbers 79,80,81

Linearity test numbers 82,83,84

Unit Roots test numbers 85,86,87,88,89,90,91

Survival analysis test numbers 92,93,94

Circular data test numbers 95,96,97,98,99,100

Each section begins with the question the statistical test addresses. This is followed by a brief guide explaining when to use the test. Three applications from the published literature are then discussed and an example of the test using R is illustrated.

We follow the R convention of giving the function used for the test statistic followed in braces by the R package required to use the function. For example correlationTest{fbasics}refers to the function correlationTest in the fbasics package.

If a package mentioned in the text in not installed on your machine you can download it by typing install.packages("package_name"). For example to download the Fbasics package you would type in the R console:

>install.packages("fbasics")

Once a package is installed, you must call it before typing in the example given in the text. You do this by typing in the R console:

>require(package_name)

You only need to type this once, at the start of your R session. For example, to call the fbasics package you would type:

>require(fbasics)

The fbasics package is now ready for use.

Let's walk through an example. The function resettest{lmtest} can be used to perform the Ramsey RESET test. If the package lmtest is not installed or

your machine you would enter:

>install.packages("lmtest")

To access the function resettest you would type

>require(lmtest)

You are now ready to perform the Ramsey RESET test. Let's give it a go right now! Enter the following data, collected, on three variables:

>dep=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,3794,3959,404

>ind.1=c(75,78,80,82,84,88,93,97,99,104,109,115,120,127)

>ind.2=c(5,8,0,2,4,8,3,7,9,10,10,15,12,12)

We begin by building a simple linear regression model.

>model <- lm(dep~ind.1+ind.2)

Now, we will use the RESET test to assess whether we should include second or third powers of the independent variables - ind.1 and ind.2. We can do this by typing:

> resettest(model, power=2:3, type="regressor")

R will respond by displaying the following:

RESET test

data: model

RESET = 1.6564, df1 = 4, df2 = 7, p-value = 0.2626

Throughout this text we use the 5% level of significance as our guide to reject the null hypothesis. This simply means if the p-value reported by R is less than 0.05 we reject the null hypothesis. Since, in this example, the p-value is greater than 5% (p-value = 0.2626), we do not reject the null hypothesis of linearity. It's that simple! Refer back to this section to refresh your memory as needed.

Now let's get started!

Back to Table of Contents

TEST 1 PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT T-TEST

Question the test addresses

Is the sample Pearson product moment correlation coefficient between two variables significantly different from zero?

When to use the test?

To assess the null hypothesis of zero correlation between two variables. Both variables are measured on either an interval or ratio scale. However, they do not need to be measured on the same scale, e.g. one variable can be ratio and one can be interval. Both variables are assumed to be a paired random sample, approximately normally distributed, their joint distribution is bivariate normal, and the relationship is linear.

Practical Applications

Sports Science: Banister's training impulse and Edwards training load are two methods, based on heart rate, commonly used to assess training intensity of an athlete's workout. Haddad et al (2012) study the convergent validity of these two methods for young taekwondo athletes. They use the Pearson product moment correlation coefficient to assess convergent validity. The correlation between the two methods was 0.89 with a p-value less than 0.05. The null hypothesis of no correlation between the two measures of training load was rejected.

Ophthalmology: Kakinoki et al (2012) compare the correlation between the macular thicknesses in diabetic macular edema measured by two different types of optical coherence tomography – spectral domain optical coherence tomography and time domain optical coherence tomography. Pearson's product moment correlation for the measure of macular thickness between the two techniques was 0.977, and significant with a p-value less than 0.001. The correlation between the best corrected visual acuity and retinal thickness measured by both techniques was 0.34, and significant with a p-value less than 0.05.

Environmental Science: Nabegu and Mustapha (2012) use the product moment correlation to explore the relationship between eight categories of solid waste in Kano Metropolis located in Northwestern Nigeria. The eight categories – food scrap, paper- cardboard, textile rubber, metals, plastic materials, glass, ash and vegetable. They find a negative correlation between food scrap and metals of -0.853 with an associated p-value of less

than 0.01. The null hypothesis of zero correlation between food scrap and metals is rejected.

How to calculate in R

Both cor.test{stats} and correlationTest{fbasics} can be used to perform this test.

Example: Two correlated samples

Enter the following data

```
> x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1)

> y <- c( 2.6,  3.1,  2.5,  5.0,  3.6,  4.0,  5.2,  2.8,  3.8)
```

The test statistic can be calculated for the above data by typing:

```
>cor.test(x,y,method="pearson",alternative="two.sided",conf.level = 0.95)

        Pearson's product-moment correlation

data:  x and y

t = 1.8411, df = 7, p-value = 0.1082

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

 -0.1497426  0.8955795

sample estimates:

      cor

0.5711816
```

The correlation between x and y is reported as 0.571. Since the p-value is 0.1082 and greater than the critical value of 0.05 do not reject the null hypothesis of zero correlation. The function also reports the 95% confidence interval as -0.149 to 0.895. It crosses zero, do not reject the null hypothesis. Note to specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater"). As an alternative we could type

```
> correlationTest(x, y)
```

Title:

Pearson's Correlation Test

Test Results:

PARAMETER:

Degrees of Freedom: 7

SAMPLE ESTIMATES:

Correlation: 0.5712

STATISTIC:

t: 1.8411

P VALUE:

Alternative Two-Sided: 0.1082

Alternative Less: 0.9459

Alternative Greater: 0.05409

CONFIDENCE INTERVAL:

Two-Sided: -0.1497, 0.8956

Less: -1, 0.867

Greater: -0.0222, 1

Notice correlationTestreports the p-value for all three alternative hypothesizes (less than, greater than and two sided). In all cases, since the p-value is greater than 0.05, do not reject the null hypothesis. The function also reports the 95% confidence interval as -0.149 to 0.895. It crosses zero, do not reject the null hypothesis.

References

Haddad, Monoem; Chaouachi, Anis; Castagna, Carlo; Wong, Del P; Chamar Karim. (2012). The Convergent Validity between Two Objective Methods for Quantifying Training Load in Young Taekwondo Athletes .Journal of Strength & Conditioning Research. Volume 26 - Issue 1 - pp 206-209.

Kakinoki ,M., Miyake, T., Sawada, O., Sawada,T. ,Kawamura,H., Ohji, M (2012).Comparison of Macular Thickness in Diabetic Macular Edema Using Spectral-Domain Optical Coherence Tomography and Time-Domain Optica

Coherence Tomography. Journal of Ophthalmology, volume 2012.

Nabegu,A.B., Mustapha,A. (2012). Using Person Product Momen Correlation to explore the relationship between different categories of Municipal solid waste in Kano Metropolis, Northwestern Nigeria. Journal o Environment and Earth Science. Volume 2. No.4. p63-67.

Back to Table of Contents

TEST 2 SPEARMAN RANK CORRELATION TEST

Question the test addresses

Is the Spearman rank correlation coefficient between two variables significantly different from zero?

When to use the test?

To assess the null hypothesis of zero correlation between two variables. A paired random sample of ordinal or ranked data; or when the data is continuous and it is unreasonable to assume the variables are normally distributed. The relationship between the variables is assumed to be linear.

Practical Applications

Physical Activity: A random sample of 177 healthy Norwegian women was recruited into a study by Borch et al (2012). The researchers compared a self-administered physical activity questionnaire to various measures of physical activity obtained from heart rate and movement sensors. The Spearman rank correlation ranged between 0.36 and 0.46 with a p-value < 0.001. The null hypothesis of no correlation between the self-administered physical activity questionnaire and objective measures obtained from heart rate and movement sensors was rejected.

Environmental Forensics: Gautheir (2001) use Spearman rank correlation to detect monotonic trends in chemical concentrations with time and space in order to evaluate the effectiveness of natural attenuation. Benzene and other chemical concentrations were recorded quarterly and then semi-annually at two petrol station wells over a period of 3.5 years. The Spearman rank correlation was -0.685 and -0.430 for the first and second well respectively. The authors used a 10% level of significance. The reported p-values were less than 0.1 for each well. The null hypothesis of no correlation between chemical concentration with time and space was rejected.

Investing: Elton and Gruber (2001) investigate the relationship between marginal tax rates of the marginal stockholder and the firms dividend yield and payout ratio. They examined all stocks listed on the New York Stock Exchange that paid a dividend during April 1, 1966 to March 31, 1967. The Spearman rank correlation between the marginal tax rates of the marginal stock holder and the dividend yield was 0.9152 with a p-value less than 0.01. The null hypothesis of no correlation between the marginal tax rates of the marginal stock holder and the dividend yield was rejected. They also

found the Spearman rank correlation between the marginal tax rates of the marginal stock holder and the payout ratio was 0.7939 with a p-value less than 0.01. The null hypothesis of no correlation between the marginal tax rates of the marginal stock holder and the payout ratio was also rejected.

How to calculate in R

Both cor.test{stats} and spearman.test{pspearman} can be used to perform this test.

Example: using cor.test

Enter the following data

> x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1)

> y <- c(2.6, 3.1, 2.5, 5.0, 3.6, 4.0, 5.2, 2.8, 3.8)

The function cor.test performs the calculation. The test statistic can be calculated for the above data by typing:

>cor.test(x,y,method="spearman",alternative="two.sided")

Spearman's rank correlation rho

data: x and y

S = 48, p-value = 0.0968

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

0.6

The Spearman rank correlation between x and y is 0.6. Since the p-value is 0.0968 and greater than the critical value of 0.05, do not reject the null hypothesis. Note to specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater").

Example: using spearman.test

Enter the following data

> x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1)

> y <- c(2.6, 3.1, 2.5, 5.0, 3.6, 4.0, 5.2, 2.8, 3.8)

The test statistic can be calculated for the above data by typing:

```
> spearman.test(x,y,alternative="two.sided",approximation ="exact")

        Spearman's rank correlation rho

data:  x and y

S = 48, p-value = 0.0968

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

0.6
```

The Spearman rank correlation between x and y is 0.6. Since the p-value is 0.0968 and greater than the critical value of 0.05, do not reject the null hypothesis. Note, spearman.test has three types of approximation – “exact”, “AS89” and “t-distribution”. For a sample size of 22 or less use “exact”. For larger sample sizes use “AS89” or “t-distribution”. To specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater").

References

Borch , Kristin B., Ekelund,Ulf. , Brage, Søren., Lun, Eiliv. (2012). Criterior validity of a 10-category scale for ranking physical activity in Norwegian women. International Journal of Behavioral Nutrition and Physical Activity . Volume 9:2.

Elton, J.T. and Gruber, M.J. (2001). Marginal Stockholder Tax Rates and the Clientele Effect. Journal of Economics and Statistics. Volume 52(1), pages 68-74.

Gautheir, Thomas D. (2012). Detecting Trends Using Spearman's Rank Correlation Coefficient. Environmental Forensics. Volume 2, Issue 4, pages 359-362.

Back to Table of Contents

TEST 3 KENDALL'S TAU CORRELATION COEFFICIENT TEST

Question the test addresses

Is the Kendall tau correlation coefficient between two variables significantly different from zero?

When to use the test?

To assess the null hypothesis of zero tau correlation between two variables. Your sample consists of a paired random sample of ordinal or ranked data; or when the data is continuous and it is unreasonable to assume the variables are normally distributed. The relationship between the variables is assumed to be linear.

Practical Applications

Pediatrics: Kyle et al (2012) use Kendall's tau to measure the correlation between an index of multiple deprivation and length of hospitalization in England for the years 2006/07. They find no statistical significant relationship between 0 to 3 day hospitalizations, Kendall's tau correlation = 0.42 (p-value = 0.089). However, for 4 or more day hospitalizations they find a significant relationship, Kendall's tau correlation = 0.64 (p-value =0.009).

Sports Science: Dayaratna and Miller (2012) test the null hypothesis that goals scored and goals allowed in North American ice hockey are independent. For the Anaheim Ducks and seasons 2008/09, 2009/10 and 2010/11 they report tau correlation of 0.075,-0.105 and 0.008 respectively. The associated p-values were 0.156, 0.078 and 0.450 respectively. The null hypothesis of no correlation between goals scored and goals allowed for the Anaheim Ducks could not be rejected.

Comparative cognition: Eighteen children (age range 5 to 12) and eighteen university undergraduates (age range 18-35) were recruited by Westphal-Fitch et al (2012) to take part in a "spot the flaw" experiment. Images of Spanish, Cuban and Portuguese tiles with both rotational and translational patterns were shown to the participants. For each image a flawed version was also shown to participants. The authors found children's performance in detecting the flawed tiles was positively correlated with age for the rotational patterns(tau correlation = 0.358 , p-value = 0.026). There was no relationship for the translational patterns in children (tau correlation = 0.2, p-value = 0.124). No age / performance correlation was found in the adults.

How to calculate in R

The function cor.test{stats} can be used to perform this test.

Example: Using cor.test

Enter the following data

```
> x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1)

> y <- c( 2.6,  3.1,  2.5,  5.0,  3.6,  4.0,  5.2,  2.8,  3.8)
```

The function cor.test performs the calculation. The test statistic can be calculated for the above data by typing:

```
> cor.test(x,y,method="kendal",alternative="two.sided")

        Kendall's rank correlation tau

data:  x and y

T = 26, p-value = 0.1194

alternative hypothesis: true tau is not equal to 0

sample estimates:

      tau

0.4444444
```

Since the p-value = 0.1194 and is greater than 0.05, do not reject the null hypothesis. To specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater").

References

Dayaratna, Kevin D.; Miller, Steven J. (2012). The Pythagorean Won-Loss Formula and Hockey: A Statistical Justification for Using the Classic Baseball Formula as an Evaluative Tool in Hockey. Hockey Research Journal: A Publication of the Society for International Hockey Research. Fall.

Kyle, R.G, Campbell, M, Powell, P, Callery, P. (2012). Relationships between deprivation and duration of children's emergency admissions for breathing difficulty, feverish illness and diarrhea in North West England: an analysis of hospital episode statistics. BMC Pediatrics. 12:22.

Westphal-Fitch , Gesche; Huber, Ludwig; Gómez, Juan Carlos ;a Fitch,W.T. (2012). Production and perception rules underlying visual patterns: effects of symmetry and hierarchy .Philos Trans R Soc Lond B Biol Sci. 2012 July 19

367(1598): 2007–2022.

Back to Table of Contents

TEST 4 Z TEST OF THE DIFFERENCE BETWEEN INDEPENDENT CORRELATIONS

Question the test addresses

Is the difference between two independent correlation coefficients significantly different from zero?

When to use the test?

To assess the null hypothesis of zero correlation between two or more sample correlation coefficients calculated from independent samples. The data are assumed to be bivariate normal. The samples may be different sizes.

Practical Applications

Obesity: Allison et al (1996) estimate heritability for body mass index of 53 pairs of monozygotic twins reared apart. They studied three cohorts – Finish (17 pairs), Japanese (10 pairs) and archival case histories (26 pairs). Heritability was measured by the correlation of body mass between pairs. For the Finish, Japanese and archival samples the correlation was estimated to be 0.54,0.77 and 0.89 respectively. Differences between the correlations were tested using the Z test of the difference between independent correlations. There was no difference between the Finish and Japanese correlations (p-value = 0.368) or the Japanese and archival correlations (p-value = 0.328). A significant correlation was found between the Finish and archival correlations (p-value = 0.015).

Cross-cultural psychology: The meaning of “being Chinese” and “being American” were compared among 119 immigrant Chinese who arrived in the United States before or at age 12, and 112 immigrant Chinese who arrived in the United States after age 12 by Tsai et al (2000). For immigrant Chinese who arrived in the United States before or at age 12 the correlation between “being Chinese” and “being American” was -0.33 (p-value <0.001). For immigrant Chinese who arrived in the United States after age 12 the correlation between “being Chinese” and “being American” was -0.26 (p-value <0.01). The authors test the equality of these two correlation coefficients. They did not find a significant difference between the two immigrant groups in the magnitude of the correlation coefficients (p-value > 0.05). The null hypothesis of equality of correlation coefficients between the two groups could not be rejected.

Financial Fraud: Elrod and Gorhum (2012) investigate financial statement

fraud using correlation analysis. For a group of 594 companies who had engaged in financial reporting fraud they calculated correlations between revenue, cash flows from operations, assets and income from continued operations. For a control group of 420 firms from the New York Stock Exchange, they calculated similar correlations. For the fraud group the correlation between cash flow from operations and income from continuing operations was 0.46. For the control group it was 0.97. The Z test of the difference between independent correlations was significant (p-value <0.0001) and the null hypothesis of equality of correlations was rejected. For the fraud group the correlation between revenue and assets was 0.69. For the control group it was 0.77. The Z test of the difference between independent correlations was significant (p-value = 0.007) and the null hypothesis of equality of correlations was rejected.

How to calculate in R

The function paired.r{psych} can be used to perform this test. Pass the function estimates of the correlations and the sample sizes using the following format paired.r(first.correlation, second.correlation, NULL, first.sample.size, second.sample.size, twotailed=TRUE)

Example monozygotic twins: Allison et al (1996) estimate heritability for body mass index of 53 pairs of monozygotic twins reared apart. They find significant correlation between the Finish and archival correlations (p-value = 0.015). Let's reproduce their results. The correlation between the Finish twins was 0.54 and 0.89 for the archival twins. The sample sizes were 17 pairs of Finish twins and 26 pairs for the archival twins. Load the package psych and enter the following:

```
> paired.r(0.54,0.89,NULL,17,26,twotailed=TRUE)

$test

[1] "test of difference between two independent correlations"

$z

[1] 2.41245

$p

[1] 0.01584569
```

The p-value is 0.015 and less than the critical value of 0.05, reject the null hypothesis.

Example: financial statement fraud

Elrod and Gorhum (2012) investigate financial statement fraud using correlation analysis. For a group of 594 companies who had engaged in financial reporting fraud they calculated correlations between revenue, cash flows from operations, assets and income from continued operations. For a control group of 420 firms from the New York Stock Exchange, they also calculated these correlations. For the fraud group the correlation between revenue and assets was 0.69. For the control group it was 0.77. The Z test of the difference between independent correlations was significant (p-value = 0.007) and the null hypothesis of equality of correlations was rejected. These results can be reproduced by entering the following:

```
> paired.r(0.69,0.77,NULL,594,420,twotailed=TRUE)

$test

[1] "test of difference between two independent correlations"

$z

[1] 2.695245

$p

[1] 0.007033692
```

The p-value is 0.007 and less than the critical value of 0.05, reject the null hypothesis. Note set twotailed=FALSE to perform a one tailed test.

References

Allison, DB; Kaprio,J; Korkeila,M; Neale, MC; Hayakawa,K. (1996). Th heritability of body mass index among an international sample of monozygotic twins read apart. International Journal of Obesity, 20, pages 501- 506.

Elrod, Henry; Gorhum, Megan Jacqueline. (2012). Fraudulent financia reporting and cash flows. Journal of Finance and Accountancy, vol 11.

Tsai, Jeanne L; Ying, Yu-Wen ; Lee, Peter A. (2012).Journal Of Cross-Cultura Psychology, Vol. 31 No. 3, May, pages: 302-332.

Back to Table of Contents

TEST 5 DIFFERENCE BETWEEN TWO OVERLAPPING CORRELATION COEFFICIENTS

Question the test addresses

Is the difference between two dependent correlations sharing a common variable significantly different from zero?

When to use the test?

To assess the null hypothesis of zero correlation between one pair of variables to that of a second, overlapping pair of variables. For example, to answer the question "is the correlation of age stronger on neuroticism than on anxiety?" If you have data on the same set of subjects for all three variables, you would use this test to compare the correlation between age and neuroticism with the correlation between age and anxiety. Notice the variable age is common to both correlations. The test was originally motivated by the selection of the better of two available predictors (x and y) for a dependent variable z. The objective was to compare the correlation between z and x with that of z and y. Since y and z are not independent, the test statistic is required to take correlation (y,z) into account. The data are assumed to be normally distributed. The test is sometimes referred to as Steiger's t-test, Meng's t-test, Meng, Rosenthall & Rubin's t-test or Williams test.

Practical Applications

Neuroradiology: Liptak et al (2008) report the Pearson correlation between upper cervical cord volume and medulla oblongata volume (MOV) from brain imaging of 45 patients with multiple sclerosis as 0.67; and the correlation between MOV and brain parenchymal fraction as 0.45. The test for difference between two overlapping dependent correlations (which they call Meng's test) was not significantly different from zero (p-value = 0.086). The null hypothesis of equality of overlapping correlation coefficients could not be rejected.

Cardiovascular health: Olkin and Fin (1990) study the correlations among measures related to cardiovascular health of 66 mothers. The objective was to determine which of a number of cardiac measures (heart rate, blood pressure, systolic blood pressure (SP) or diastolic blood pressure) is the best indicator of body mass index (BMI). The correlation (BMI,SP) = 0.396 correlation (BMI, heart rate) = 0.179. Testing the difference between these two correlations involves comparing two overlapping dependent correlations (as BMI is a common variable). The test for the difference

between two overlapping dependent correlations was not significantly different from zero (p-value = 0.291). The null hypothesis of equality of overlapping correlation coefficients could not be rejected.

Neuropsychology: Crawford (2000) et al explore whether aging is associated with a differential deficit in executive function, compared with deficits in general cognitive ability. The 123 participants aged between 18 and 75 were given a range of general cognitive ability, executive function, and memory tests. Scaled scores for all subtests were summed to produce a Full Scale measure. Executive function tests included the Modified Card Sorting Test, Controlled Oral Word Association Test, and the Stroop Test. The correlation between the Stroop Test and age was -0.2, the correlation between the Full Scale measure and age was -0.28. The test for difference between these two overlapping dependent correlations (common variable is age) was not significantly different from zero (p-value = 0.44). The null hypothesis of equality of dependent correlation coefficients could not be rejected.

How to calculate in R

The functions compOverlapCorr{compOverlapCorr}, paired.r{psych} or r.test{psych} can be used to perform this test.

Example: Stroop Test and age.

Crawford (2000) et al report the correlation between the Stroop Test and age as -0.2, the correlation between the Full Scale measure and age as -0.28. The common variable is age. The correlation between the Stroop Test and the Full Scale measure was 0.3. The study had 123 participants aged between 18 and 75. Load the package compOverlapCorr and enter the following:

```
> compOverlapCorr(123, r13=-0.2, r23=-0.28, r12=0.30)

[1] 0.7713865 0.4404779
```

The first number is the value of the t-test statistic (0.77), the second is the p-value (0.44). The p-value is greater than the critical value of 0.05, do not reject the null hypothesis.

Example: using paired.r:

Continuing with the above example, load the package psych and enter the following:

```
> paired.r(xy=-0.2,xz=-0.28,yz=0.30, 123,twotailed=TRUE)
```

$test

[1] "test of difference between two correlated correlations"

$t

[1] 0.7732426

$p

[1] 0.4408868

The first number is the value of the t-test statistic (0.77), the second is the p-value (0.44). The p-value is greater than the critical value of 0.05, do not reject the null hypothesis. Note, set twotailed=FALSE to perform a one tailed test.

Example: using r.test:

Continuing with the above example, load the package psych and enter the following:

> r.test(123,r12=-.2,r34=-.28,r23=.3,twotailed = TRUE)

Correlation tests

Call:r.test(n = 123, r12 = -0.2, r34 = -0.28, r23 = 0.3, twotailed = TRUE)

Test of difference between two correlated correlations

 t value 0.77   with probability < 0.44

The first number is the value of the t-test statistic (0.77), the second is the p-value (0.44). The p-value is greater than the critical value of 0.05, do not reject the null hypothesis. Note, set twotailed=FALSE to perform a one tailed test.

References

Liptak Z. ; Berger A. M. ; Sampat M. P. ; et al. Medulla oblongata volume: A biomarker of spinal cord damage and disability in multiple sclerosis. American journal of neuroradiology Volume: 29 Issue: 8 Pages: 1465-1470

Olkin, I., Finn, J.D., 1990. Testing correlated correlations. Psych. Bull. 108, 330–333.

Crawford, J. R., Bryan, J., Luszcz, M. A., Obonsawin, M. C., & Stewart, (2000). The executive decline hypothesis of cognitive aging: Do executive

deficits qualify as differential deficits and do they mediate age-related memory decline? Aging, Neuropsychology, and Cognition, 7, 9–31.

Back to Table of Contents

TEST 6 DIFFERENCE BETWEEN TWO NON-OVERLAPPING DEPENDENT CORRELATION COEFFICIENTS

Question the test addresses

Is the difference between two non-overlapping correlation coefficients significantly different from zero?

When to use the test?

You have data on the same set of subjects for four variables and want to compare the null hypothesis of zero correlation between one pair of variables and a second non-overlapping pair of variables. This test is frequently used to compare the difference in correlation between two variables at two different points in time. The data are assumed to be normally distributed.

Practical Applications

Family psychology: One-hundred and eighty-seven married couples' dieting behaviors, marital quality, body mass index, weight concerns, depression, and self-esteem were assessed in a study by Markey, Markey, and Birch (2008). The authors report the correlation between body mass index and the wife's healthy dieting behavior as 0.26; and the correlation between body mass index and the husband's healthy dieting behavior as 0.15. Both correlations are significantly different from zero (p-value < 0.05). They ask whether the difference between these two correlations is significantly different from zero. Since participants in this study were married, the correlations of husbands and wives are related, but non-overlapping. The test for difference between two non-overlapping dependent correlation coefficients was used. The authors report a p-value greater than 0.05; the null hypothesis of no difference between the two correlations cannot be rejected.

Psychological Trauma: Dekel, Solomon and Ein-Dor (2012), in a longitudinal study, examine the relationship between posttraumatic growth (PTG) and posttraumatic stress disorder (PTSD) for a sample of Israeli ex-prisoners of war. The participants were followed over 17 years with assessments at three time periods 1991, 2003 and 2008. The correlations between PTSD and PTG for the years 2003 and 2008 were calculated and used to test the difference between the correlations PTG with PTSD. The test for difference between two non-overlapping dependent correlation coefficients was not

significantly different from zero (p-value = 0.19). The null hypothesis of no difference between the two correlations cannot be rejected.

Verbal achievement: Steiger (1980) reports 103 observations on a hypothetical longitudinal study of sex stereotypes and verbal achievement. The three variables of masculinity, femininity and verbal ability are measured at two different time points. The question is whether the correlation between femininity and verbal achievement was the same at both time points. The Pearson correlations between femininity and verbal achievement for the two time periods were calculated, and then used to test the difference between the correlations of femininity with verbal achievement. The authors report a test statistic of 1.4 (p-value = 0.16), the null hypothesis cannot be rejected.

How to calculate in R

The function r.test{psych} can be used to perform this test.

Example: using r.test in family psychology:

Markey, Markey, and Birch (2008) in a study of 187 participants report the correlation between understanding from spouse and the wife's healthy dieting behavior as -0.11; and the correlation between understanding from spouse and the husband's healthy dieting behavior as 0.06. The correlation between the wife and husbands understanding from spouse score is 0.41. To test for difference between the correlation coefficients (wife = -0.11, and husband = 0.06) load the package psych and enter the following:

```
> r.test(187, r12 = -0.11, r34 = 0.06, r23 = 0.41)

Correlation tests

Call:r.test(n = 187, r12 = -0.11, r34 = 0.06, r23 = 0.41)

Test of difference between two correlated  correlations

 t value -2.15    with probability < 0.033
```

The first number is the value of the value of the test statistic (-2.15), the second is the p-value (p< 0.033). The p-value is less than the critical value of 0.05, reject the null hypothesis.

Example: using r.test to assess verbal achievement:

Steiger (1980) reports 103 observations on a hypothetical longitudinal study of sex stereotypes and verbal achievement. The question is whether the correlation between femininity and verbal achievement was the same

at both time points. The Pearson correlations between femininity (F) and verbal achievement (V) for the two time periods were calculated, and then used to test the difference between the correlations of femininity with verbal achievement. Steiger report the correlations as follows:

Correlation (F at time 1,V at time 1) = 0.5. We refer to this as r12 in the code below.

Correlation (F at time 1 ,F at time 1) = 0.7. We refer to this as r13 in the code below.

Correlation (V at time 2 ,F at time 1) = 0.5. We refer to this as r14 in the code below.

Correlation (F at time 2 ,V at time 1) = 0.5. We refer to this as r23 in the code below.

Correlation (V at time 2 ,V at time 1) = 0.8. We refer to this as r24 in the code below.

Correlation (V at time 2 ,F at time 2) = 0.6. We refer to this as r34 in the code below.

The authors report a test statistic of 1.4 (p-value = 0.16). To replicate their results load the package psych and enter the following:

```
> r.test(n=103,r12=0.5,r34=0.6,r13=0.7,r23=0.5,r14=0.5,r24=0.8)

Correlation tests

Call:r.test(n = 103, r12 = 0.5, r34 = 0.6, r23 = 0.5, r13 = 0.7, r14 = 0.5, r24 =
0.8)

Test of difference between two dependent correlations

 z value -1.4     with probability  0.16
```

The first number is the value of the test statistic (-1.4), the second is the p-value (0.16). The p-value is greater than the critical value of 0.05, do not reject the null hypothesis.

References

Dekel S, Ein-Dor T, Solomon Z. (2012). Posttraumatic growth and posttraumatic distress: a longitudinal study. Psychol Trauma: Theory, Res, Prac and Pol, 4:94–101.

Markey CN, Markey PM, Birch LL. Interpersonal predictors of dietin

practices among married couples. Journal of Family Psychology. 2001;15:464–475.

Steiger, J.H. (1980), Tests for comparing elements of a correlation matrix, Psychological Bulletin,87, 245-251.

Back to Table of Contents

TEST 7 BARTLETT'S TEST OF SPHERICITY

Question the test addresses

Is the correlation matrix an identity matrix?

When to use the test?

The test is used to assess whether a correlation matrix is an identity matrix (all diagonal terms are one and all off-diagonal terms are zero). It is often used in factor analysis studies where rejection of the null hypothesis of identity is an indication that the data are suitable for the Factor Analysis model.

Practical Applications

Electromyographic walking speeds: Ivanenko et al (2004) apply factor analysis to the set of electromyographic records obtained at different walking speeds and gravitational loads from 18 subjects. Participants were asked to walk on a treadmill at speeds of 1, 2, 3 and 5kmh as well as when 35–95% of the body weight was supported using a harness. Between12–16 ipsilateral leg and trunk muscles using both surface and intramuscular recording were taken. Bartlett's test of sphericity was applied to the correlation matrix of the 4 different speeds across 6 subjects and the overall average across subjects (p-value <0.001).

Brazilian general health questionnaire: Carvalho et al (2011) investigate the structural coherency of the 60-item version of the Brazilian general health questionnaire using factor analyses. A random sample of 146 individuals were recruited onto the study. To evaluate the suitability of the dataset for factor analysis, the researchers applied Bartlett's test of sphericity (p-value < 0.001 in all cases). The researchers conclude their dataset is suitable for exploratory and confirmatory factor analyses.

Self-regulatory skills in Greek school children: Konstantinopoulou and Metallidou (2012) examine the psychometric properties of a behavioral computerized test for measuring self-regulatory skills, in a sample of Greek primary school children. A total of 88 fourth grade girls (44) and boys (44) participated in the study. As part of the assessment, a child' number of key presses for each distracter condition (visual, audiovisual, and forced) was subtracted from the baseline number of key presses. This number was divided by the baseline number of key presses and then multiplied by 100. Bartlett's Test of Sphericity was applied on the correlation matrix (p-value <0.001).

How to calculate in R

The function cortest.bartlett{psych} can be used to perform this test. It takes the form cortest.bartlett(r, n = 100). Where r is a correlation matrix and n the sample size (default value is 100).

Example: using correlated data

We begin by generating correlated data, to do so enter the following

```
set.seed(1234)

n=1000

y1 <- rnorm(n)

y2<- rnorm(n)

y3<-y1+y2

data<-matrix(c(y1,y2,y3) , nrow = n, ncol=3, byrow=TRUE,)

correlation.matrix <- cor(data)
```

The above code creates a correlation matrix from 3 random variables, each containing 1000 observations. The variable y3 is dependent on the random variable y1 and y2. We can perform Bartlett's test of sphericity by entering

```
> cortest.bartlett(correlation.matrix,n)

$chisq

[1] 9.379433


$p.value

[1] 0.02464919


$df

[1] 3
```

The second number the p-value (p-value = 0.0246). Since it is less than the critical value of 0.05, reject the null hypothesis.

References

Carvalho, H. W. D., Patrick, C. J., Jorge, M. R., & Andreoli, S. B. (2011) Validation of the structural coherency of the General Health Questionnaire. Revista Brasileira de Psiquiatria, 33(1), 59-63.

Ivanenko, Y. P., Poppele, R. E., & Lacquaniti, F. (2004). Five basic muscle activation patterns account for muscle activity during human locomotion. The Journal of physiology, 556(1), 267-282.

Konstantinopoulou, E., & Metallidou, P. (2012). Psychometric properties of the self-regulation and concentration test for children (srtc) in a Greek sample of fourth grade students. Hellenic Journal of Psychology, 9, 158-178.

Back to Table of Contents

TEST 8 JENNRICH TEST OF THE EQUALITY OF TWO MATRICES

Question the test addresses

Are a pair of correlation matrices equal?

When to use the test?

To test for equality between two correlation matrices computed over independent subsamples. The test specifies the correlation coefficients as piecewise constant over time and then verifies whether the constants coincide. The underlying observations are assumed to be independent and normally distributed.

Practical Applications

Correlation between stock returns: Chesnay and Jondeau study correlations between international equity markets using a multivariate Markov-switching framework. The model allows the correlation matrix to vary across regime. Using weekly returns for the S&P, DAX and FTSE ove the period 1988 to 1999, the researchers investigate whether or not correlations are regime independent. The Jennrich test is used to assess this. The correlation matrix for the period 1988-1991 is compared to the correlation matrix for the period 1992-1995 (p-value = 0.2608); for this period the null hypothesis cannot be rejected. The correlation matrix for the period 1992-1995 is compared to the correlation matrix for the period 1995-1999 (p-value = 0.0.001); for this period the null hypothesis of equality of correlation matrices was rejected.

Real Estate Investment Trusts: Kim et al (2007) investigates structural breaks in the returns on Real Estate Investment Trusts (REIT). Usin monthly data measured over the period 1971–2004. A vector auto-regression model is constructed using time-series data on REIT returns, stock market returns and other macroeconomic variables. The variables considered in the analysis are REITs (total) return rates, SP S&P 500 retur rates, industrial production growth rates, US consumer price index, term spread (the difference between the 10 year US Treasury bond rate and the one month US Treasury bill rates), credit spread (the difference between Baa and Aaa corporate bond rates) and the first difference of one month US Treasury bill rates. The correlation matrices of model innovations for the periods November 1980 to November 2004 and December 1971 to

October 1980 are compared using the Jennrich test (p-value <0.01). The researchers conclude the matrices of innovations between the two periods are not homogeneous.

Analysis of 120 years of industrial production: Annual data on real Gross Domestic Product for sixteen industrial countries over 120 years was studied by Bordo and Helbling (2003). The focus is on four distinct eras with different international monetary regimes. The four eras covered are 1880-1913 when much of the world adhered to the classical Gold Standard, the interwar period (1920-1938), the Bretton Woods regime of fixed but adjustable exchange rates (1948-1972), and the modern period of managed floating among the major currency areas (1973 to 2001). The Jennison test was used to compare the correlation matrices of different time periods within and between these regimes. The following correlation matrices were tested: 1880-1913 versus 1926-38 (p-value = 0.86), 1926-38 versus 1952-72 (p-value = 0.01), 1952-72 versus 1973-2001 (p-value = 0.01), 1880-1913 versus 1952-72 (p-value = 0.28), 1880-1913 versus 1973-2001 (p-value =<0.01), 1226-38 versus 1973-2001 (p-value <0.01).

How to calculate in R

The function cortest.jennrich{psych} can be used to perform this test. It can be used in the form cortest.jennrich(sample.1, sample.2). Where sample.1 and sample.2 are the sample observations from which the first and second correlation matrix will be constructed.

Example:

We begin by generating two correlation matrices with a similar structure. To do this we will use the standard normal distribution:

```
set.seed(1234)

n1 =1000

n2=500

sample.1 <- matrix(rnorm(n1),ncol=10)

sample.2 <- matrix(rnorm(n2),ncol=10)
```

The test can be applied as follows:

```
> cortest.jennrich(sample.1, sample.2)

$chi2
```

[1] 54.16223

$prob

[1] 0.1644589

The second number the p-value (p-value = 0.164). Since it is greater than the critical value of 0.05, we cannot reject the null hypothesis.

Example:

Let's take a look at the case where we have different correlation structures. To do this generate a standard normal sample and a sample from the lognormal distribution as follows:

```
set.seed(1234)

n1 =1000

n2=500

sample.1 <- matrix(rnorm(n1),ncol=10)

sample.2 <- matrix(rlnorm(n2, meanlog = 0, sdlog = 10),ncol=10)
```

The test can be applied as follows:

```
> cortest.jennrich(sample.1, sample.2)

$chi2

[1] 95.69021

$prob

[1] 1.613946e-05
```

Since the p-value is less than 0.05, we reject the null hypothesis of equality of correlation matrices.

References

Bordo, M. D., & Helbling, T. (2003). Have national business cycles become more synchronized? (No. w10130). National Bureau of Economic Research.

Chesnay, F., & Jondeau, E. (2001). Does correlation between stock returns really increase during turbulent periods?. Economic Notes, 30(1), 53-80.

Kim, J. W., Leatham, D. J., & Bessler, D. A. (2007). REITs' dynamics under

structural change with unknown break points. Journal of Housing Economics, 16(1), 37-58.

Back to Table of Contents

TEST 9 GRANGER CAUSALITY TEST

Question the test addresses

Is one time series useful in forecasting another?

When to use the test?

Granger causality is a statistical concept of causality. Whenever a "surprise" in an independent variable leads to a later increase in the dependent variable we call this variable "Granger causal." The test is based on prediction, if an independent variable "Granger-causes" a dependent variable, then past values of the independent variable should contain information that helps predict the dependent variable above and beyond the information contained in past values of the dependent variable alone. For example, a time series X is said to Granger-cause Y if it can be shown—usually through a series of t-tests and F-tests on lagged values of X (and with lagged values of Y also included)—that those X values provide statistically significant information about future values of Y.

Practical Applications

Posterior-anterior connectivity in the brain: Granger causality tests have been used to identify bi-directional, posterior-anterior connectivity in the brain. Using magneto-encephalography and Granger causality analysis, Lou et al (2011) tested in a paralimbic network the hypothesis that stimulation may enhance causal recurrent interaction between higher-order, modality non-specific regions. The network includes anterior cingulate/medial prefrontal and posterior cingulate/medial parietal cortices together with pulvinar thalami, a network known to be effective in autobiographic memory retrieval and self-awareness. The test variables were computed on single trials in 100 ms time windows that covered the time range between −1 s and +1 s with respect to stimulus onset. Time series were computed for each single trial, for each region of interest and for each individual participant. The researchers observed Granger causality (p-value <0.05) was determined during 1 s in the prestimulus condition and during 1 s in the stimulus condition. It was also observed that Granger causality is bi-directional and approximately symmetrical between regions in almost all 100 ms epochs with few exceptions, independent of frequency band, and in both conditions.

Herbicide-tolerant soybeans and tillage practices: Fernandez-Cornejo et al (2013) examine the extent to which adopting herbicide-tolerant (HT) soybeans affects conservation tillage practices and herbicide use. The

model is estimated using a state-level panel dataset extending across 12 major soybean-producing states in the US from 1996 to 2006. The researchers investigate the granger causality between HT soybean adoption rates and conservation tillage adoption rates. They find state-level HT soybean adoption rates Granger-cause conservation tillage adoption rates at the 5% level (p-value = 0.014), but conservation tillage rates do not Granger-cause HT soybean adoption rates (p-value = 0.17).

Metabolic reaction network: Stern and Enflo (2013) develops a new approach to identify and predict a probable metabolic reaction network from time-series data of metabolite concentrations. Their analysis starts with smoothing noisy time-series data using locally estimated scatter plot smoothing. Then, bivariate Granger causality is calculated to examine causal relationships between all pairs of metabolites, with unrelated metabolite pairs removed from further consideration. The researchers observed that each metabolite is Granger-caused by other metabolites (p-value <0.01 in all cases).

How to calculate in R

The function granger.test{MSBVAR} can be used to perform this test. It takes the form granger.test(data, p=1), where data are the time-series data and p is the order of the test.

Example:

Enter the following data

sample1=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,3794,3959,

sample2=c(75,78,80,82,84,88,93,97,99,104,109,115,120,127)

sample3=c(5,8,0,2,4,8,3,7,9,10,10,15,12,12)

data=cbind(sample1,sample2,sample3)

To apply the Granger causality test of order one to all combinations enter:

> granger.test(data, p=1)

	F-statistic	p-value
sample2 -> sample1	15.5918633	0.002736246
sample3 -> sample1	0.4599351	0.513040905
sample1 -> sample2	0.9111302	0.362318992

sample3 -> sample2 0.2544981 0.624854894

sample1 -> sample3 7.0543816 0.024063590

sample2 -> sample3 7.5968675 0.020260056

The function returns the F-statistic and p-value for all six possible combinations of Granger causality. The results indicate sample2 Granger causes sample1 (p-value =0.0027), sample1 Granger causes sample3 (p-value =0.024) and sample2 Granger causes sample3 (p-value =0.0202).

To carry out the test for Granger causality test of order two enter

> granger.test(data, p=2)

	F-statistic	p-value
sample2 -> sample1	2.7304983	0.132864335
sample3 -> sample1	1.6266345	0.262920776
sample1 -> sample2	0.6167666	0.566619190
sample3 -> sample2	0.3415193	0.721900939
sample1 -> sample3	9.8334762	0.009266939
sample2 -> sample3	10.4923267	0.007827505

It appears sample1 Granger causes sample3 (p-value =0.009) and sample2 Granger causes sample3 (p-value =0.0078).

References

Fernandez-Cornejo, J., Hallahan, C., Nehring, R., Wechsler, S., & Grube, A (2013). Conservation Tillage, Herbicide Use, and Genetically Engineered Crops in the United States: The Case of Soybeans.

Lou, H. C., Joensson, M., Biermann-Ruben, K., Schnitzler, A., Østergaard, L Kjaer, T. W., & Gross, J. (2011). Recurrent activity in higher order, modality non-specific brain regions: a Granger causality analysis of autobiographic memory retrieval. PloS one, 6(7), e22286.

Stern, D. I., & Enflo, K. (2013). Causality Between Energy and Output in the Long-Run (No. 126). Department of Economic History, Lund University.

Back to Table of Contents

TEST 10 DURBIN-WATSON AUTOCORRELATION TEST

Question the test addresses

Is there serial correlation in the sample?

When to use the test?

To investigate whether the residuals from a linear or multiple regression model are independent. It is assumed the residuals from the regression model are stationary and normally distributed with zero mean. It tests the null hypothesis that the errors are uncorrelated against the alternative that they are autoregressive. The test is not valid if there are lagged values of the dependent variable on the right hand side of the equation (in this case use Breusch-Godfrey test).

Practical Applications

Corrected Anion Gap as Surrogates: Mallat et al (2013) investigate whether the difference between sodium and chloride and anion gap corrected for albumin and lactate (AGcorr) could be used as a strong ion gap (SIG surrogate in critically ill patients. A total of 341 patients were prospectively recruited on to the study; 161 were allocated to the modeling group, and 180 to the validation group. A linear regression model was constructed between SIG and AGcorr. The assumption of independent residuals was tested with the Durbin-Watson test. The Durbin-Watson test of the model was 1.9, which the researchers report as indicative of independence of residuals (p-value > 0.05).

Decline in new drug launches: Ward et al (2013) carry out a retrospective observational study of new drug launches in the UK. A linear regression model of new drugs introduced (dependent variable) on time (independent variable) is built using data on new drugs launched from 1971 to 2011. There was a significant positive first-order autocorrelation in the residuals (Durbin-Watson statistic = 1.10, p-value <0.01).

Predicting gas flaring: Gas flaring is the burning of excessive gas associated with crude oil production. John and Friday (2012) build a linear regression model to predict gas flaring in the Niger Delta, Nigeria. Data on gas production, oil production and flaring over the period 1980 to 2000 were analyzed in the study. The basic model related flaring (dependent variable) to gas and oil production (independent variables). The Durbin-Watson test was significant (p-value <0.05) and the researchers reject to null hypothesis of independent residuals.

How to calculate in R

The function durbinWatsonTest{car} can be used to perform this test. It takes the form durbinWatsonTest (residual,lag). The parameter lag refers to the number of autocorrelations you wish to test for.

Example:

Enter the following data, collected, on two variables:

dependent.variable=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,

independent.variable=c(75,78,80,82,84,88,93,97,99,104,109,115,120,127)

To carry out the test enter

durbinWatsonTest(lm(dependent.variable ~ independent.variable))

lag	Autocorrelation	D-W Statistic	p-value
1	-0.5487192	3.031143	0.062

Alternative hypothesis: rho != 0

Since the p-value is greater than 0.05, do not reject the null hypothesis. If you want to test for 3 lags you can enter:

> durbinWatsonTest(lm(dependent.variable ~ independent.variable),3)

lag	Autocorrelation	D-W Statistic	p-value
1	-0.5487192	3.031143	0.064
2	0.3218803	1.116702	0.072
3	-0.1545470	1.782968	0.934

Alternative hypothesis: rho[lag] != 0

Since the p-value at each lag is greater than 0.05, the null hypothesis cannot be rejected.

References

John, O., & Friday, U. E. (2012). Model for predicting gas flaring in Niger delta. Continental Journal of Engineering Sciences, 6(2).

Mallat, J., Barrailler, S., Lemyze, M., Pepy, F., Gasan, G., Tronchon, L., & Thevenin, D. (2013). Use of Sodium-Chloride Difference and Corrected Anion Gap as Surrogates of Stewart Variables in Critically Ill Patients. PloS

one, 8(2), e56635.

Ward, D. J., Martino, O. I., Simpson, S., & Stevens, A. J. (2013). Decline i new drug launches: myth or reality? Retrospective observational study using 30 years of data from the UK. BMJ open, 3(2).

Back to Table of Contents

TEST 11 BREUSCH–GODFREY AUTOCORRELATION TEST

Question the test addresses

Is there serial correlation in the sample?

When to use the test?

To investigate whether the residuals from a linear or multiple regression model are independent. It is assumed the residuals from the regression model are stationary and normally distributed with zero mean. It tests the null hypothesis that the errors are uncorrelated against the alternative that they are autoregressive.

Practical Applications

Enhanced nutrient concentrations in streams: Artigas et al (2013) analyzed the biological responses of stream ecosystems to experimental nutrient enrichment in three bioclimatic regions (Mediterranean, Pampean and Andean). In each stream, the researchers enhanced nutrient concentrations 2–4 fold over a 50 meter length. An upstream reach of similar morphological and hydrological characteristics was kept as the control. The experiment followed a BACIPS (before–after, control–impact paired series) design. An important assumption of BACIPS designs is the lack of seria correlation. This was tested using the Breusch–Godfrey test. The researchers report one of the 23 analyses (combinations of variables and streams) showed significant serial correlation (p-value < 0:05). One case of the 23 raw variables showed serial correlation (p-value < 0:05) in the enriched zones.

Sales and marketing communication: Nogueira (2013) investigate the relationship between sales and marketing investments of a shopping mall in Brazil. A regression model with monthly sales as the dependent variable and marketing communication investments as the independent variable is constructed. The period of the study was January 2001 to December 2008, a total of 96 monthly observations. The linear regression error term was modeled using an auto-regressive of order 1 term. The Breusch-Godfrey test was then applied to the full model (p-value = 0.428). The null hypothesis that the residuals are serially uncorrelated could not be rejected (up to 12 lags).

US total Oil Consumption: Huntington (2010) explores US total oi consumption over the period 1950-2005. A regression model was constructed with per-capita total oil consumption as the dependent variable. The independent variables were the change in Gross Domestic

Product, change in price, and oil demand in the previous period. The residual term from the model was assessed using the Breusch-Godfrey test (p-value <0.05).

How to calculate in R

The function bgtest{lmtest} can be used to perform this test. It takes the basic form bgtest(model, order = 1). The parameter order refers to the number of autocorrelations you wish to test for.

Example:

Enter the following data, collected, on two variables:

dependent.variable=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,

independent.variable=c(75,78,80,82,84,88,93,97,99,104,109,115,120,127)

To carry out a first order test enter

> bgtest(lm(dependent.variable ~ independent.variable),order=1)

Breusch-Godfrey test for serial correlation of order up to 1

data: lm(dependent.variable ~ independent.variable)

LM test = 4.2591, df = 1, p-value = 0.03904

Since the p-value is less than 0.05, reject the null hypothesis. If you want to test forth order serial correlation you can enter:

> bgtest(lm(dependent.variable ~ independent.variable),order=4)

Breusch-Godfrey test for serial correlation of order up to 4

data: lm(dependent.variable ~ independent.variable)

LM test = 5.1459, df = 4, p-value = 0.2727

Since the p-value is greater than 0.05, the null hypothesis cannot be rejected.

References

Artigas, J., García-Berthou, E., Bauer, D. E., Castro, M. I., Cochero, J Colautti, D. C., ... & Sabater, S. (2013). Global pressures, specific responses: effects of nutrient enrichment in streams from different biomes. Environmental Research Letters, 2013, vol. 8, núm. 1, p. 014002.

Huntington, H. G. (2010). Short-and long-run adjustments in US petroleum

consumption. Energy Economics, 32(1), 63-72.

Nogueira, C. A. G., Pinheiro, M. F., Neto, A. R., & Gomes, D. M. D. O. A (2013). Do Marketing Communication Investments Always Pay Off?. Revista FSA (Faculdade Santo Agostinho), (9), 33-54.

Back to Table of Contents

TEST 12 ONE SAMPLE T-TEST FOR A HYPOTHESIZED MEAN

Question the test addresses

Is the mean of a sample significantly different from a hypothesized mean?

When to use the test?

You want to compare the sample mean to a hypothesized value. The test assumes the sample observations are normally distributed and the population standard deviation is unknown.

Practical Applications

Predator avoidance: Ings, Wang and Chittka (2012) study how bees' past experience of predator spiders influenced their responses to the presence of predator spiders. One sample t tests were used to determine if, by the end of predator avoidance training, the bees had learnt to avoid flowers harbouring predator spiders. The authors report visitation rates in trained bees of 0.03 visits per flower choice. This was significantly different from the value of 0.25 expected if the bees were choosing flowers at random (p-value <0.001). The null hypothesis that the mean number of visits of trained bees was equal to 0.25 (and hence random) was rejected.

Positive psychology: Proyer, Ruch and Buschor (2012), investigate the impact of character strengths-based positive intervention. An experimental group of 56 individuals underwent interventions on the strengths of curiosity, gratitude, hope, and zest. Participants were given a 240 item self-assessment questionnaire on 24 character strengths prior to the intervention. A self-evaluation sheet was completed by participants after the completion of the program. A score of 4.87 for curiosity post intervention was compared to the pre-intervention score of 4 using the one sample t-test. The p-value was less than 0.01, the authors reject the null hypothesis of no change in self- perception of curiosity post intervention.

Psychiatry: Ayoughi, Missmahl, Weierstall (2012) examine the benefits of psychosocial counseling. Thirty one Afghan women suffering from mental health problems received five counseling sessions of between 45 minutes to an hour. The mean pre-treatment depression score was 41.65. A one sample t-test was used to investigate if the post treatment depression score is significantly different from the pre-treatment score. The post counseling depression score was 20.26 with a p-value < 0.001, the null hypothesis of no difference is rejected. The authors also examined a

control group of thirty women who received medication but no counseling. The mean pre-treatment score for this group was 43. To investigate if the mean post treatment depression score is significantly different from the post treatment score, a one sample t-test was used. The null hypothesis was not rejected (p-value = 0.90).

How to calculate in R

The function t.test{stats} is used to perform this test. It takes the form:

t.test(x,mu=3, alternative = "two.sided", conf.level = 0.95), where mu is the hypothesized value to be tested (in this case the value 3), alternative is the type of test to be conducted and conf.level is the confidence level of the test, in this case 95%. Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Example: two sided test using t.test:Enter the following data

>x <- c(59.3,14.2,32.9,69.1,23.1,79.3,51.9,39.2,41.8)

To test the null hypothesis that the sample mean equals 40, type

>t.test(x,mu=40, alternative = "two.sided", conf.level = 0.95)

One Sample t-test

data: x

t = 0.7956, df = 8, p-value = 0.4492

alternative hypothesis: true mean is not equal to 40

95 percent confidence interval:

29.28381 62.00508

sample estimates:

mean of x

45.64444

The p-value is equal to 0.4492 and greater than 0.05, do not reject the null hypothesis. The function also reports the 95% confidence interval as 29.2 to 62.0 As it crosses the test value of 40, do not reject the null hypothesis.

Example: one sided test using t.test

Enter the following data

```
x <- c(59.3,14.2,32.9,69.1,23.1,79.3,51.9,39.2,41.8)
```

To test the null hypothesis the sample mean is greater than 30, type

```
> t.test(x,mu=30, alternative = "greater", conf.level = 0.95)

        One Sample t-test

data:  x

t = 2.2051, df = 8, p-value = 0.02927

alternative hypothesis: true mean is greater than 30

95 percent confidence interval:

 32.45133      Inf

sample estimates:

mean of x

 45.64444
```

The p-value is equal to 0.029 and less than 0.05, reject the null hypothesis. The function also reports the 95% confidence interval as 32.45 to infinity. It lies above the hypotheisesd value of 30, reject the null hypothesis.

References

Ayoughi S, Missmahl I, Weierstall R, Elbert T (2012). Provision of menta health services in resource-poor settings: a randomised trial comparing counselling with routine medical treatment in North Afghanistan (Mazar-e-Sharif). BMC Psychiatry 12: 14.

Ings, T. C; Wang, M.Y; Chittka, L. (2102). Colour-independent shape recognition of cryptic predators by bumblebees. Behavioral Ecology and Sociobiology. Volume 66, Number 3 (2012), 487-496.

Proyer, R. T., Ruch, W., & Buschor, C. (2012). Testing strengths-based interventions: A preliminary study on the effectiveness of a program targeting curiosity, gratitude, hope, humor, and zest for enhancing life satisfaction. Journal of Happiness Studies.

Back to Table of Contents

TEST 13 ONE SAMPLE WILCOXON SIGNED RANK TEST

Question the test addresses

Is the median of a sample significantly different from a hypothesized value?

When to use the test?

To test whether the median of sample is equal to a specified value. The null hypothesis is that the median of observations is zero (or some other specified value). It is a nonparametric test and therefore requires no assumption for the sample distribution. It is an alternative to the one-sample t-test when the normal assumption is not satisfied.

Practical Applications

Dog- human communication: The visual communication between humans and dogs is analyzed by Lakatos, Gacsi, Topal and Miklosi (2012). The authors conduct three studies to assess the ability of dogs to comprehend a variety of human pointing gestures. Sixteen dogs were recruited into the study. One experiment investigated whether dogs chose the correct side (left or right) based on a momentary distal pointing gesture of a human. The null hypothesis that side selection by the dogs was random was rejected by the one sample Wilcoxon signed rank test (p-value <0.001).

Higher Education Studies: Ahmad, Farley and Naidoo (2012) administer a survey of 120 senior academic administrators at Malaysian public universities. Their objective is to assess how Malaysian higher education reforms have been perceived by this group. A questionnaire with a seven point scale was completed by the respondents. The authors use the one sided Wilcoxon signed rank test to assess the null hypothesis that the sample mean response to the question “Improved overall quality of teaching and learning”. The null hypothesis is that the sample mean is equal to 4 (neutral on their scale). The alternative hypothesis was that it is greater than 4. The sample mean response was 5.33 with p-value less than 0.001, the null hypothesis was rejected.

Brain oscillations: Continuous electroencephalogram oscillations in humans exhibit a power-law decay of temporal correlations in the fluctuations of oscillation amplitudes. Long range temporal correlations are often characterized using estimates of the Hurst exponent. Hartley et al (2012) analyzed the electroencephalogram of 11 preterm babies. The authors use the Wilcoxon signed rank test to assess whether the Hurst exponents of inter-event interval sequences for the sample are different from random. Using the one sample Wilcoxon signed rank test, the exponents were found

to be significantly different from random (p-value <0.001). The authors reject the null hypothesis and conclude the results indicate long range temporal correlations in the inter-event interval sequences of all subjects studied.

How to calculate in R

The function wilcox.test{stats} is used to perform this test. It takes the form:

wilcox.test(x,mu=3, alternative = "two.sided"), where mu is the hypothesized value to be tested (in this case the value 3), alternative is the type of test to be conducted. Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Example: two sided test using wilcox.test

Enter the following data

>x <- c(59.3,14.2,32.9,69.1,23.1,79.3,51.9,39.2,41.8)

To test the null hypothesis the sample mean = 40, type

> wilcox.test (x,mu=40, alternative = "two.sided")

Wilcoxon signed rank test

data: x

V = 29, p-value = 0.4961

alternative hypothesis: true location is not equal to 40

Since the p-value is greater than 0.05, do not reject the null hypothesis.

Example: one sided test using wilcox.test

Enter the following data

x <- c(59.3,14.2,32.9,69.1,23.1,79.3,51.9,39.2,41.8)

To test the null hypothesis the sample mean is greater than 30, type

> wilcox.test (x,mu=30, alternative = "greater")

Wilcoxon signed rank test

data: x

V = 38, p-value = 0.03711

alternative hypothesis: true location is greater than 30

Since the p-value is less than 0.05, reject the null hypothesis.

References

Ahmad, A. R., Farley, A., & Naidoo, M. (2012). Impact of the governmen funding reforms on the teaching and learning of Malaysian public universities. Higher Education Studies, 2(2), p114.

Hartley C, Berthouze L, Mathieson SR, Boylan GB, Rennie JM, et al. (201 Long-Range Temporal Correlations in the EEG Bursts of Human Preterm Babies. PLoS ONE 7(2): e31543. doi:10.1371/journal.pone.0031543

Lakatos, G., Gacsi, M., Topal, J., & Miklosi, A . (2012) Comprehension and utilisation of pointing gestures and gazing in dog human communication in relatively complex situations. Animal Cognition, 15, 201-213.

Back to Table of Contents

TEST 14 SIGN TEST FOR A HYPOTHESIZED MEDIAN

Question the test addresses

Is the median of a sample significantly different from a hypothesized median?

When to use the test?

To test whether the median of sample is equal to a specified value. The null hypothesis is that the median of observations is zero (or some other specified value). It is a nonparametric test and therefore requires no assumption for the sample distribution. It is often used alongside the Wilcoxon signed rank test or as an alternative to the one-sample t-test when the normal assumption is not satisfied.

Practical Applications

Accounting: Productivity change, technical progress, and relative efficiency change in the United States public accounting industry is investigated by Banker, Chang and Natarajan. (2005). The authors collect revenue and human resources data for 64 large accountancy firms over the period 1995-1999. The sign test for a hypothesized median is used to assess whether the annual rate of change in the factors of productivity, technical efficiency and relative efficiency are greater than zero. For the year 1995-96 the estimated median for productivity change was 0.034, with a p-value =0.01. The null hypothesis of no change in productivity for the year 1995-1999 was rejected.

Lipid storage: Spalding et al (2008) study the role of increased lipid storage in already developed fat cells (adipocytes). The authors construct various death rate estimates of adipocytes and use the sign test to assess the reliability of their estimates against their sample. They find estimates of the death rate do not differ from the observed sample median (p-value > 0.3).

Horticultural Science: Davis (2009) reviews the evidence supporting declines over the past 100 years in the concentration of certain nutrients in vegetables and fruits available in the United Kingdom and United States The sign test is used to assess whether the median nutrient values are different than historical levels. Of 33 nutrients comparisons for various common fruits and vegetables, 11 showed statistically significant declines (p-value < 0.05).

How to calculate in R

The functions simple.median.test{UsingR} and SIGN.test{BSDA}can be use to perform this test.

Example: using simple.median.test

This function takes the form simple.median.test(x, median=3), where median is the hypothesized value to be tested (in this case the value 3).

Enter the following data

> x<-c(12,2,17,25,52,8,1,12)

To test the null hypothesis the sample median is equal to 20, type

> simple.median.test(x, median = 20)

[1] 0.2890625

The p-value = 0.289, do not reject the null hypothesis.

Example: using SIGN.test

This function takes the form

SIGN.test(x,md = 3, alternative = "two.sided", conf.level = 0.95), where md is the hypothesized value of the median to be tested (in this case the value 3) . Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Enter the following data

> x<-c(12,2,17,25,52,8,1,12)

To test the null hypothesis the sample median is equal to 20, type

> SIGN.test(x,md=20, alternative = "two.sided", conf.level = 0.95)

One-sample Sign-Test

data: x

s = 2, p-value = 0.2891

alternative hypothesis: true median is not equal to 20

95 percent confidence interval:

1.675 33.775

sample estimates:

median of x

12

Conf.Level L.E.pt U.E.pt

Lower Achieved CI 0.9297 2.000 25.000

Interpolated CI 0.9500 1.675 33.775

Upper Achieved CI 0.9922 1.000 52.000

The p-value = 0.289, do not reject the null hypothesis.

References

Banker, R. D., H. Chang, R. Natarajan. (2005). Productivity change, technical progress, and relative efficiency change in the public accounting industry. Management Sci. 51 291–304.

Davis, D. (2009). Declining Fruit and Vegetable Nutrient Composition: What Is the Evidence? HortScience, 44, 15-19.

Spalding, K. L. et al. Dynamics of fat cell turnover in humans. (2008) Nature 453, 783–787.

Back to Table of Contents

TEST 15 TWO SAMPLE T-TEST FOR THE DIFFERENCE IN SAMPLE MEANS

Question the test addresses

Is the difference between the mean of two samples significantly different from zero?

When to use the test?

You want to assess the extent to which the mean of two independent samples are different from each other. The test assumes the sample observations are normally distributed, and the sample variances are equal.

Practical Applications

Software Engineering: New releases of existing software bring with them additional features, and also some software bugs. Mohagheghi, Conradi and Schwarz (2004) report on the impact of reuse on defect density in a large scale telecom software system. The authors investigate whether reused software components are modified more than non-reused ones. The mean modification for reused components was 43%, and 57% for non-reused components. The authors us a two sample t-test (two tailed) for the difference in means and report a p-value of 0.001. The authors conclude that non-reused components are modified more than reused ones.

Dental Practice: Hashim and AlBarakati (2003) investigate the cephalometric soft tissue profile of 56 Saudi participants (30 males and 26 females). The authors investigate the difference between male and female facial angle of convexity. The average angle of convexity score was 2.6 and 4.2 for males and females respectively. The authors report a p-value of 0.28, and therefore cannot reject the null hypothesis of no difference between male and female angle of convexity.

Avian biodiversity: Tolbolka, Sparks and Tryjanowski (2012) investigate avian biodiversity near the towns of Gostyn and Koscian in Western Poland. Their study compared avian biodiversity between sites occupied by nesting White Storks and sites that were formerly occupied but were unoccupied during the two years (2007-2008) of the study. The researches use the Shannon Wiener index as a measure of avian biodiversity. The two sample t test was used to determine if there was a statistically significant difference in the Shannon Wiener index between occupied and unoccupied territories. For the 2007 samples, the p-value was 0.034. The authors reject the null hypothesis of no difference. The authors also use the two-sample

t-test to investigate the difference in mean bird species between occupied and unoccupied sites during 2007, with p-value of -0.62; the null hypothesis of no difference cannot be rejected.

How to calculate in R

The function t.test{stats} is used to perform this test. It takes the form:

t.test(x,y, alternative = "two.sided", var.equal=TRUE), Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Example: two sided test

Enter the following data

```
> x<-c(0.795,.864,.841,.683,.777,.720)

> y<-c(.765,.735,1.003,.778,.647,.740,.612)

> t.test(x,y, alternative = "two.sided", var.equal=TRUE)

   Two Sample t-test

data:  x and y

t = 0.4446, df = 11, p-value = 0.6652

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -0.1015879  0.1530164

sample estimates:

mean of x mean of y

0.7800000 0.7542857
```

Since the p-value is = 0.6652, do not reject the null hypothesis.

Example: one sided test using t.test

Enter the following data

```
> x<-c(0.795,.864,.841,.683,.777,.720)

> y<-c(.765,.735,1.003,.778,.647,.740,.612)

> t.test(x,y, alternative = "greater", var.equal=TRUE)

        Two Sample t-test
```

data: x and y

t = 0.4446, df = 11, p-value = 0.3326

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

 -0.07815739   Inf

sample estimates:

mean of x mean of y

0.7800000 0.7542857

The p-value is equal to 0.3326 therefore do not reject the null hypothesis.

References

Hashim HA, AlBarakati SF. (2003). Cephalometric soft tissue profile analysis between two different ethnic groups: a comparative study. J Contemp Dent Pract;4:60-73.

Marcin Tobolka, Tim H. Sparks & Piotr Tryjanowski. (2012). Does the White Stork Ciconia ciconia reflect farmland bird diversity? Ornis Fennica, vol 89.

Mohagheghi P, Conradi R, Killi OM, Schwarz H (2004) An empirical study o software reuse vs. defect density and stability. In: Proc. 26th Int'l Conf. on Software Engineering (ICSE'04), pp 282–292.

Back to Table of Contents

TEST 16 PAIRWISE T-TEST FOR THE DIFFERENCE IN SAMPLE MEANS

Question the test addresses

Is the difference between the mean of three or more samples significantly different from zero?

When to use the test?

You want to assess the extent to which the pairwise mean of more than two samples are different from each other. The test assumes the sample observations are normally distributed.

Practical Applications

Subgingival microbial flora during pregnancy: The subgingival bacterial flora from 2 gingival sites from twenty pregnant women, was cultured and characterized monthly in twenty periodontitis-free women during pregnancy and again post-partum by Kornman and Loesche (1980). At each visit the Gingival Index was estimated for the mesial of the maxillary right second premolar (Site 1) and of the mandibular left cuspid (Site 2) in each subject. Differences between the mean from different time periods were evaluated by means of a paired t-test. The mean Gingival Index at Site 1 and Site 2 increased significantly (pairwise t-test p-value < 0.05) from 1.16 initially to between 1.44 and 1.61 by 13-28 weeks.

Voice quality: Campbell and Mokhtari (2003) analyze voice quality using the normalized amplitude quotient (NAQ). Participants NAQ was assessed across a number of classes representing speaking style - “polite”, “friendly”, “casual”, “family”, “friends”, “others”, and “self-directed”. The researchers identify and assess 24 speech-act categories such as giving information, exclamations, requesting information, muttering and so on. The p-value between “family” and “friends” was less than 0.01. The p-value between “family” and “child” was 0.58. The p-value between “other” and “child” was 0.16. Overall, the researchers find all but the child-directed voice-quality differences are significant.

Covenants on bondholder wealth: Asquith and Wizman (1990) analyze the effect of covenants on bondholder wealth. Using data on 214 publicly traded bonds over the years 1980-1988 they categorize bond protection in three ways: strong, weak, and no protection. The pairwise t-statistic is used to assess statistical relationship of returns between these categories. The authors report the pairwise t-statistic between strong covenant protection

and both weak and no covenant protection are both significant (p-value <0.01). The pairwise t-statistic between weak and no protection was not significant (p-value >0.05). Asquith and Wizman conclude these results demonstrate that bonds with strong covenant protection have significantly larger abnormal returns in buyouts than bonds with weak or no protection.

How to calculate in R

The function pairwise.t.test{stats} is used to perform this test. It takes the form pairwise.t.test(sample, g, p.adjust.method ="holm", pool.sd = FALSE,alternative = "two.sided") where sample refers to the sample data and g represents the sample groups or levels. Note, to specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater"). The parameter p.adjust.method refers to the p-value adjustment due to the multiple comparisons. The adjustment methods include the Bonferroni correction ("bonferroni") in which the p-values are multiplied by the number of comparisons. Less conservative corrections include "holm", "hochberg", "hommel", "BH" (Benjamini & Hochberg adjustment), and "BY"(Benjamini & Yekutieli adjustment).

Example: using the “holm” adjustment

Suppose you have collected the following experimental data on three samples:

group	Value
1	2.9
1	3.5
1	2.8
1	2.6
1	3.7
2	3.9
2	2.5
2	4.3
2	2.7
3	2.9
3	2.4

3 3.8

3 1.2

3 2

Enter this data into R by typing:

```
sample_1 <- c(2.9, 3.5, 2.8, 2.6, 3.7)

sample_2 <- c(3.9, 2.5, 4.3, 2.7)

sample_3<- c(2.9, 2.4, 3.8, 1.2, 2.0)

sample <- c(sample_1, sample_2, sample_3)

g <- factor(rep(1:3, c(5, 4, 5)),

          labels = c("sample_1",

                     " sample_2",

                     " sample_3"))
```

To conduct the test type:

```
> pairwise.t.test(sample, g, p.adjust.method ="holm", pool.sd = FALSE,alternative = "two.sided")

    Pairwise comparisons using t tests with non-pooled SD

data:  sample and g

         sample_1  sample_2

 sample_2 0.64     -

 sample_3 0.59     0.59

P value adjustment method: holm
```

The p-value of sample 1 and sample 2 is 0.64 and not significant at the 5% level. The p-value between sample 2 and sample 3 is also not significant with a p-value of 0.59.

References

Asquith, P., & Wizman, T. A. (1990). Event risk, covenants, and bondholder returns in leveraged buyouts. Journal of Financial Economics, 27(1), 195-

213.

Campbell, N., & Mokhtari, P. (2003, August). Voice quality: the 4th prosodic dimension. In 15th ICPhS (pp. 2417-2420).

Kornman, K. S., & Loesche, W. J. (1980). The subgingival microbial flora during pregnancy. Journal of periodontal research, 15(2), 111-122.

Back to Table of Contents

TEST 17 PAIRWISE T-TEST FOR THE DIFFERENCE IN SAMPLE MEANS WITH COMMON VARIANCE

Question the test addresses

Is the difference between the mean of three or more samples significantly different from zero?

When to use the test?

You want to assess the extent to which the pairwise mean of more than two samples are different from each other. The test assumes the sample observations are normally distributed, and it uses a pooled estimate of sample variance, implying variances are equal across samples.

Practical Applications

Men's artistic gymnastics: Čuk and Forbes (2010) study exercise difficulty content of 49 gymnasts who competed at a European Championship qualification event. The researchers construct six variables of difficulty from official results. The six variables were, Floor Exercise (FX) , Pommel Horse (PH) , Rings (RI), Vault (VT), Parallel Bars (PB) and Horizontal Bar (HB Pairwise t-tests were used to assess the relationship between scores for each exercise. The researchers observe a p-value of 0.01 for PH and FX, and a p-value of 0.248 for FX and RI. Čuk and Forbes conclude the scores between the vault and other apparatus, and between the pommel horse and other apparatus were significantly different.

Green mussel growth: The effect of temperature on the development, growth, survival and settlement of Perna viridis (green mussel) was studied by Manoj and Appukuttan (2003). Settlement of samples was observed over a range of temperatures. The pairwise t-tests indicated that the settlement percentages did not differ significantly between 31 degrees Celsius and 29 degrees Celsius (p-value > 0.05). However, they did differ significantly from 27 degrees Celsius and 24 degrees Celsius. The researchers conclude settlement was best in the temperature range of 29 to 31 degrees Celsius.

Antagonistic coevolution: Using Tribolium castaneum and the microsporidian parasite Nosema whitei, five random populations were chosen as experimental lines for cross-infection by Bérénos et al (2012). In the "coevolution" regime, lines were subjected to coevolution with the Nosema whitei. In the "control" regime lines of identical origin and genetic background were maintained in the absence of parasites. The regimes were

maintained for a total of 16 generations. The researchers report variation in mortality among host lines upon exposure to parasite isolates did not differ significantly between control and coevolution treatment (pairwise t-test p-value = 0.253). Furthermore, variation in induced host mortality among parasite isolates did not differ between selection regimes (pairwise t-test p-value = 0.551).

How to calculate in R

The function pairwise.t.test{stats} is used to perform this test. It takes the form:

pairwise.t.test(sample, g, p.adjust.method ="holm", pool.sd = TRUE,alternative = "two.sided") where sample refers to the sample data and g represents the sample groups or levels.

Note, to specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater"). The parameter p.adjust.method refers to the p-value adjustment due to the multiple comparisons. The adjustment methods include the Bonferroni correction ("bonferroni") in which the p-values are multiplied by the number of comparisons. Less conservative corrections include "holm", "hochberg", "hommel", "BH" (Benjamini & Hochberg adjustment), and "BY"(Benjamin & Yekutieli adjustment).

Example: using the “holm” adjustment

Suppose you have collected the following experimental data on three samples:

group	Value
1	2.9
1	3.5
1	2.8
1	2.6
1	3.7
2	3.9
2	2.5

2 4.3

2 2.7

3 2.9

3 2.4

3 3.8

3 1.2

3 2

Enter this data into R by typing:

```
sample_1 <- c(2.9, 3.5, 2.8, 2.6, 3.7)
sample_2 <- c(3.9, 2.5, 4.3, 2.7)
sample_3<- c(2.9, 2.4, 3.8, 1.2, 2.0)
sample <- c(sample_1, sample_2, sample_3)
g <- factor(rep(1:3, c(5, 4, 5)),
          labels = c("sample_1",
                     " sample_2",
                     " sample_3"))
```

To conduct the test type:

```
> pairwise.t.test(sample, g, p.adjust.method ="holm", pool.sd = TRUE,alternative = "two.sided")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  sample and g

         sample_1  sample_2
 sample_2 0.65     -
 sample_3 0.46     0.38

P value adjustment method: holm
```

The p-value of sample 1 and sample 2 is 0.65 and not significant at the 5%

level. The p-value between sample 2 and sample 3 is also not significant with a p-value of 0.38.

References

Bérénos, C., Schmid-Hempel, P., & Wegner, K. M. (2012). Complex adaptive responses during antagonistic coevolution between Tribolium castaneum and its natural parasite Nosema whitei revealed by multiple fitness components. BMC evolutionary biology, 12(1), 11.

Čuk, I., & Forbes, W. (2010). How apparatus difficulty scores affect all around results in men's artistic gymnastics. Science of Gymnastics Journal, 2(3), 57-63.

Manoj Nair, R., & Appukuttan, K. K. (2003). Effect of temperature on the development, growth, survival and settlement of green mussel Perna viridis (Linnaeus, 1758). Aquaculture Research, 34(12), 1037-1045.

Back to Table of Contents

TEST 18 WELCH T-TEST FOR THE DIFFERENCE IN SAMPLE MEANS

Question the test addresses

Is the difference between the mean of two samples significantly different from zero?

When to use the test?

You want to assess the extent to which the mean of two independent samples are different from each other. The test assumes the sample observations are normally distributed, and the sample variances are not equal.

Practical Applications

Satellite telemetry: Satellite telemetry is commonly used to track marine animals whilst at sea. Vincent et al (2002) assess the accuracy of the Argos satellite system fixes on location categories from tags mounted on female grey seals. A total of 367 reliable fixes were captured over 61 seal days. A two sample Welch t-test is used to assess whether location errors by category were drawn from a population with mean equal to zero. For location category A, a p-value of 0.37 is observed for unfiltered data, the null hypothesis cannot be rejected. For location category B, a p-value of 0.048 is reported and the null hypothesis of zero mean is rejected at the 5% significance level.

Circadian rhythms: Lahti et al (2006) recruited ten healthy individuals into to a study to evaluate the effects of transition into daylight saving time on circadian rhythm activity. The participants all lived in Helsinki, Finland and were assigned to the groups “morning” and “intermediate”, based on daily activity patterns. The authors use a two sample Welch t-test to assess whether the rest-activity cycle was increased after the transition to daylight time. The p-value is 0.01, and the authors conclude the average level of rest activity was increased after transition among the morning group.

Ballistics: Barker et al (2012) analyze mine experimental impulse data for v-shaped structures constructed with a top floor plate. Eight normalized impulse measurements were obtained using centerline shots, with a mean of 0.780 and standard error of 0.028. Seven normalized impulse measurements were obtained using off-center shots, with a mean of 0.754 and standard error of 0.048. The Welch test for the difference between sample means had a p-value of 0.62. The null hypothesis could not be

rejected, and the authors conclude the mean observed impulse for the centerline and offset shots were not significantly different.

How to calculate in R

The function t.test{stats} is used to perform this test. It takes the form:

t.test(x,y, alternative = "two.sided", var.equal=FALSE),

Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Example: two sided test using t.test

Enter the following data

```
> sample1<-c(0.795,.864,.841,.683,.777,.720)
> sample2<-c(.765,.735,1.003,.778,.647,.740,.612)
> t.test(sample1,sample2, alternative = "two.sided", var.equal=FALSE)

        Welch Two Sample t-test

data:  sample1 and sample2
t = 0.4649, df = 9.565, p-value = 0.6524
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09829187  0.14972044
sample estimates:
mean of x mean of y
0.7800000 0.7542857
```

Since the p-value is = 0.6524, do not reject the null hypothesis.

Example: one sided test using t.test

Enter the following data

```
> sample1<-c(0.795,0.864,0.841,0.683,0.777,0.720)
> sample2<-c(0.765,0.735,1.003,0.778,0.647,0.740,0.612)
> t.test(sample1,sample2, alternative = "greater", var.equal=FALSE)
```

Welch Two Sample t-test

data: sample1 and sample2

t = 0.4649, df = 9.565, p-value = 0.3262

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.07500081 Inf

sample estimates:

mean of x mean of y

0.7800000 0.7542857

The p-value is 0.3262, do not reject the null hypothesis.

References

Barker Craig; Howle, Douglas; Holdren, Terry; Koch, Jeffrey; Ciappi, Raquel (2011). Results and Analysis from Mine Impulse Experiments Using Stereo Digital Image Correlation. 26th International Ballistics Symposium, 2011. Lancaster, PA: DEStech Publications, Inc.

Lahti, T.A., Leppämäki, S., Ojanen, S.-M., Haukka, J., Tuulio-Henriksson, A. Lönnqvist, J., and Partonen, T. (2006). Transition into daylight saving time influences the fragmentation of the rest-activity cycle. J. Circ. Rhythms 4, 1–6.

Vincent, C., B. J. McConnell, M. A. Fedak, and V. Ridoux.(2002). Assessmen of ARGOS location accuracy from satellite tags deployed on captive grey seals. Marine Mammal Science 18:301–322.

Back to Table of Contents

TEST 19 PAIRED T-TEST FOR THE DIFFERENCE IN SAMPLE MEANS

Question the test addresses

Is the difference between the mean of two samples significantly different from zero?

When to use the test?

This test is used when each subject in a study is measured twice, before and after a treatment. Alternatively, in a matched pairs experimental design, where subjects are matched in pairs and different treatments are given to each subject pair. Subjects are assumed to be drawn from a population with a normal distribution.

Practical Applications

Obesity: Flechtner-Mors et al (2000) investigate the long term contribution of meal and snack replacements on various health risk factors for obese patients. One hundred patients were divided into two groups; Group A received a calorie controlled diet and Group B an isoenergetic diet. After three months of weight loss, all patients were transferred to the calorie controlled diet. The authors use a paired t-test to assess the impact of the diet after 51 months on various health risk factors. For both groups, glucose and insulin levels were significantly improved compared to baseline values (p-value <0.01). For group B, significant reductions in systolic blood pressure and triacylglycerol were also observed (p-value < 0.01).

Shark Behavior: The social preferences of forty two juvenile lemon and nurse sharks was studied by Guttridge et al (2009). In one experiment each shark was given a choice between two empty compartments. The researchers hypothesized the sharks would have no preference for either compartment. A paired t-test could not reject the null hypothesis (p-value $= 0.517$). In another experiment, the researches investigate whether shark species differ in their degree of sociality. A paired t-test could not reject the null hypothesis of no difference (p-value $= 0.57$).

Stress Reduction: Mindfulness based stress reduction was analyzed in a controlled longitudinal study by Holzel et al (2011). Sixteen healthy participants undertook an eight week mindfulness program consisting of yoga, body scan and sitting meditation. Relative to baseline, participants experienced significant increases in awareness, observing and non-judging,

with paired t- test p-values of 0.003, 0.0001 and 0.003 respectively. A further seventeen individuals formed the control group and did not undergo the mindfulness program. For this group no significant difference in awareness, observing and non-judging was observed; the respective p-values were 0.498, 0.068, 0.523.The authors conclude, the participants who learnt the stress reduction techniques significantly increased their mindfulness.

How to calculate in R

The function t.test{stats} is used to perform this test. It takes the form:

t.test(final_value, initial_value, alternative = "two.sided", paired =TRUE),

Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Example: two sided test using t.test

Enter the following data

initial_value <- c(16,20,21,22,23,22,27,25,27,28)

final_value <- c(19,22,24,24,25,25,26,26,28,32)

To test the two-sided null hypothesis that the sample means are equal type

> t.test(final_value,initial_value, alternative = "two.sided", paired =TRUE)

Paired t-test

data: final_value and initial_value

t = 4.4721, df = 9, p-value = 0.00155

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.9883326 3.0116674

sample estimates:

mean of the differences

2

The p-value is equal to 0.001 and less than 0.05, reject the null hypothesis. The function also reports the 95% confidence interval as 0.98 to 3.0, as it

does not cross 0 (no difference between the two samples), reject the null hypothesis.

Example: one sided test using t.test

Enter the following data

initial_value <- c(16,20,21,22,23,22,27,25,27,28)

final_value <- c(19,22,24,24,25,25,26,26,28,32)

To test the two-sided null hypothesis that the sample means are equal type

> t.test(final_value,initial_value, alternative = "greater", paired =TRUE)

Paired t-test

data: final_value and initial_value

t = 4.4721, df = 9, p-value = 0.0007749

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

1.180207 Inf

sample estimates:

mean of the differences

2

The p-value is less than 0.01, reject the null hypothesis.

References

Fletchner-Mors, M., Ditschuneit, H. H., Johnson, T. D., Suchard, M. A., & Adler, G. (2000). Metabolic and weight-loss effects of a long-term dietary intervention in obese patients: A four-year follow-up. Obesity Research, 8, 399–402.

Guttridge, T.L., Gruber, S.H., Gledhill, K.S., Croft, D.P., Sims, D.W. and Krause, J. (2009) Social preferences of juvenile lemon sharks Negaprion brevirostris. Animal Behaviour, doi:10.1016/j.anbehav.2009.06.009.

Hölzel, B.K., Carmody, J., Vangel, M., Congleton, C., Yerramsetti, S.M., Gard T., & Lazar, S.W. (2011). Mindfulness practice leads to increases in regional brain gray matter density. Psychiatry Research, 191, 36–43.

Back to Table of Contents

TEST 20 MATCHED PAIRS WILCOXON TEST

Question the test addresses

Is the difference between the mean of two samples significantly different from zero?

When to use the test?

This test is used when each subject in a study is measured twice, before and after a treatment. Alternatively, in a matched pairs experimental design, where subjects are matched in pairs and different treatments are given to each subject pair. The test assumes the subjects are measured on a scale that allows rank ordering of observations. It is typically used when subjects cannot be assumed to be drawn from a population with a normal distribution.

Practical Applications

Baboon paternal care: The paternal care characteristics of wild savannah baboons living in the foothills of Mount Kilimanjaro, Kenya, was studied by Buchan et al (2003). The researchers were interested in whether male baboons helped their own genetic offspring more often than other offspring. For a sample of 15 males and a null hypothesis of no difference in help rendered, the authors find the Wilcoxon matched pairs test p-value is less than 0.01. They conclude adult males differentiate their offspring from unrelated juveniles.

Punishment: Traulsen, Rohl and Milinski (2012) investigate whether humans prefer pool or peer punishment. They design an economic game in which individuals make decisions as to whether or not to punish certain actions. In one game they form eight groups of five subjects and play with and without second order punishment. The researchers observe 87.5% of subjects, during last 10 rounds, choose pool over peer punishment. The matched pairs Wilcoxon test returned a p-value of 0.034. The researchers reject the null hypothesis in favor of the conclusion humans prefer pool punishment.

Bariatic surgery: Clark et al (2005) report the effects of bariatric surgery on non-alcoholic fatty liver disease. They measured liver histology (steatosis, inflammation, ballooning, perisinusoidal fibrosis and portal fibrosis) at the time of Roux-en-Y gastric bypass surgery after weight loss for sixteen patients. At baseline all patients had steatosis. At biopsy, which averaged 305 days after the first procedure, 18.8% of patients showed steatosis. The researchers used a matched pairs Wilcoxon test (p-value < 0.01), and

conclude surgery resulted in improvements in steatosis. The authors also used the matched pairs Wilcoxon test for inflammation, ballooning, perisinusoidal fibrosis and portal fibrosis. The resultant p-values were <0.001,<0.001,0.01 and 0.01 respectively. The authors conclude Roux-en-Y gastric bypass surgery improves liver histology in patients with non-alcoholic fatty liver disease.

How to calculate in R

The function wilcox.test{stats} is used to perform this test. It takes the form:

wilcox.test(initial_value, final_value, paired = TRUE, alternative = "two.sided").Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Example: two sided test using wilcox.test

Enter the following data

initial_value <- c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)

final_value <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)

To test the two-sided null hypothesis that the sample means are equal type

> wilcox.test(initial_value, final_value, paired = TRUE, alternative = "two.sided")

Wilcoxon signed rank test

data: initial_value and final_value

V = 40, p-value = 0.03906

alternative hypothesis: true location shift is not equal to 0

The p-value is equal to 0.039 and less than 0.05, reject the null hypothesis of no difference.

Example: two sided test using wilcox.test

Enter the following data

initial_value <- c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)

final_value <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)

To test the two-sided null hypothesis that the sample means are equal

type

> wilcox.test(initial_value, final_value, paired = TRUE, alternative = "greater")

Wilcoxon signed rank test

data: initial_value and final_value

V = 40, p-value = 0.01953

alternative hypothesis: true location shift is greater than 0

The p-value = 0.019 and less than 0.05, reject the null hypothesis.

References

Buchan, J. C., Alberts, S. C., Silk, J. B., & Altmann, J. (2003). True paterna care in a multi-male primate society. Nature, 425, 179–181.

Clark JM, Alkhuraishi AR, Solga SF, et al. (2005). Roux-en-Y gastric bypas improves liver histology in patients with non-alcoholic fatty liver disease. Obesity Research;13:1180–1186.

Traulsen A, Rohl T, Milinski M. (2012). An economic experiment reveals that humans prefer pool punishment to maintain the commons. Proc Biol Sci. Sep 22;279(1743):3716-21.

Back to Table of Contents

TEST 21 PAIRWISE PAIRED T-TEST FOR THE DIFFERENCE IN SAMPLE MEANS

Question the test addresses

Is the difference between the mean of two samples significantly different from zero?

When to use the test?

This test is used when you have multiple samples. It is used to assess the extent to which the pairwise mean differ from each other. The test is applied where each subject in a study is measured twice, before and after a treatment. Alternatively, in a matched pairs experimental design, where subjects are matched in pairs and different treatments are given to each subject pair. Subjects are assumed to be drawn from a population with a normal distribution.

Practical Applications

Effective teeth whitening treatments: Twenty four subjects were recruited into a study to compare the subjective clinical effects of three commercial 10% carbamide peroxide teeth bleaching systems by Tam (1999). Subjects were given a journal and asked to record daily the duration of bleaching and any subjective evaluations or effects of each bleaching agent. A pre-study photograph of the subjects teeth with and without a matching Vita shade guide tab was taken. After the bleaching treatment, a post-study photograph of each patient was taken and the daily logs were analyzed. Tam reports pairwise paired t-test indicate no statistical differences in the time of onset of subjective tooth whitening and the onset, frequency and duration of tooth sensitivity among the three commercial bleaching systems.

Autistic dialogue: Heeman et al (2010) explore differences in the interactional aspects of dialogue between children with Autistic Spectrum Disorder (ASD) and those with typical development (TD). A total of 22 TI children and 26 with ASD were assessed on three types of activity: converse is when there is no non-speech task; describe is when the child is doing a mental task, such as describing a picture; and play is when the child is interacting with the clinician in a play session. The assessment focused on pauses in activity. The paired pairwise t-test was used. The researchers report the p-value for TD and the converse activity as 0.3, for ASD anc converse activity it was 0.34. Neither are significant at the 5% level.

Phonetic realization: Reutskaja (2011) undertook experiments to investigate the effect of contextually salient neighbors on the phonetic realization of vowels and initial consonant aspiration. In one experiment target words were presented in the context of neighbors that differed only in onset , vowel, or coda positions. Twenty four participants spoke each of 48 target words twice in one of the four conditions: onset, vowel, coda, or filler word. Different neighbor types were matched for frequency using pairwise paired t-tests (p-value > 0:3 for all).

How to calculate in R

The function pairwise.t.test{stats} is used to perform this test. It takes the form:

pairwise.t.test(sample, g, p.adjust.method ="holm", paired= TRUE,alternative = "two.sided") where sample refers to the sample data and g represents the sample groups or levels. Note, to specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater"). The parameter p.adjust.method refers to the p-value adjustment due to the multiple comparisons. The adjustment methods include the Bonferroni correction ("bonferroni") in which the p-values are multiplied by the number of comparisons. Less conservative corrections include "holm", "hochberg", "hommel", "BH" (Benjamini & Hochberg adjustment), and "BY"(Benjamini & Yekutieli adjustment).

Example: using the “holm” adjustment

Suppose you have collected the following experimental data on three samples:

group	Value
1	2.9
1	3.5
1	2.8
1	2.6
1	3.7
1	4.0
2	3.9
2	2.5

2 4.3

2 2.7

2 2.6

2 3.0

3 2.9

3 2.4

3 3.8

3 1.2

3 2

3 1.97

Enter this data into R by typing:

```
sample_1 <- c(2.9, 3.5, 2.8, 2.6, 3.7,4.0)

sample_2 <- c(3.9, 2.5, 4.3, 2.7,2.6,3.0)

sample_3<- c(2.9, 2.4, 3.8, 1.2, 2.0,1.97)

sample <- c(sample_1, sample_2, sample_3)

g <- factor(rep(1:3, c(6, 6, 6)),

          labels = c("sample_1",

                     " sample_2",

                     " sample_3"))
```

To conduct the test type:

```
> pairwise.t.test(sample, g, p.adjust.method ="holm", paired= TRUE,alternative = "two.sided")

       Pairwise comparisons using paired t tests

data:  sample and g

         sample_1  sample_2

 sample_2 0.864    -

 sample_3 0.245    0.033
```

P value adjustment method: holm

The p-value of sample 1 and sample 2 is 0.864 and not significant at the 5% level. The p-value between sample 1 and sample 3 is also not significant with a p-value of 0.245; However, the p-value between sample 2 and sample 3 is significant at the 5% level.

References

Heeman, P. A., Lunsford, R., Selfridge, E., Black, L., & Van Santen, J. (2010 September). Autism and interactional aspects of dialogue. In Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 249-252). Association for Computational Linguistics.

Reutskaja, E., Nagel, R., Camerer, C. F., & Rangel, A. (2011). Search dynamics in consumer choice under time pressure: An eye-tracking study. The American Economic Review, 101(2), 900-926.

Tam, L. (1999). Clinical trial of three 10% carbamide peroxide bleaching products. Journal-Canadian Dental Association, 65, 201-207.

Back to Table of Contents

TEST 22 PAIRWISE WILCOX TEST FOR THE DIFFERENCE IN SAMPLE MEANS

Question the test addresses

Is the difference between the mean of two samples significantly different from zero?

When to use the test?

This test is used when you have multiple samples to assess the extent to which the pairwise mean differ from each other. It is applied where each subject in a study is measured twice, before and after a treatment. Alternatively, in a matched pairs experimental design, where subjects are matched in pairs and different treatments are given to each subject pair. The test is frequently used when subjects cannot be assumed to be drawn from a population with a normal distribution.

Practical Applications

Pollination Biology: The pollination biology of an annual endemic herb, physaria filiformis (brassicaceae), in the Missouri ozarks following controlled burns is considered by Edens-Meier et al (2011). To compare rates of self-compatibility, buds on each plant were divided into three experimental categories: The control group (mechanical self-pollination), Hand self-pollination (HSP), and Hand cross-pollination (HCP). Due to the non-normality of the sample the Pairwise Wilcox test was used. The results indicated significant differences in the number of pollen grains on the stigmas between the Control and HCP (p-value <0.01), Control and HSP (p value <0.01), and HCP and HSP (p-value <0.05). The researchers also tested for differences in the number of pollen tubes in the styles, between the control group and HCP (p-value <0.01), and between HCP and HSP (p-valu <0.01).

Soil strength: Graf, Frei and Böll (2009) tested three different soil samples - planted soil ,pure soil at low dry unit weight and pure compacted soil. The pairwise Wilcox test was used to make comparisons. The researchers report dry unit weights before consolidation of the compacted soil samples were significantly higher than those of both the planted soil samples (p-value< 0.05) and the pure soil samples at low dry unit weight (p-value< 0.05).

Living genera: Sites et al (1996) reconstruct phylogenetic relationships among the genera of the lizard family Iguanidae using various

morphological characters and molecular data. Two trees were constructed and then tested for significant differences in topologies. The researchers use the Pairwise Wilcox test for the difference in sample means to determine whether the most parsimonious topology obtained from each data set constitute suboptimal topologies for the other data sets. In the comparison of gene ND4 versus morphology between Tree 1 and Tree 2, the researchers report a p-value < 0.01.

How to calculate in R

The function pairwise.wilcox.test{stats} is used to perform this test. It takes the form: pairwise.wilcox.test (sample, g, p.adjust.method ="holm", paired= TRUE,alternative = "two.sided") where sample refers to the sample data and g represents the sample groups or levels. Note, to specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater"). The parameter p.adjust.method refers to the p-value adjustment due to the multiple comparisons. The adjustment methods include the Bonferroni correction ("bonferroni") in which the p-values are multiplied by the number of comparisons. Less conservative corrections include "holm", "hochberg", "hommel", "BH" (Benjamini & Hochberg adjustment), and "BY"(Benjamini & Yekutieli adjustment).

Example: using the “holm” adjustment

Suppose you have collected the following experimental data on three samples:

group	Value
1	2.9
1	3.5
1	2.8
1	2.6
1	3.7
1	4.0
2	3.9
2	2.5
2	4.3
2	2.7

2 2.6

2 3.0

3 2.9

3 2.4

3 3.8

3 1.2

3 2

3 1.97

Enter this data into R by typing:

```
sample_1 <- c(2.9, 3.5, 2.8, 2.6, 3.7,4.0)

sample_2 <- c(3.9, 2.5, 4.3, 2.7,2.6,3.0)

sample_3<- c(2.9, 2.4, 3.8, 1.2, 2.0,1.97)

sample <- c(sample_1, sample_2, sample_3)

g <- factor(rep(1:3, c(6, 6, 6)),

          labels = c("sample_1",

                     " sample_2",

                     " sample_3"))
```

To conduct the test type:

```
> pairwise.wilcox.test(sample, g, p.adjust.method ="holm", paired= TRUE,alternative = "two.sided")

       Pairwise comparisons using Wilcoxon signed rank test

data:  sample and g

         sample_1  sample_2

 sample_2 1.000    -

 sample_3 0.211    0.094

P value adjustment method: holm
```

The p-value of sample 1 and sample 2 is 1 and not significant at the 5% level. The p-value between sample 1 and sample 3 is also not significant with a p-value of 0.211; Finally, the p-value between sample 2 and sample 3 is significant at the 10% level.

References

Edens-Meier, R., Joseph, M., Arduser, M., Westhus, E., & Bernhardt, P. (2011). The Pollination Biology of an Annual Endemic Herb, Physaria filiformis (Brassicaceae), in the Missouri Ozarks Following Controlled Burns 1. The Journal of the Torrey Botanical Society, 138(3), 287-297.

Graf, F., Frei, M., & Böll, A. (2009). Effects of vegetation on the angle of internal friction of a moraine. For. Snow Landsc. Res, 82(1), 61-77.

Sites, J. W., Davis, S. K., Guerra, T., Iverson, J. B., & Snell, H. L. (1996 Character congruence and phylogenetic signal in molecular and morphological data sets: a case study in the living iguanas (Squamata, Iguanidae). Molecular Biology and Evolution, 13(8), 1087-1105.

Back to Table of Contents

TEST 23 TWO SAMPLE DEPENDENT SIGN RANK TEST FOR DIFFERENCE IN MEDIANS

Question the test addresses

Is the difference between the median of two samples significantly different from zero?

When to use the test?

This test is used when each subject in a study is measured twice, before and after a treatment. Alternatively, in a matched pairs experimental design, where subjects are matched in pairs and different treatments are given to each subject pair. The test assumes the underlying distribution, of the variables of interest, is continuous.

Practical Applications

Plant palatability: Increased herbicory at low latitudes is hypothesized, by Morrison and Hay (2012), to select for more effective plant defenses. To assess this hypothesis the researchers carried out a number of experiments involving feeding live or freeze dried low and high latitude plants to crayfish and the apple snail. The researches scored the number of times a high-latitude plant was significantly preferred to a low latitude one. Across the 66 sample assays run by the experimenters, the crayfish and snails preferred the high-latitude plant material 30% of the time and the low-latitude plant material 15% of the time. The null hypothesis of no difference in selection choice was assessed using the two sample dependent sign rank test. The p-value was 0.05, and the authors could not reject the null hypothesis of no difference.

Soil Science: Miller, Galbraith and Daniels (2004) investigate soil organic carbon in the Ridge and Valley of southwest Virginia. At various sites samples were taken in the litter layer (A horizon and B horizon) to a depth of one meter or bedrock. A sample of 12 measurements of bulk density each in the A and B horizons was collected. The researchers use a two sample sign rank test to assess the null hypothesis of no difference, with a significance level of 0.1. They report the p-value from the test as less than 0.1. The null hypothesis is rejected.

Ecological immunity: Otti et al (2011) analyze the relationship between immune response and predation in field crickets. As part of the study, immune challenged and control crickets were placed into artificial burrows. The researchers observed that control crickets were sitting 68% of the time

and the immune challenged 82% of the time. A sign test on 12 matched pairs of crickets resulted in a p-value of 0.39, and the null hypothesis of no difference in sitting time could not be rejected.

How to calculate in R

The function SIGN.test{BSDA}can be used to perform this test. It takes th form:

SIGN.test(initial_value,final_value, alternative = "two.sided"), Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Example: two sided test using SIGN.test

Enter the following data

initial_value <- c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)

final_value <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)

To test the two-sided null hypothesis that the sample medians are equal type

> SIGN.test(initial_value,final_value, alternative = "two.sided")

Dependent-samples Sign-Test

data: initial_value and final_value

S = 7, p-value = 0.1797

alternative hypothesis: true median difference is not equal to 0

95 percent confidence interval:

-0.0730000 0.9261778

sample estimates:

median of x-y

0.49

Conf.Level L.E.pt U.E.pt

Lower Achieved CI 0.8203 0.010 0.6200

Interpolated CI 0.9500 -0.073 0.9262

Upper Achieved CI 0.9609 -0.080 0.9520

Since the p-value is greater than 0.05, do not reject the null hypothesis.

Example: one sided test using SIGN.test:

Enter the following data

initial_value <- c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)

final_value <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)

To test the two-sided null hypothesis that the sample medians are equal type

> SIGN.test(initial_value,final_value, alternative = "greater")

Dependent-samples Sign-Test

data: initial_value and final_value

S = 7, p-value = 0.08984

alternative hypothesis: true median difference is greater than 0

95 percent confidence interval:

-0.041 Inf

sample estimates:

median of x-y

0.49

Conf.Level L.E.pt U.E.pt

Lower Achieved CI 0.9102 0.010 Inf

Interpolated CI 0.9500 -0.041 Inf

Upper Achieved CI 0.9805 -0.080 Inf

The p-value is equal to 0.089, do not reject the null hypothesis.

References

Miller, J.O., J.M. Galbraith and W.L. Daniels. 2004. Organic carbon conten and variability in frigid Southwest Virginia mountain soils. Soil Sci. Soc. Am J. 68:194–203.

Morrison, W. E. and Hay, M. E. (2012). Are lower latitude plants better defended?: Palatability of freshwater macrophytes. Ecology 93: 65–74.

Otti, O.; Gantenbein-Ritter, I.; Jacot, A.; Brinkhof, M.G.W. (2012). Immune response increases predation risk. Evolution, 66, 732-739.

Back to Table of Contents

TEST 24 WILCOXON RANK SUM TEST FOR THE DIFFERENCE IN MEDIANS

Question the test addresses

Is the difference between the median of two samples significantly different from zero?

When to use the test?

You want to assess the extent to which the median of two independent samples are different from each other. The test is less sensitive to outliers than the two sample t-test. Note, the test is sometimes referred to as the Mann–Whitney U test, or the Mann–Whitney–Wilcoxon test.

Practical Applications

Robot foraging: Wischmann, Floreano and Keller (2012) study the evolution of communication systems in populations of cooperatively foraging simulated robots. Each population consisted of 100 groups of 20 simulated robots who evolved over 1,000 generations whilst foraging for a food source. The researchers observed two main signaling systems evolved surrounding the food source: one signal communication and two signal communication. The one signal populations had a foraging food score of 0.196, whilst the two signal populations scored 0.168. The researchers use the Wilcoxon rank sum test to assess the statistical significance of the difference. They report a p-value of less than 0.001, and reject the null hypothesis of no difference.

Wild crops: The flowering time of ten populations of wild wheat and ten populations of wild barely growing in Israel over the period 1980 to 2008 was analyzed by Nevo et al (2012). The researchers observed a shorting in flowering time of both crops over the 28 year time frame of the study. The average shortening in wild wheat was 8.53 days, and in wild barley 10.94 days. The difference between the species was significant (Wilcoxon ran sum test p-value less than 0.01).

Ant foraging: Dolezal et al (2012) compare the foraging behavior of single age cohort colonies of harvester ants to mature colonies. The researchers created four single age cohort colonies by removing ants of differential age and replacing them with same age ants. They observed, on average, single cohort initiated foraging five times earlier than mature colony ants. The Wilcoxon rank sum test (which they call the Mann-Whitney U-test) is used to assess the statistical significance of the difference in initiation of

foraging time. They report a p-value of less than 0.001, and reject the null hypothesis of no difference.

How to calculate in R

The function wilcox.test{stats} is used to perform this test. It takes the form:

wilcox.test(x,y, alternative = "two.sided"), Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Example: two sided test using wilcox.test

Enter the following data

> x<-c(0.795,0.864,0.841,0.683,0.777,0.720)

> y<-c(0.765,0.735,1.003,0.778,0.647,0.740,0.612)

> wilcox.test(x,y, alternative = "two.sided")

Wilcoxon rank sum test

data: x and y

W = 27, p-value = 0.4452

alternative hypothesis: true location shift is not equal to 0

The p-value is equal to 0.4452, do not reject the null hypothesis.

Example: two sided test using wilcox.test

Enter the following data

> x<-c(0.795,0.864,0.841,0.683,0.777,0.720)

> y<-c(0.765,0.735,1.003,0.778,0.647,0.740,0.612)

> wilcox.test(x,y, alternative = "greater")

Wilcoxon rank sum test

data: x and y

W = 27, p-value = 0.2226

alternative hypothesis: true location shift is greater than 0

The p-value is equal to 0.2226, do not reject the null hypothesis.

References

Dolezal AG, Brent CS, Hölldobler B, Amdam GV (2012) Worker division c labor and endocrine physiology are associated in the harvester ant, Pogonomyrmex californicus.. J Exp Biol 215: 454–460.

Nevo, E.; Fu, Y.B.; Pavlicek, T.; Khalifa, S.; Tavasi, M.; Beiles, A. Evolution o wild cereals during 28 years of global warming in Israel. Proc. Natl. Acad Sci. USA 2012, 109, 3412-3415.

Wischmann, S., Floreano, D. & Keller, L. (2012) Historical contingency affects signaling strategies and competitive abilities in evolving populations of simulated robots. Proc. NatlAcad. Sci. USA109, 864–868.

Back to Table of Contents

TEST 25 WALD-WOLFOWITZ RUNS TEST FOR DICHOTOMOUS DATA

Question the test addresses

Is the sequence of binary events in a sample randomly distributed?

When to use the test?

To test the hypothesis that the elements of the sequence of dichotomous data in a sample are random. A run is defined as a series of increasing values or a series of decreasing values. The number of increasing, or decreasing, values is the length of the run.

Practical Applications

Indian Stock Market: Kurmar and Kurmar (2012) study the efficiency of the National Stock Exchange in India, which is the market index of India's largest stock exchange. Daily closing values of the index are collected over the period 1 January 2003 to 31 March 2011. The researchers test for randomness of daily stock prices. They report a p-value of less than 0.001, and reject the null hypothesis that daily fluctuations in stock prices are random.

Pulsar nulling: Pulsar nulling is the sudden cessation in pulsar emission. Redman and Rankin discuss how the Wald-Wolfowitz runs test can be used by astronomers to identify pulsars that have non-random nulls. Observations on eighteen pulsars are collected. For pulsar B0834+06, the null hypothesis could not be rejected and the authors conclude this pulsar has a random null. However, the vast majority of pulsars rejected the null hypothesis, leading the researchers to conclude the majority of pulsars null non-randomly.

Mosquito Feeding: Oliveira et al (2012) compare mosquito feeding patterns in the Tuskegee National Forest in south-central Alabama. Mosquito's in this region feed on both avian (yellow-crowned night heron, great blue heron) and mammalian hosts (white-tailed deer). A total of 1099 meals of the Culex erraticus mosquito were collected and analyzed. A runs test revealed the patterns of feeding were not randomly distributed over time (p-value <0.05). The researchers also applied the runs test to assess the feeding patterns upon different hosts. The p-value for the Yellow-crowned night heron, Great blue heron and white-tailed deer were 0.0141,0.0101 and 0.0001 respectively. In all cases the null hypothesis of randomness was rejected.

How to calculate in R

The function runs.test{tseries} is used to perform this test. It takes the form:

runs.test(binary_factor, alternative = "two.sided"), To conduct a one sided test set alternative = “less” or alternative =”greater”. There are two types of non-random sequences: those that are ‘over-clustered’ (set alternative =”greater”) and those that are ‘over-scattered’ (alternative = “less”).

Example: two sided test using runs.test

Enter the following data

```
> binary_factor<-factor(c(1,0,0,0,0,0,0,0,1,1,1,1,1,0,1,1,1,1,1,1,1,1,0,0,0,0,0))

> runs.test(binary_factor,alternative ="two.sided")

        Runs Test

data:  binary_factor

Standard Normal = -3.2026, p-value = 0.001362

alternative hypothesis: two.sided
```

Since the p-value is less than 0.05, reject the null hypothesis of randomness.

Example: one sided test using runs.test:

Enter the following data

```
> binary_factor<-factor(c(1,0,0,0,0,0,0,0,1,1,1,1,1,0,1,1,1,1,1,1,1,1,0,0,0,0,0))

> runs.test(binary_factor,alternative ="less")

        Runs Test

data:  binary_factor

Standard Normal = -3.2026, p-value = 0.0006811

alternative hypothesis: less
```

Since the p-value is less than 0.05, reject the null hypothesis.

References

Kumar, A., & Kumar, S. (2012). Weak Form Efficiency of Indian Stock Market: A Case of National Stock Exchange (NSE). International Journal o

Management Sciences, 12(1), 27-31.

Oliveira, A., C. R. Katholi, N. Burkett-Cadena, H. K. Hassan, S. Kristensen and T. R. Unnasch. 2011. Temporal analysis of feeding patterns of Culex erraticus in Central Alabama. Vector Borne Zoon. Dis. 11: 413–421.

Redman , Stephen L. and Rankin, Joanna M. (2009). On the randomness o pulsar nulls. Mon. Not. R. Astron. Soc. 395, 1529–1532. doi:10.1111/j.1365-2966.2009.14632.x

Back to Table of Contents

TEST 26 WALD-WOLFOWITZ RUNS TEST FOR CONTINUOUS DATA

Question the test addresses

Is the sequence of observations in a sample randomly distributed?

When to use the test?

To test the hypothesis that the elements of the sequence of data in a sample are random. A run is defined as a series of increasing values or a series of decreasing values. The number of increasing, or decreasing, values is the length of the run.

Practical Applications

Blue whale communication: As part of a survey into the vocal behavior of blue whales during seismic surveys, Di lorio and Clark (2009), collected vocal activity data of blue whales. Over four days when no seismic activity was taking place sound data was collected. The data from each day was broken into 10 minute intervals, and the number of whale calls determined. The same procedure was repeated for four days where seismic activity was present. The Wald-Wolfowitz runs test was used to determine the randomness within a sample. For all daily samples (with and without seismic activity) , the researchers fail to reject the null hypothesis (p-value >0.05). The researchers conclude the Wald-Wolfowitz runs test revealed that the samples were independent.

Surgical site infection: Hollenbeak et al (2000) examine how deep chest surgical site infections following coronary artery bypass graft surgery impact hospital inpatient length of stay, costs and mortality. In total 41 patients, from a community medical center, developed deep chest infection. The researchers used the Wald-Wolfowitz runs test to investigate whether infections were randomly distributed across time. The p-value was 0.31, the null hypothesis cannot be rejected and the researchers conclude there is no evidence that the infections occurred in clusters.

Honeybees: The effect of Nosema caranae infection on honeybee sensitivity to subleathal doses of insecticides fiponil and thiacloprid was investigated by Vidau et al (2011). The Wald-Wolfowitz runs test was used to assess whether the uptake of insecticide in bees infected with Nosema caranae was random. Two separate samples were investigated. Infected bees exposed to fipronil, and infected bees exposed to thiacloprid. The runs test revealed the consumption of insecticide in infected bees was not

non-random. For the sample of infected bees exposed to fipronil the p-value was less than 0.01, and for infected bees exposed to thiacloprid the p-value was less 0.01.

How to calculate in R

The function runs.test{lawstat} is used to perform this test. It takes the form:

runs.test(y, alternative = "two.sided"). Note to conduct a one sided test set alternative = "positive.correlated" or alternative ="negative.correlated".

Example: two sided test using runs.test

Enter the following data

```
> y=c(1.8,2.3,3.5,4,5.5,6.3,7.2,8.9,9.1)

> runs.test(y, alternative = "two.sided")

        Runs Test - Two sided

data:  y

Standardized Runs Statistic = -2.49, p-value = 0.01278
```

Since the p-value is less than 0.05, reject the null hypothesis of randomness. Notice this data only has one run (each value is higher than the last) and so is highly unlikely to be random.

Example: one sided test using runs.test

Enter the following data

```
> y=c(1.8,2.3,3.5,4,5.5,6.3,7.2,8.9,9.1)

> runs.test(y, alternative = "positive.correlated")

        Runs Test - Positive Correlated

data:  y

Standardized Runs Statistic = -2.49, p-value = 0.006388
```

Since the p-value is less than 0.05, reject the null hypothesis.

References

Di Iorio, L. & Clark, C. W. 2009 Exposure to seismic survey alters blue whale communication. Biol. Lett. 6, 51–54. (doi:10.1098/rsbl.2009.0651).

Hollenbeak CS, Murphy DM, Koenig S, Woodward RS, Dunagan WC, Frase VJ.(2000) The clinical and economic impact of deep chest surgical site infections following coronary artery bypass graft surgery. Chest;118:397—402.

Vidau C, Diogon M, Aufauvre J, Fontbonne R, Vigues B, et al. (2011 Exposure to sublethal doses of fipronil and thiacloprid highly increases mortality of honey bees previously infected by Nosema ceranae. PLoS One 6: e21550.

Back to Table of Contents

TEST 27 BARTELS TEST OF RANDOMNESS IN A SAMPLE

Question the test addresses

Is the sequence of observations in a sample randomly distributed?

When to use the test?

To test the hypothesis that the elements of the sequence of data in a sample are random. A run is defined as a series of increasing values or a series of decreasing values. The number of increasing, or decreasing, values is the length of the run.

Practical Applications

Corporate income: Bartels (1982) investigated the distribution of undistributed income of companies in Australia over the years 1959 to 1978. Adjusting by the Gross Domestic Product price index as a deflator, Bartels test is significant at the 1% level. The author concludes undistributed income of companies in Australia does not follow a random walk.

Soil Biology: Spatial and seasonal variation of gross nitrogen transformations in grasslands near Hancock, Pennsylvania were studied by Corre, Schnabel and Stout (2002). The researchers created three topographic units based on soil and drainage. Within each type ten measurements, equally spaced at 10 meters apart, were obtained. The researchers used Bartels test of randomness to determine whether the samples within topographic units were random. They could not reject the null hypothesis (p-value > 0.05) and therefore conclude the ten sampling points in each topographic were statistically independent.

Psychomotor vigilance: Rajaraman et al (2012) develop a psychomotor vigilance metric for quantifying the effects of sleep loss on performance impairment. Measurements on twelve adults subjected to 85 hours of extended wakefulness, followed by 12 hours of recovery, were used to construct various process models for a psychomotor vigilance metric. Bartels test was used on the residual of fitted models to assess the goodness of fit. For the two-process model a p-value of 0.01 was reported, and the null hypothesis of random residuals (and the model) was rejected.

How to calculate in R

The function bartels.test{lawstat} is used to perform this test. It takes the form:

runs.test(y, alternative = "two.sided"). Note to conduct a one sided test set alternative = "positive.correlated" or alternative = "negative.correlated".

Example: two sided test using bartels.test

Enter the following data

> y<-c(-82.29,-31.14,136.58,85.42,42.96,-122.72,0.59,55.77,117.62,-10.95,-211.38,-304.02,30.72,238.19,140.98,18.88,-48.21,-63.7)

> bartels.test(y,alternative="two.sided")

Bartels Test - Two sided

data: y

Standardized Bartels Statistic = -1.8915, RVN Ratio = 1.108, p-value = 0.05856

Since the p-value is greater than 0.05, do not reject the null hypothesis of randomness.

Example: one sided test using bartels.test:

Enter the following data

> y<-c(-82.29,-31.14,136.58,85.42,42.96,-122.72,0.59,55.77,117.62,-10.95,-211.38,-304.02,30.72,238.19,140.98,18.88,-48.21,-63.7)

> bartels.test(y,alternative = "positive.correlated")

Bartels Test - Positive Correlated

data: y

Standardized Bartels Statistic = -1.8915, RVN Ratio = 1.108, p-value = 0.02928

Since the p-value is less than 0.05, reject the null hypothesis.

References

Bartels, R. (1982), "The Rank Version of von Neumann's Ratio Test for Randomness," Journal of the American Statistical Association, 77, 40-46.

Herbst ,Anthony F. Slinkman; and Craig W. (1984). Political-Economic Cycles in the U.S. Stock Market. Financial Analysts Journal. Vol. 40, No. 2, pp. 38 44.

Rajaraman, Srinivasan et al (2012). A new metric for quantifying

performance impairment on the psychomotor vigilance test. Journal of Sleep Research. March. doi: 10.1111/j.1365-2869.2012.01008.x.

Back to Table of Contents

TEST 28 LJUNG-BOX TEST

Question the test addresses

Is the sequence of observations in a sample randomly distributed?

When to use the test?

To test the hypothesis that the elements of the sequence of data in a sample are random. The Ljung-Box test is based on the autocorrelation plot. If the autocorrelations are very small, we conclude that series is random. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a pre-specified number of lags. There are a number of rules of thumb for choosing the lag length. The first is to set it to ln(n), where n is the number of observations in the sample and ln() is the natural logarithm. An alternative rule sets it to 20, if the sample size is reasonable large.

Practical Applications

Onchocerciasis cases in Mexico: Lara-Ramírez et al (2013) study data or onchocerciasis cases in Chiapas and Oaxaca, Mexico. Monthly data of onchocerciasis cases between 1988 and 2010 were modeled using time-series models. The researchers developed two models, one for Chiapas and the other for Oaxaca. The best-fit model for Oaxaca was a mixed Auto-regressive integrated moving average (ARIMA) seasonal non-stationary model. The Ljung–Box test was used to assess the independence of the residuals (p- value = 0.93); It did not reject the null hypothesis of independence in the residuals of the Oaxaca time series model. The best-fit model for Chiapas was a mixed ARIMA seasonal non-stationary model. The Ljung–Box test was used to assess the independence of the residuals (p-value = 0.34); It did not reject the null hypothesis of independence in the residuals of the Oaxaca time series model.

Seed dispersal: Maurer et al (2013) investigate seed dispersal by the tropical tree, Luehea seemannii in the Parque Natural Metropolitano, Panama. A nine-month data set of wind speed in three dimensions and turbulence (February through October, 2007) was used in the analysis. In addition long-term measurements of above-canopy wind (hourly mean horizontal wind speed and temperature from 2000 to 2010). A multivariate regression model between seed abscission and the observed environmental factors is constructed. The goodness-of-fit of the final model was evaluated by testing the residuals for independence using the Box–Ljung test. The researchers report the best-fit model and the second-

best-fit model, the residuals can be considered independent (Ljung-Box test p-value >0.05).

Angelman Syndrome: Allen et al (2013) evaluate the effectiveness of a behavioral treatment package to reduce chronic sleep problems in children with Angelman Syndrome. Five children (Annie,Bobby ,Eddie, Cindy and Darcy) between the ages of 2 to 11 years old were recruited onto the study. Sleep and disruptive nighttime behaviors were logged by parents in sleep diaries. Actigraphy was added to provide independent evaluations of sleep–wake activity. The researchers report that Annie,Bobby and Eddie had no statistically significant autocorrelations (Ljung–Boxtest p-value >0.05). Cindy showed significant auto correlation at lag 1 (Ljung–Boxtest p-value <0.05). Darcy showed autocorrelation at lag 1 (Ljung–Boxtest p-value <0.01), lag 2 (Ljung–Boxtest p-value <0.01), lag 3 (Ljung–Boxtest p-value <0.01), lag 4 (Ljung–Boxtest p-value <0.05), lag 5 (Ljung–Boxtest p-value <0.05) and lag 6 (Ljung–Boxtest p-value <0.05).

How to calculate in R

The function Box.test{stats} is used to perform this test. It takes the form:

Box.test (series, lag = 1, type = "Ljung-Box"). Note series refers to the time-series you wish to test, lag refers to the number of autocorrelation coefficients you want to test.

Example:

Enter the following data

y<-c(-82.29,-31.14,136.58,85.42,42.96,-122.72,0.59,55.77,117.62,-10.95,-211.38,-304.02,30.72,238.19,140.98,18.88,-48.21,-63.7)

> Box.test (y, lag = 3,type = "Ljung-Box")

Box-Ljung test

data: y

X-squared = 18.9507, df = 3, p-value = 0.0002799

Since the p-value is less than 0.05, reject the null hypothesis of randomness.

References

Allen, K. D., Kuhn, B. R., DeHaai, K. A., & Wallace, D. P. (2013). Evaluation of a behavioral treatment package to reduce sleep problems in children with

Angelman Syndrome. Research in developmental disabilities, 34(1), 676-686.

Lara-Ramírez, E. E., Rodríguez-Pérez, M. A., Pérez-Rodríguez, M. A., Adeleke M. A., Orozco-Algarra, M. E., Arrendondo-Jiménez, J. I., & Guo, X. (2013 Time Series Analysis of Onchocerciasis Data from Mexico: A Trend towards Elimination. PLOS Neglected Tropical Diseases, 7(2), e2033.

Maurer, K. D., Bohrer, G., Medvigy, D., & Wright, S. J. (2013). The timing of abscission affects dispersal distance in a wind-dispersed tropical tree. Functional Ecology, 27(1), 208-218.

Back to Table of Contents

TEST 29 BOX-PIERCE TEST

Question the test addresses

Is the sequence of observations in a sample randomly distributed?

When to use the test?

To test the hypothesis that the elements of the sequence of data in a sample are random. The test is based on the autocorrelation. If the autocorrelations are very small, we conclude that series is random. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a pre-specified number of lags. There are a number of rules of thumb for choosing the lag length. The first is to set it to the ln(n), where n is the number of observations in the sample. An alternative rule sets it to 20, if the sample size is reasonable large.

Practical Applications

Monthly rainfall in Dhaka: Mahsin (2011) builds a seasonal time-series model for monthly rainfall data in Dhaka, Bangladesh over the period 1981-2010. The researcher reports a seasonal cycle in the raw data and plot the autocorrelation and partial autocorrelation functions. The Box-Pierce statistic (lag 4) rejects the null hypothesis (Box-Pierce p-value <0.05).The author performs a log transformation and first order difference of the original data. The transformed series is reported as independent (Box-Pierce p-value >0.05).

United States macroeconomics 1860-1988: Darné and Charles (2011) study 14 U.S. macroeconomic and financial time-series - Real GNP, nominal GNP, real per capita GNP industrial production, employment, unemployment, GNP deflator, consumer price, nominal wages, real wages, money stock, velocity, interest rate, and stock price. The data consists of annual observations which begins between 1860 and 1909 and end 1988. In testing for independence the researchers report the Box-Pierce statistics are not significant for all series (p-value >0.05). Removing outliers from the stock price variable results in a rejection of the null hypothesis (p-value <0.05). The researchers conclude there is no serial linear correlation in the data, except in the stock price, when the data are corrected of outliers.

Volatility in US housing: Li (2012) compare in-sample estimation of the real estates related financial data relative to out-of-sample conditional mean and volatility forecast using a variety of Generalized Auto-regressive Conditional Heteroskedasticity models. Five housing market variables were used in the analysis, housing price index (HPI), total home market amount

(RHMA) , loan to price ratio (LTP),consumer loans (CL) and inter-bank loans (IL). For the data series on RHMA, LTP, CL and IL, the sample period wen from January 1988 to February 2009. For HPI it covered the period from January 1991 to February 2009. The Box-Pierce test, with lag set to 5, was used to assess the serial correlation in each of the five variables. The null hypothesis of no serial correlation was rejected for HPI (p-values <0.000), RHMA (p-values <0.000), LTP(p-values <0.0012) and CL (p-values <0.000) The null hypothesis could not be rejected for IL (p-value =0.422). The researchers also apply the test to the squared returns of the five variables. The null hypothesis of no serial correlation was rejected for HPI (p-values <0.000), RHMA(p-values <0.000), LTP(p-values <0.025) and IL (p-values <0.007). The null hypothesis could not be rejected at the 5% level of significance for CL (p-value =0.095).

How to calculate in R

The function Box.test{stats} is used to perform this test. It takes the form:

Box.test (series, lag = 1, type = "Box-Pierce"). Note the parameter series refers to the time-series you wish to test, lag refers to the number of autocorrelation coefficients you want to test.

Example:

Enter the following data

y<-c(-82.29,-31.14,136.58,85.42,42.96,-122.72,0.59,55.77,117.62,-10.95,-211.38,-304.02,30.72,238.19,140.98,18.88,-48.21,-63.7)

To carry out the test with a lag of 3 enter.

> Box.test (y, lag = 3,type = "Box-Pierce")

Box-Pierce test

data: y

X-squared = 14.7694, df = 3, p-value = 0.002025

Since the p-value is less than 0.05, reject the null hypothesis of randomness.

References

Darné, O., & Charles, A. (2011). Large shocks in US macroeconomic time series: 1860–1988. Cliometrica, 5(1), 79-100.

Li, K. W. (2012). A study on the volatility forecast of the US housing market

in the 2008 crisis. Applied Financial Economics, 22(22), 1869-1880.`

Mahsin, M. (2011). Modeling Rainfall in Dhaka Division of Bangladesh Usin Time Series Analysis. Journal of Mathematical Modelling and Application 1(5), 67-73.

Back to Table of Contents

TEST 30 BDS TEST

Question the test addresses

Is the sample independent and identical distributed?

When to use the test?

To test for independence in a time series, it is frequently used as a diagnostic for residuals in statistical models. This procedure tests for the joint null hypothesis of independence and identical distribution. It tests the null hypothesis by measuring the degree of spatial correlation in the sequence. In essence, this is achieved by searching for sub-sequences of length m that are significantly different from other m-long sub-sequences in the sample; the value of m is referred to as the 'embedding dimension'. Rejection of the null hypothesis implies non-stationarity of the sample (e.g., existence of trends), or the fact that there are linear or non-linear dependencies in the sample.

Practical Applications

Nonlinearity in the Istanbul Stock Exchange: Özer and Ertokatlı (2010 examine nonlinearity in the Istanbul Stock Exchange (ISE) all share equity indices. The sample consisted of 3,036 observations of the daily ISE closing price over the period 02 February 1997 to 16 March 2009. Daily returns were calculated as the change in logarithm of closing stock market indices of successive days. The best fitting autoregressive integrated moving average (ARIMA) model is fitted to data. The researchers do this to eliminate linearity from the data. Then the BDS test is applied to the residuals of that ARIMA model, which by default must be linearly independent, so that any dependence found in the residuals must be nonlinear in nature. The researchers report the besting fitting model is an ARIMA (0,1,3) . The researchers use an embedding dimension up to 5, with the distance between points ranging from 0.5 to 2 standard deviations. This results in a grid of 16 p-values, all of which are significant at the 5% level. The researchers observe the rejection could be due to linear serial dependencies in the residuals, non-stationary in the residuals or a nonlinear serial dependency in the residuals (chaotic or stochastic).

Modeling aperiodic traffic flow: Khan et al (2009) derive a model of inter-arrival arrival patterns for aperiodic traffic. Data was collected from measurements taken on-board of a PSA Peugeot-Citroën vehicle. The researchers used the BDS test to assess whether aperiodic inter-arrivals are independent and identically distributed. They carried out the BDS test for

various combinations of embedding dimension. For many combinations they could not reject the null hypothesis at the 1% confidence level. The researchers conclude that it is possible to model aperiodic inter-arrival traffic by a random variable obeying a memory-less probabilistic distribution.

Indian rupee- US dollar exchange rate: Pal (2011) investigates the non-linearity property of the real exchange rate of the Indian rupee-US dollar over the period 1959-2001. The logarithm of the annuals spot exchange rate is assessed with the BDS test. The researcher chooses the best fitting autoregressive (AR) model for this data, and identifies an AR(1) to be optimal. For the BDS test an embedding dimension of 2 to 4 is specified, with the distance between points ranging from 0.5 to 2 standard deviations. The resultant grid of p-values are all significant at the 1% level. Due to the limited sample size, the author also performs a bootstrapped BDS test using 1,000 observations per bootstrap. The resultant grid of p-values are all significant at the 1% level. The researcher concludes the Indian-US dollar real exchange is non-linear.

How to calculate in R

The function bds.test{tseries}can perform this test. It can be used in the form:

bds.test(data, m , eps).Common practice is to test for a range of embedding dimensions (typically m = 2 through 8). However, the fewer the number of observations, the lower the maximum embedding dimension. Given an embedding dimension, eps should be selected such that the expected number of m-histories is large enough and varies little to achieve reliable estimation of the probability that two m length vectors are within eps. A common practice is to set using the standard deviation of the data (sd) so that eps = seq(0.5 * sd(data), 2 * sd(data). Note that Brock, Hsieh and LeBaron (1991) point out that samples with fewer than 500 observations are generally not reliable.

Example:

We illustrate the use of this test statistics on data which we know to be independent and identically distributed

```
set.seed(1234)

x <- rnorm(5000)
```

To carry out the test with to the 6th dimension enter.

```
> bds.test(x,m=6)

        BDS Test

data:  x

Embedding dimension =  2 3 4 5 6

Epsilon for close points =  0.4957 0.9913 1.4870 1.9827

Standard Normal =

     [ 0.4957 ] [ 0.9913 ] [ 1.487 ] [ 1.9827 ]

[ 2 ]    -0.3192    -0.4794   -0.5354    -0.4395

[ 3 ]    -0.8099    -0.8871   -0.9134    -0.8516

[ 4 ]    -1.0534    -1.0290   -0.9986    -0.8969

[ 5 ]    -1.5586    -1.6091   -1.4851    -1.2751

[ 6 ]    -1.7497    -1.6681   -1.4979    -1.2518

p-value =

     [ 0.4957 ] [ 0.9913 ] [ 1.487 ] [ 1.9827 ]

[ 2 ]     0.7495     0.6316    0.5924     0.6603

[ 3 ]     0.4180     0.3750    0.3610     0.3945

[ 4 ]     0.2921     0.3035    0.3180     0.3698

[ 5 ]     0.1191     0.1076    0.1375     0.2023

[ 6 ]     0.0802     0.0953    0.1341     0.2106
```

The function reports the p-values for the second to sixth dimension, and for a range of values. Notice that all the p-values are greater than 0.05, so we can feel confident in not rejecting the null hypothesis. The data appear to be independently and identically distributed.

Let's apply the test to the daily closing first difference of the DAX stock market index using data from 1991-1998. This data is contained in the dataframe EuStockMarkets:

```
> DAX<-EuStockMarkets[,1]

> diff_DAX = diff(DAX,1)
```

```
> bds.test(diff_DAX, m=6)

        BDS Test

data:  diff_DAX

Embedding dimension =  2 3 4 5 6

Epsilon for close points =  16.2486 32.4973 48.7459 64.9945

Standard Normal =

     [ 16.2486 ] [ 32.4973 ] [ 48.7459 ] [ 64.9945 ]

[ 2 ]     12.5683     14.7624     13.8202     10.9687

[ 3 ]     16.9602     19.6209     18.9546     15.7706

[ 4 ]     21.0833     23.3879     22.3583     18.6625

[ 5 ]     25.8893     26.8156     24.9345     20.6816

[ 6 ]     31.8667     30.3848     27.1510     22.3024

p-value =

     [ 16.2486 ] [ 32.4973 ] [ 48.7459 ] [ 64.9945 ]

[ 2 ]         0         0         0         0

[ 3 ]         0         0         0         0

[ 4 ]         0         0         0         0

[ 5 ]         0         0         0         0

[ 6 ]         0         0         0         0
```

Notice, in this case, all the p-values are reported as zero, and we strongly reject the null hypothesis that the daily difference in the DAX index is jointly independent and identically distributed.

References

Brock, W. A., D. Hsieh, and B. LeBaron (1991): Nonlinear Dynamics, Chaos and Instability: Statistical Theory and Economic Evidence. MIT Press Cambridge, Massachusetts.

Khan, D. A., Navet, N., Bavoux, B., & Migge, J. (2009, September). Aperiodi traffic in response time analyses with adjustable safety level. In Emerging Technologies & Factory Automation, 2009. ETFA 2009. IEEE Conference or

(pp. 1-9). IEEE.

Özer, G., & Ertokatlı, C. (2010). Chaotic processes of common stock index returns: An empirical examination on Istanbul Stock Exchange (ISE) market African Journal of Business Management, 4(6), 1140-1148.

Pal, S. (2011). Productivity Differential and Bilateral Real Exchange Rate between India and US. Journal of Quantitative Economics, 9(1), 146-155.

Back to Table of Contents

TEST 31 WALD-WOLFOWITZ TWO SAMPLE RUN TEST

Question the test addresses

Do two random samples come from populations having the same distribution?

When to use the test?

The test is used to detect differences such as averages or spread between two populations. A run is defined as a series of increasing values or a series of decreasing values. The number of increasing, or decreasing, values is the length of the run.

Practical Applications

Interplanetary dust: The distribution of interplanetary dust is investigated by Davis et al (2012). Observations on dust impacts over the period 1 April 2007 to 6 February 2010 were obtained by STERO ahead and STERO behin spacecraft. The researchers use a runs test to assess whether the distributions of dust observed by the STERO ahead and STERO behin spacecraft are random in nature. The runs test indicated the observed dust distributions are not statistically distinct from a random distribution (p-value greater than 0.05).

Down syndrome: Ramano et al (2002) compared a range of clinical and biochemical variables and zinc levels in 120 Down syndrome patients. Two groups, one with normal zinc levels, and the second with low zinc levels, were compared in the analysis. The Wald-Wolfowitz runs test for randomness was used to assess whether there were significant differences between the two samples. The authors report a p-value of less than 0.02, and therefore reject the null hypothesis of no difference between the two groups.

Stellar luminosity: Whether binary stars lead to significant bias in photometric parallax-based measurements of the stellar luminosity function is investigated by Reid and Gizis (1997). The researchers compile a catalogue of photometry and binary statistics for stars known to be north of minus thirty degrees declination and within eight parsecs of the Sun. As part of their analysis, the researchers investigate whether binary stars amongst the field M-dwarfs have semi-major axis and mass-ratio distributions consistent with those of the nearby stars. Two samples of binaries are compared using a runs test. The null hypothesis is not rejected (p-value > 0.05) and the researchers conclude there is no statistical difference between either the overall binary fraction or the mass-ratio

distributions of the two samples.

How to calculate in R

The function runs.test{tseries} is used to perform this test. It takes the form:

runs.test(combined_sample, alternative = "two.sided"), Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Example: two sided test using runs.test

Suppose you have collected data as follows:

First sample {3.18, 3.28, 3.92, 3.6, 3.0, 3.45, 3.74}

Second sample {3.55, 2.76, 2.13, 2.48, 3.67, 3.0}

Denoting sample the first sample by 1 and the second sample by 0, the data are combined and ordered as follows:

value	2.13	2.48	2.76	3.0	3.0	3.18	3.28	3.45	3.55	3.6	3.67	3.74	3.92
sample	0	0	0	0	1	1	1	1	0	1	0	1	1

Now enter the following combined data as follows:

```
> combined_sample<-factor(c(0,0,0,0,1,1,1,1,0,1,0,1,1))

> runs.test(combined_sample)

        Runs Test

data:  combined_sample

Standard Normal = -0.8523, p-value = 0.3941

alternative hypothesis: two.sided
```

Since the p-value is greater than 0.05, do not reject the null hypothesis of randomness.

Example: one sided test using runs.test

Enter the following data

```
> combined_sample<-factor(c(0,0,0,0,1,1,1,1,0,1,0,1,1))

> runs.test(combined_sample,alternative ="less")
```

Runs Test

data: combined_sample

Standard Normal = -0.8523, p-value = 0.197

alternative hypothesis: less

Since the p-value is greater than 0.05, do not reject the null hypothesis.

References

Davis, C. J. et al .(2012). Predicting the arrival of high-speed solar wind streams at Earth using the STEREO Heliospheric Imagers.Space Weather the international journal of research and applications, vol. 10, S02003, 18 PP.. doi:10.1029/2011SW000737.

Reid, I. N., & Gizis, J. E. (1997). Loss-Mass Binaries and the Stella Luminosity Function. The Astronomical Journal, 113, 2246

Romano C, Pettinato R, Ragusa L, et al. (2002). Is there a relationship between zinc and the peculiar comorbidities of Down syndrome? Downs Syndr Res Pract 2002;8:25–8.

Back to Table of Contents

TEST 32 MOOD'S TEST

Question the test addresses

Do two independent samples come from the same distribution?

When to use the test?

To test the null hypothesis that two population distribution functions corresponding to the two samples are identical against the alternative hypothesis that they come from distributions that have the same median and shape but different dispersions (scale). It is assumed the data are collected from two independent random samples. The underlying population distributions are continuous and the data are measured on at least an ordinal scale.

Practical Applications

Genetics: Thirty nine (ten female and twenty nine male) disease-free adults were recruited into a marked impairment of Fc receptor-dependent mononuclear phagocyte system study by Kimberly et al (1983). Participants were divided into four groups – those individuals with an HLA haplotype containing either DR2, MT1, or B8/ DR3 and those without such haplotypes (other).The researchers report the DR2 group is significantly more dispersed than both the non-DR2 groups, B8/DR3 and MTI (p–value using Mood's test of dispersion < 0.01 for all comparisons). They also find DR2 group is significantly more dispersed than the "other" subgroup (p-value using Mood's test of dispersion < 0.04).

Cattle prices: Basmann (2003), as part of a wider study into the legal case "Paul F. Engler and Cactus Feeders, Inc., v. Oprah Winfrey et al", investigate whether first differences of future cattle prices are statistically independent with respect to their temporal order. Chicago mercantile exchange June futures price from April 1, 1996 to June 28, 1996, were first differenced and then divided into two samples. For Mood's test of equality of dispersions the p-value was less than 0.01 and the null hypothesis is rejected.

Patient predictions: Boos (1985) report on the percentages of correct predictions of patient disorders by trainees at veteran hospitals and undergraduate psychology majors. A two-sample comparison of trainees and undergraduates predictions using Mood's test was not significant (p-value > 0.5). The authors observe the analysis indicate no scale differences between the two samples.

How to calculate in R

The functions mood.test{stats}and scaleTest{fBasics} can be used to perform this test.

Example: two sided test using mood.test

The function takes the form mood.test (sample_1, sample_2, alternative = "two.sided"). Note to conduct a one sided test set alternative = "less" or alternative ="greater".

Enter the following data

> sample_1 <-c(3.84,2.6,1.19,2)

> sample_2<-c(3.97,2.5,2.7,3.36,2.3)

> mood.test (sample_1, sample_2, alternative ="two.sided")

Mood two-sample test of scale

data: sample_1 and sample_2

Z = 0.7928, p-value = 0.4279

alternative hypothesis: two.sided

Since the p-value is greater than 0.05, do not reject the null hypothesis of randomness.

Example: two sided test using scaleTest

The function scaleTest takes the form scaleTest(sample_1, sample_2, method = "mood")

Enter the following data

> sample_1 <-c(3.84,2.6,1.19,2)

> sample_2<-c(3.97,2.5,2.7,3.36,2.3)

> scaleTest(sample_1,sample_2,method = "mood")

Title:

Mood Two-Sample Test of Scale

Test Results:

STATISTIC:

Z: 0.7928

P VALUE:

Alternative Two-Sided: 0.4279

Alternative Less: 0.7861

Alternative Greater: 0.2139

The function reports the two sided p-value equal to 0.4279. It is greater than 0.05, do not reject the null hypothesis.

References

Basmann, R.L. (2003). Statistical outlier .analysis in litigation support: the case of Paul F. Engler and Cactus Feeders, Inc., v. Oprah Winfrey et al. Journal of Econometrics 113, 159-200.

Boos, Dennis D. (1985)."Rank analysis of k samples." Institute of Statistics Mimeograph Series No. 1670.

Kimberly, R.P., A. Gibofsky, J.E. Salmon, and M. Fotino. 1983. Impaired Fc mediated mononuclear phagocyte system clearance in HLA-DR2 and MT1 positive healthy young adults. J. Exp. Med. 157:1698–1703.

Back to Table of Contents

TEST 33 F-TEST OF EQUALITY OF VARIANCES

Question the test addresses

Are the variances of two samples equal?

When to use the test?

This test is used to test the null hypothesis that two independent samples have the same variance. The test is sensitive to departures from normality.

Practical Applications

Rotational hip profile: Staheli (1985) studied 1,000 lower extremities of healthy children and adults in order to establish normal values for their rotational profile. As part of the study measurements of the rotation of the hip were made using both clinical methods (gravity goniometer, protractor) and photographic techniques (camera located distally and directed cephalad in line with the axes of the thighs). An F test of equality of variances of the photo graphic and clinical measurements showed no significant differences ($p > 0.05$).

Prudent sperm use: Sperm use during egg fertilization of the leaf-cutter ant Atta colombica was investigated by den Boer et al (2009). They find that queens are able to fertilize close to 100 per cent of the eggs and that the average sperm use per egg is very low, but increases with queen age. Variation in median sperm use among founding queens was observed to be much higher than among established queens (F-test of equality of variances p-value < 0.001).

Metal hip replacement: Georgiou et al (2012) examined the effect of head diameter and neck geometry on migration at two years of follow-up in a case series of 116 patients (125 hips), who have undergone primary metal-on-metal total hip arthroplasty The determination of bone and prosthesis landmarks were assessed by hand by two observers. The researchers assess inter-observer variability in measurements using the F-test of equality of variances. They comment that inter-observer variability was negligible and the measurements of the two observers were highly correlated (Pearson correlation coefficient = 0.970) with equal variances (F-test of equality of variances p-value =0.641).

How to calculate in R

The function var.test{stats}can be used to perform this test. It takes the form:

var.test(x, y, ratio = 1, alternative = "two.sided",conf.level = 0.95).Note, x and y are the data samples, ratio is the hypothesized ratio of variance. For a one sided test set alternative =”less” or alternative =”greater”.

Example: testing the weight of rolled oats:

The following have been collected on the weight of packets of rolled oats filled by two different machines.

machine.1=c(10.8,11.0,10.4,10.3,11.3)

machine.2=c (10.8,10.6,11,10.9,10.9,10.7,1.8)

The variance ratio test can be carried out as follows:

> var.test(machine.1,machine.2, ratio =1, alternative="two.sided",conf.level = 0.95)

F test to compare two variances

data: machine.1 and machine.2

F = 0.0149, num df = 4, denom df = 6, p-value = 0.001142

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.002388292 0.136784965

sample estimates:

ratio of variances

0.01487228

Since the p-value is less than 0.05, reject the null hypothesis. The variances are significantly different from each other.

References

Den Boer, S. P., Baer, B., Dreier, S., Aron, S., Nash, D. R., & Boomsma, J. J (2009). Prudent sperm use by leaf-cutter ant queens. Proceedings of the Royal Society B: Biological Sciences, 276(1675), 3945-3953.

Georgiou, C. S., Evangelou, K. G., Theodorou, E. G., Provatidis, C. G., & Megas, P. D. (2012). Does Choice of Head Size and Neck Geometry Affec Stem Migration in Modular Large-Diameter Metal-on-Metal Total Hip Arthroplasty? A Preliminary Analysis. The open orthopaedics journal, 6 593.

Staheli, L. T., Corbett, M., Wyss, C., & King, H. (1985). Lower-extremity rotational problems in children. J Bone Joint Surg [Am], 67(1), 39-47.

Back to Table of Contents

TEST 34 PITMAN-MORGAN TEST

Question the test addresses

Are the variances of two correlated samples equal?

When to use the test?

To test for equality of the variances of the marginal distributions of two correlated variables. The test involves testing the correlation between the sum and difference of the two responses, with zero correlation corresponding to equality of the two variances. The test is known to be optimal for testing equality of the variances of components of a bivariate normal distribution. It is however, sensitive to departures from normality.

Practical Applications

Gallbladder ejection fraction: Ziessman et al (2001) determine normal gallbladder ejection fraction (GBEF) values for two sincalide (cholecystokinin) infusion dose rates, 0.01 µg per kilogram of body weight infused for 3 minutes and 0.01 µg/kg infused for 60 minutes. Twenty participants were recruited and GBEFs were calculated for the 3-minute infusion and for each 15-minute interval for the 60-minute infusion. The researchers test whether inter-subject variability of GBEF was less for the 60-minute infusion than it was for the 3-minute infusion. With the 3-minute infusion method, the GBEF was significantly more variable than GBEFs at 45 or 60 minutes with the 60-minute infusion (Pitman-Morgan test p-values =0 .013 and 0.022 respectively).

Pig weight: Jones et al (2009) investigate the effect of feed withdrawal on live weight pigs. Three different age groups (“weaners”, “growers” and “finishers”) were split randomly into control and treatment groups. The pigs in each group were weighed in the evening and again the following morning after a time lapse of 11 hours for the weaners and 17 hours for the other groups. Those in the control group were fed normally, but food was withheld from the treatment group. The Pitman-Morgan test was used to assess the variability between live weight in the evening and live weight the following morning for control and treatment groups. For weaners in the control group (p-value = 0.73, n = 66), for weaners in the withheld group (p-value = 0.0008, n = 52). For growers in the control group (p-value = 0.9485, n = 52), for growers in the withheld group (p-value = 0.0014, n = 51). For finishers in the control group (p-value = 0.7216, n = 50), for finishers in the withheld group (p-value = 0.0484, n = 52). The null hypothesis of no difference between the variability of live weight in the

evening and live weight the following morning was rejected for the food withheld group of pigs.

Inter-rater reliability for job seekers: Baugher et al (2011) investigates Inter-rater reliability for candidates seeking in-line promotions in a State Agency to financial analyst (FA) and upper management (UM) positions. The sample consisted of 64 candidates seeking positions for a FA post, and 35 candidates seeking promotion to upper management. Three rating approaches were analyzed: one rater, two raters, and two raters with hybrid consensus. The Pittman-Morgan t-test for comparing correlated variances showed that the variance from the three approaches did not differ significantly for the UM position (p-value > 0.05).

How to calculate in R

The function pitman.morgan.test{PairedData}can be used to perform this test. It takes the form pitman.morgan.test(x, y , alternative = "two.sided" or "less" or "greater")omega = 1, conf.level = 0.95). Note, x and y are the data samples, omega is the hypothesized ratio of variance. For a one sided test set alternative =”less” or alternative =”greater”.

Example: testing the weight of rolled oats

The following data have been collected on the weight of packets of rolled oats filled by the same machine during the morning and the evening shift. Are the variances of the two correlated samples equal?

machine.am=c(10.8,11.0,10.4,10.3,11.3,10.2,11.1)

machine.pm=c (10.8,10.6,11,10.9,10.9,10.7,1.8)

The test can be carried out as follows:

> pitman.morgan.test(machine.am,machine.pm)

Paired Pitman-Morgan test

data: machine.1 and machine.2

t = -9.4346, df = 5, p-value = 0.0002258

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.002516232 0.101298431

sample estimates:

variance of x variance of y

0.1857143 11.6323810

Since the p-value is less than 0.05, reject the null hypothesis. The variances are significantly different.

References

Baugher, D., Weisbord, E., & Eisner, A. (2011) .Evaluating training and experience: do multiple raters or consensus make a difference? Proceedings of ASBBS. Volume 18 Number 1,page 516-528.

Jones, G., Noble, A. D. L., Schauer, B., & Cogger, N. (2009). Measuring the Attenuation in a Subject-specific Random Effect with Paired Data. Journal of Data Science, 7, 179-188.

Ziessman, H. A., Muenz, L. R., Agarwal, A. K., & ZaZa, A. A. (2001). Norma Values for Sincalide Cholescintigraphy: Comparison of Two Methods1. Radiology, 221(2), 404-410.

Back to Table of Contents

TEST 35 ANSARI-BRADLEY TEST

Question the test addresses

Do two independent samples come from the same distribution?

When to use the test?

To test the null hypothesis that the two population distribution functions corresponding to the two samples are identical against the alternative hypothesis that they come from distributions that have the same median and shape but different dispersions (scale). It is assumed the data are collected from two independent random samples. The underlying population distributions are continuous and the data are measured on at least an ordinal scale.

Practical Applications

Genomics: Wang et al (2012) explore the genes associated with MUC5AC expression in small airway epithelium of smokers and non-smokers. For the samples obtained from non-smokers the Ansari-Bradley test was used to assess the variation in the degree of MUC5AC gene expression compared to the housekeeping genes ACTB, GAPDH, B2M, RPLPO and PPIA. In all cas the p-value was less than 0.01 and the null hypothesis is rejected. A similar finding was reported for smokers.

Climate change: The impact of climate change in winter wheat and grain maize production in two study regions of the Swiss Plateau, which differed in their climate and soil types, are studied by Lehmann et al (2012). Using yield distributions of 25 weather years and a bio-economic model the researchers used the Ansari-Bradley test to assess changes in crop yield variability. For grain maize cultivated in the Greifensee-Watershed region a significant change (p-value < 0.05) in the variability of yield between the baseline and their regional climate model scenario was reported.

Emotional speech: Pribil, Pribilova and Durackova (2012) investigate the effect of the fixed and removable orthodontic appliances on spectral properties of emotional speech. The researchers apply the Ansari-Bradley test to sets of spectrograms with different configurations of orthodontic appliances in neutral and emotional styles. For a neutral style the Ansari-Bradley test for the comparison “without orthodontic appliance” to “the lower fixed orthodontic brackets” returned a p-value less than 0.05.

How to calculate in R

The functions ansari.test{stats}, ansari.exact{exactRankTests} and scaleTest{fBasics} can be used to perform this test.

Example: two sided test using ansari.test

The function ansari.test takes the form ansari.test(sample_1, sample_2, alternative = "two.sided"). Note to conduct a one sided test set alternative = "less" or alternative ="greater".

Enter the following data

```
> sample_1 <-c(3.84,2.6,1.19,2)

> sample_2<-c(3.97,2.5,2.7,3.36,2.3)

> ansari.test(sample_1, sample_2, alternative ="two.sided")

        Ansari-Bradley test

data:  sample_1 and sample_2

AB = 10, p-value = 0.7937

alternative hypothesis: true ratio of scales is not equal to 1
```

Since the p-value is greater than 0.05, do not reject the null hypothesis of randomness.

Example: two sided test using ansari.exact

The function ansari.exact takes the form ansari.exact(sample_1, sample_2, alternative = "two.sided"). Note to conduct a one sided test set alternative = "less" or alternative ="greater".

Enter the following data

```
> sample_1 <-c(3.84,2.6,1.19,2)

> sample_2<-c(3.97,2.5,2.7,3.36,2.3)

> ansari.exact(sample_1, sample_2, alternative ="two.sided")

        Ansari-Bradley test

data:  sample_1 and sample_2

AB = 10, p-value = 0.6587

alternative hypothesis: true ratio of scales is not equal to 1
```

Since the p-value is greater than 0.05, do not reject the null hypothesis.

Example: two sided test using scaleTest

The function scaleTest takes the form scaleTest(sample_1, sample_2, method = "ansari")

Enter the following data

```
> sample_1 <-c(3.84,2.6,1.19,2)

> sample_2<-c(3.97,2.5,2.7,3.36,2.3)

> scaleTest(sample_1,sample_2,method = "ansari")
```

Title:

Ansari-Bradley Test for Scale

Test Results:

STATISTIC:

AB: 10

P VALUE:

Alternative Two-Sided : 0.593

Alternative Two-Sided | Exact: 0.7937

Alternative Less : 0.7035

Alternative Less | Exact: 0.7778

Alternative Greater : 0.2965

Alternative Greater | Exact: 0.3968

The function reports the exact two sided p-value equal to 0.7937. It is greater than 0.05, do not reject the null hypothesis.

References

Pribil,J; Pribilova, A; and Durackova, D. (2012).An experiment with spectra analysis of emotional speech affected by orthodontic appliances. Journal of Electrical Engineering, Vol. 63, No. 5, 2012, 296–302.

Lehmann, Niklaus et al. (2012). Adapting Towards Climate Change: A Bioeconomic Analysis of Winterwheat and Grain Maize. Internationa Association of Agricultural Economists Conference, August 18-24, 2012, Foz do Iguaçu, Brazil.

Wang, G et al. (2012). Genes associated with MUC5AC expression in smal airway epithelium of human smokers and non-smokers. BMC Medical Genomics, 5:21

Back to Table of Contents

TEST 36 BARTLETT TEST FOR HOMOGENEITY OF VARIANCE

Question the test addresses

Do k samples come from populations with equal variances?

When to use the test?

This test is used to test the null hypothesis that multiple independent samples have the same variance. The test, is sensitive to departures from normality.

Practical Applications

Human mercury accumulation: The traditional Arctic diet involves the consumption of a high intake of mercury primarily from marine mammals. Johansen et al (2007) investigate whether the mercury is accumulated in humans. Autopsy samples of liver, kidney and spleen from adult ethnic Greenlanders (57 men, 45 women) who died between 1990 and 1994 was analyzed. Liver, kidney and spleen samples from randomly selected case subjects were analyzed for total mercury and methylmercury. Liver samples were analyzed for selenium. Barlett test was used to test for homogeneity of variance between samples of the sexes. The researchers report in no cases did the variance differ between sexes (p-value > 0.05).

Electromagnetic wave propagation: Esperante et al (2012) study the behavior electromagnetic waves radiated from an indoor wireless fidelity access point with two different antenna positions (vertical and horizontal). Measurements of signal strength were taken for vertical and horizontal antenna positions at 3 meter increments, starting at 3 meters away from the access point, ending at 30 meters. The researchers test for equal variance for all the measurements and distances for the two antenna positions. The Barlett test p-value was greater than 0.05 and null hypothesis was not rejected.

Transcriptomic analysis of autistic brain: Voineagu et al (2011) investigate differences in transcriptome organization between the autistic and normal brain using gene co-expression network analysis. Ribonucleic acid samples from the cortex for 13 autism and 13 control cases were obtained. For each of the 510 genes that were differentially expressed the researchers compared the variance of autism and control expression. The homogeneity of variance was assessed using the Barlett test. A total of fifty one genes showed a significant difference in variance (p-value < 0.05). This result is

consistent with their overall finding of significant differences in transcriptome organization between the autistic and normal brain.

How to calculate in R

The function bartlett.test{stats}can be used to perform this test.

Example: two sided test using bartlett.test

Suppose you have collected the following data on five samples.

The first column is sample A, second sample B, third sample C, forth Sample D and the final column sample E

250	100	250	340	250
260	330	230	270	240
230	280	220	300	270
270	360	260	320	290

Enter the data as follows:

```
> count_data<-
c(250,260,230,270,310,330,280,360,250,230,220,260,340,270,300,320,250,240
> sample<-
c("A","A","A","A","B","B","B","B","C","C","C","C","D","D","D","D","E","E","E","
> data<-data.frame((list(count= count_data, sample=sample)))
> bartlett.test(data$count, data$sample)

        Bartlett test of homogeneity of variances

data:  data$count and data$sample
Bartlett's K-squared = 1.8709, df = 4, p-value = 0.7595
```

Since the p-value is greater than 0.05, do not reject the null hypothesis.

Example: alternative approach to conducting a two sided test

Using the above data we can also use a slightly different specification to conduct the test.

```
> bartlett.test(count ~ sample, data = data)
```

Bartlett test of homogeneity of variances

data: count by sample

Bartlett's K-squared = 1.8709, df = 4, p-value = 0.7595

We obtain the same p-value as the previous example, and do not reject the null hypothesis.

References

Johansen, P.,Mulvad,G.,Pedersen,H.S.,Hansen,J.C.,Riget,F.,(2007).Humar accumulation ofmercuryinGreenland.Sci.TotalEnviron.377,173–178.

Esperante, P. G., Cymrot, R., Garcia, P. A., Vieira, M. S., & Perotoni, M (2012, April). Analysis of electromagnetic wave propagation in indoor environments. In Proceedings of the 11th international conference on Telecommunications and Informatics, Proceedings of the 11th international conference on Signal Processing (pp. 101-105). World Scientific and Engineering Academy and Society (WSEAS).

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., ... & Geschwind, D. H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature, 474(7351), 380-384.

Back to Table of Contents

TEST 37 FLIGNER-KILLEEN TEST

Question the test addresses

Do k samples come from populations with equal variances?

When to use the test?

This test is used to test the null hypothesis that multiple independent samples have the same variance. The test, is robust to departures from normality.

Practical Applications

Human spinal cord injury: Behrman (2012) develop a neuromuscular recovery scale for classification of functional motor recovery after spinal cord injury. Ninety five individuals with spinal injury were recruited into the study. At enrollment participants were allocated into one of three groups based on their neuromuscular recovery scale. Each participant took part in intensive loco-motor training. The Fligner-Killeen test was used to investigate the variability in outcome measures (Berg balance scale, six-minute walk test, and ten-meter walk test) of each group. The authors report a p-value <0.01 for all measures. They conclude their neuromuscular recovery scale classification is able to discriminate patients with respect to functional performance.

Pathogen load in plants: The incidence of fungal pathogens in dioecious versus hermaphroditic plant species was investigated by Williams, Antonovics and Rolff (2011). One hundred and twenty eight pairs, in thirty two families of flowering plants, were studied. To test for differences in variation of pathogen diversity between hermaphroditic and dioecious species, a Fligner-Killeen test was used. The researchers observe the variances of pathogen load tended to be greater in dioecious species, although the difference was not significant (Fligner–Killeen p-value = 0.0541).

Sea trout growth: Marco-Rius et al (2012) used scale analysis to reconstruct growth trajectories of migratory sea trout from six neighboring populations in Spain. The researchers compared the size individuals attained in freshwater with their subsequent growth at sea. The coefficient of variation (CV) was used to examine how much body size varied across populations and life stages. The researchers used the Fligner-Killen test to compare differences in variation of body size among stages of development and between rivers. Individual variation in growth increased significantly over time (Fligner-Killen Test p-value < 0.01). The CV on body size, calculated for

returning adults that had spent two winters in freshwater and one winter at sea, varied significantly among life stages (Fligner-Killen p-value <0.01). The researchers also find populations from different locations differed significantly in CV for body size during the first winter in freshwater (Fligner-Killen p-value = 0.013).

How to calculate in R

The function fligner.test{stats}can be used to perform this test.

Example: two sided test using fligner.test

Suppose you have collected the following data on five samples.

The first column is sample A, second sample B, third sample C, forth Sample D and the final column sample E

250	100	250	340	250
260	330	230	270	240
230	280	220	300	270
270	360	260	320	290

Enter the following data as follows:

```
> count_data<-
c(250,260,230,270,310,330,280,360,250,230,220,260,340,270,300,320,250,24(

> sample<-
c("A","A","A","A","B","B","B","B","C","C","C","C","D","D","D","D","E","E","E","
```

```
> data<-data.frame((list(count= count_data, sample=sample)))

> fligner.test (data$count, data$sample)
```

Fligner-Killeen test of homogeneity of variances

data: data$count and data$sample

Fligner-Killeen:med chi-squared = 2.8973, df = 4, p-value = 0.5752

Since the p-value is greater than 0.05, do not reject the null hypothesis.

Example: alternative approach to conducting a two sided test

Using the above data we can also use a slightly different specification to conduct the test.

```
> fligner.test (count ~ sample, data = data)

Fligner-Killeen test of homogeneity of variances

data:  count by sample

Fligner-Killeen:med chi-squared = 2.8973, df = 4, p-value = 0.5752
```

We obtain the same p-value as the previous example, and do not reject the null hypothesis.

References

Behrman, A. L., Ardolino, E., VanHiel, L. R., Kern, M., Atkinson, D., Lorenz, D J., & Harkema, S. J. (2012). Assessment of functional improvement without compensation reduces variability of outcome measures after human spinal cord injury. Archives of Physical Medicine and Rehabilitation, 93(9), 1518-1529.

Marco-Rius, F., Caballero, P., Morán, P., & de Leaniz, C. G. (2012). And the Last Shall Be First: Heterochrony and Compensatory Marine Growth in Se Trout (Salmo trutta). PloS one, 7(10), e45528.

Williams, A., Antonovics, J., & Rolff, J. (2011). Dioecy, hermaphrodites and pathogen load in plants. Oikos, 120(5), 657-660.

Back to Table of Contents

TEST 38 LEVENE'S TEST OF EQUALITY OF VARIANCE

Question the test addresses

Do k samples come from populations with equal variances?

When to use the test?

To test the null hypothesis that multiple independent samples have the same variance. The test is more robust to departures from normality than Bartlet's test for homogeneity of variances.

Practical Applications

Reproductive success: Brown et al (2009) examine data on mating behavior and reproductive success in current and historic human populations. As part of their study the researchers investigate the lifetime reproductive success of monogamous Pitcairn Islanders (145 males and 127 females). The results indicated male and female variances are not significantly different (Levene's test p-value >0.05).

Delinquents and mental health: Timmons-Mitchel et al (1997) study the prevalence of mental disorder in a juvenile justice population. A total of 173 delinquents, (121 males and 52 females) were recruited at random from a male institution and female institution in the State of Ohio. A random sub-sample of fifty (25 male, 25 female) was subject to a battery of tests including clinician-rated and self-report measures. The researchers analyses employed independent samples t-tests. Levene's test was used to determine whether to use an equal or unequal variances estimate of the t-test. In cases where the Levene's test was significant at the 0.05 level, the unequal variance estimate of the t test was selected.

Sexual conflict in insects: Arnqvist et al (2000) assess the general importance of post mating sexual conflict for the rate of speciation, by comparing extant species richness in pairs of related clades of insects differing in the opportunity for post mating sexual conflict. The researchers identify 25 phylogenetic contrasts, representing five different orders, all of which were independent in the sense that no clade was represented in more than one contrast. The researchers find neither the variance nor the magnitude of species richness depended significantly on whether the clades in a contrast were sister groups or more distantly related (Levene's test p-value = 0.354 or whether the contrast involved a within- or a between-family comparison (Levene's test p-value = 0.469).

How to calculate in R

The function leveneTest{outliers}can be used to perform this test.

Example: two sided test using leveneTest

Suppose you have collected the following data on five samples.

The first column is sample A, second sample B, third sample C, forth Sample D and the final column sample E

250	100	250	340	250
260	330	230	270	240
230	280	220	300	270
270	360	260	320	290

Enter the following data as follows:

```
> count_data<-
c(250,260,230,270,310,330,280,360,250,230,220,260,340,270,300,320,250,24(

> sample<-
c("A","A","A","A","B","B","B","B","C","C","C","C","D","D","D","D","E","E","E","

> data<-data.frame((list(count= count_data, sample=sample)))

> leveneTest (data$count, data$sample)

Levene's Test for Homogeneity of Variance (center = median)

      Df F value Pr(>F)

group  4  0.7247 0.5886
```

Since the p-value is greater than 0.05, do not reject the null hypothesis.

Example: alternative approach to conducting a two sided test

Using the above data we can also use a slightly different specification to conduct the test.

```
> leveneTest (count ~ sample, data = data)

Levene's Test for Homogeneity of Variance (center = median)

      Df F value Pr(>F)

group  4  0.7247 0.5886
```

We obtain the same p-value as the previous example, and do not reject the

null hypothesis.

References

Arnqvist, G., Edvardsson, M., Friberg, U., & Nilsson, T. (2000). Sexual conflict promotes speciation in insects. Proceedings of the National Academy of Sciences, 97(19), 10460-10464.

Brown, G. R., Laland, K. N., & Mulder, M. B. (2009). Bateman's principle and human sex roles. Trends in ecology & evolution, 24(6), 297-304.

Timmons-Mitchel, J., Brown, C., Schulz, S. C., Webster, S. E., Underwood, l A., & Semple, W. E. (1997). Comparing the mental health needs of female and male incarcerated juvenile delinquents. Behavioral Sciences and the Law, 15, 195-202.

Back to Table of Contents

TEST 39 COCHRAN C TEST FOR INLYING OR OUTLYING VARIANCE

Question the test addresses

Do k samples come from populations with equal variances?

When to use the test?

To test the null hypothesis of equality of variances against the alternative that one variance is larger (or smaller) than the rest. The sample data on each factor should all be equal length. It is assumed that each individual data series is normally distributed. The statistic compares the largest (or smallest) sample variance with the sum of all variances to determine whether or not an outlier exists.

Practical Applications

Fecal bacteria along the coast: Total and fecal coliforms over along 50 km of the Marche coasts (Adriatic Sea) were analyzed by Luna et al (2010). Samples were collected at depths ranging from 2 to 5 meters. Total and fecal coliforms (FC) were counted by culture-based methods. Differences in the microbiological variables (total prokaryotes, total and fecal coliforms) between different areas and sampling depths were investigated. In total seven, sampling areas and two depths constituted a major part of the sample. Cochran's C test was used to test for homogeneity of variance with significance level set to a very conservative 0.001. Where samples failed Cochran's C test (p-value <0.001) the data were log transformed in an attempt to induce homogeneity.

Micropredators of the sea urchin: Bonaviri et al (2012) identified several potential invertebrate micropredators of settlers of the sea urchin (Paracentrotus lividus) and measured their predation activity. For the predator Hermit crab (Calcinus tubularis), Cochran's C test was used to assess homogeneity of variances of per capita predation rates on sea urchins given Hermit crab size (small and large) and the Urchin size (small and large). No significant differences were identified (p-value >0.05). For the predator Shrimp (Alpheus dentipes), Cochran's C test was also used to assess homogeneity of variances of per capita predation rates on sea urchins given Hermit crab size (small and large) and the Urchin size (small and large). No significant differences were identified (p-value >0.05).

Urge to cough: Lavorini et al (2010) study how exercise and voluntary isocapnic hyperpnea affect the sensitivity of the cough reflex and the

sensation of a urge to cough evoked by ultrasonically nebulized distilled water inhalation in healthy subjects. Twelve nonsmoker participants were recruited onto the study and induced to cough via the nebulizer output. Experiments consisted of adjusting the range of nebulizer outputs ranged from 30% to 100%. The researchers report variances calculated for each set of experiments were homogeneous (Cochran's C-test p-value = 0.49).

How to calculate in R

The function cochran.test{outliers} or C.test{GAD} can be used to perform this test.

Example: Testing for outlying variance

Suppose you have collected the following data on five samples.

The first column is sample A, second sample B, third sample C, forth Sample D and the final column sample E

250	100	250	340	250
260	330	230	270	240
230	280	220	300	270
270	360	260	320	290

Enter the following data as follows:

```
> count_data<-
c(250,260,230,270,310,330,280,360,250,230,220,260,340,270,300,320,250,24(
> sample<-
c("A","A","A","A","B","B","B","B","C","C","C","C","D","D","D","D","E","E","E","
> data<-data.frame((list(count= count_data, sample=sample)))
```

To carry out a test of for the largest variance enter:

```
> cochran.test(count~sample,data,inlying=FALSE)

        Cochran test for outlying variance

data:  count ~ sample

C = 0.3607, df = 4, k = 5, p-value = 0.6704

alternative hypothesis: Group B has outlying variance

sample estimates:
```

```
     A        B        C        D        E
 291.6667 1133.3333  333.3333  891.6667  491.6667
```

The function identifies B as the largest value against which to conduct the test. Since the p-value is greater than 0.05 we cannot reject the null hypothesis of equality of variances. As an alternative you can also use C.test, to do so enter

```
> C.test(lm(count~sample,data =data))

        Cochran test of homogeneity of variances

data:  lm(count ~ sample, data = data)

C = 0.3607, n = 4, k = 5, p-value = 0.6704

alternative hypothesis: Group B has outlying variance

sample estimates:

     A        B        C        D        E
 291.6667 1133.3333  333.3333  891.6667  491.6667
```

Again, the p-value is greater than 0.05, so we cannot reject the null hypothesis.

Example: Testing for inlying variance

We can also test for the smallest variance by entering:

```
> cochran.test(count~sample,data,inlying=TRUE)

        Cochran test for inlying variance

data:  count ~ sample

C = 0.0928, df = 4, k = 5, p-value < 2.2e-16

alternative hypothesis: Group A has inlying variance

sample estimates:

     A        B        C        D        E
 291.6667 1133.3333  333.3333  891.6667  491.6667
```

In this case the smallest variance is A, since the p-value is less than 0.05 reject the null hypothesis of equality of variances.

References

Bonaviri, C., Gianguzza, P., Pipitone, C., & Hereu, B. (2012). Micropredatior on sea urchins as a potential stabilizing process for rocky reefs. Journal of Sea Research.

Lavorini, F., Fontana, G. A., Chellini, E., Magni, C., Duranti, R., & Widdicombe, J. (2010). Desensitization of the cough reflex by exercise and voluntary isocapnic hyperpnea. Journal of Applied Physiology, 108(5), 1061-1068.

Luna, G. M., Vignaroli, C., Rinaldi, C., Pusceddu, A., Nicoletti, L., Gabellin M., ... & Biavasco, F. (2010). Extraintestinal Escherichia coli carrying virulence genes in coastal marine sediments. Applied and environmental microbiology, 76(17), 5659-5668.

Back to Table of Contents

TEST 40 BROWN-FORSYTHE LEVENE-TYPE TEST

Question the test addresses

Do k samples come from populations with equal variances?

When to use the test?

To test the null hypothesis that multiple independent samples have the same variance. The test is more robust to departures from normality than Bartlet's test for homogeneity of variances.

Practical Applications

Snow density measurement: Conger and McClung (2009) use a randomized block design to measure variance, measurement errors and sampling error in snow density measurement using five common snow cutters. Data for analysis were collected during February and March 2006 in the Parks Canada Mount Fidelity Station in Glacier Park, British Columbia. In tota five snow layers were analyzed per cutter. Brown–Forsythe, Levene's, test was used to evaluate the assumption of equal homogeneity of variance. Only in layer 1 did the p-value (<0.05) suggest unequal variances. For all other layers, the null hypothesis could not be rejected.

Butterfly male mate preferences: Heliconius butterflies are well known for their brightly colored patterns which are used both as warnings and as mate recognition cues. Merrill et al (2011) investigated male mate preferences within a single polymorphic population as well as between three pairs of sister taxa in the melpomene-cydno clade of Heliconius. The mate preference data, representing different stages of divergence, allowed the researchers to compare diverging mate preferences across the continuum of Heliconius speciation. The researchers observed the extent of variance in preference among individual butterflies differed significantly among populations based on the Brown–Forsythe Levene-type test for equality of variances (p-value < 0.000001).

Touch based mobile interaction: A user evaluation was conducted by Hayes et al (2011) using a tablet type computer to present a target selection task within a map-based interface. Participants interacted with the mobile device while seated or while walking in an uncontrolled environment. There were 329 total target selections while walking and 299 in the while seated. The Brown-Forsythe Levene-type test was used to test for differences in the variance of the data between being seated and walking. The researchers find a significant difference between the seated and walking position (Brown-Forsythe Levene-type test p-value <0.01). They

also find a significant difference between male and female participants (Brown-Forsythe Levene-type test p-value <0.005).

How to calculate in R

The function levene.test{lawstat}can be used to perform this test. The function takes the form levene.test(count, sample.group, location="median", correction.method="zero.correction"). Note sample.group refers to the category or group label, count refers to the number of events observed per individual or participant.

Example:

Suppose you have collected the following data on five samples.

The first column is sample A, second sample B, third sample C, forth Sample D and the final column sample E

250	100	250	340	250
260	330	230	270	240
230	280	220	300	270
270	360	260	320	290

Enter the following data as follows:

```
> count_data<-
c(250,260,230,270,310,330,280,360,250,230,220,260,340,270,300,320,250,240
> sample<-
c("A","A","A","A","B","B","B","B","C","C","C","C","D","D","D","D","E","E","E","
> data<-data.frame((list(count= count_data, sample=sample)))
> levene.test (data$count, data$sample, location="median",
correction.method="zero.correction")
        modified robust Brown-Forsythe Levene-type test based on the
absolute
    deviations from the median with modified structural zero removal
    method and correction factor
data:  data$count
Test Statistic = 2.2051, p-value = 0.1416
```

Since the p-value is greater than 0.05, do not reject the null hypothesis.

References

Conger, S. M., & McClung, D. M. (2009). Comparison of density cutters for snow profile observations. Journal of Glaciology, 55(189), 163-169.

Hayes, S. T., Hooten, E. R., & Adamsψ, J. (2011).A. Touch-based Target Selection for Mobile Interaction Technical Report HMT-11-01.

Merrill, R. M., Gompert, Z., Dembeck, L. M., Kronforst, M. R., McMillan, W O., & Jiggins, C. D. (2011). Mate preference across the speciation continuum in a clade of mimetic butterflies. Evolution, 65(5), 1489-1500.

Back to Table of Contents

TEST 41 MAUCHLY'S SPHERICITY TEST

Question the test addresses

Are the variances of the differences between all possible pairs of groups in a repeated measures analysis of variance equal?

When to use the test?

To investigate whether the variances of the differences between all combinations of related groups are equal. Sphericity can be likened to homogeneity of variances in a between-subjects analysis of variance study.

Practical Applications

Vocal expression of emotion: Patel et al (2011) analyzed short affect bursts (sustained/a/vowels), produced by 10 professional actors for five emotions, according to physiological variations in phonation. The researchers investigate using a repeated measures ANOVA design for each of 12 acoustic parameters. Mauchly's sphericity test was used to assess sphericity. The null hypothesis could not be rejected for eight of their twelve acoustic parameters. The remaining four parameters had p – values less than 0.05 – (equivalent sound level p-value = 0.001, jitter p-value =0.002, MFO p-value = 0.016, pulse amp p-value = 0.012).

Pupil diameter: Atchison et al (2011),using a repeated-measures ANOVA design, investigated the interaction between adapting field size and luminance on pupil diameter when cones alone or rods and cones were active. Six male and two female subjects were recruited into the study. The researchers observe pupil size show individual differences in mean diameter, but little variation in size with increasing stimulus area. Mauchly's test of sphericity for field size was not significant (p-value = 0.14).

Cuttlefish Vision: Whether cuttlefish use their vision to perform adaptive camouflage in dim light was investigated by Allen et al (2010). In one experiment the cuttlefish were presented with a small check substrate that was changed to either a large check or to a grey substrate at a light intensity of 0.003 lux (to simulate starlight). The distributions of mean granularity statistics for each light level were tested for sphericity using Mauchly's test of sphericity and were compared using a repeated measures ANOVA. The p-value was 0.44, and the authors conclude that sphericity was not violated for these data.

How to calculate in R

The function mauchly.test{stats} can be used to perform this test.

Example:

Enter the data and perform the test as follows:

>dependent_variable <- c (-5, -10, -5, 0, -3, -3, -5, -7, -2, 4, -1, -5, -4, -8, -4,-5,-12,-7)

>mlm <- matrix (dependent_variable, nrow = 6, byrow = TRUE)

>mauchly.test (lm (mlm ~ 1), X = ~1)

Mauchly's test of sphericity

Contrasts orthogonal to

~1

data: SSD matrix from lm(formula = mlm ~ 1)

W = 0.4545, p-value = 0.2065

Since the p-value is greater than 0.05 do not reject the null hypothesis of sphericity.

References

Allen, J. J., Mäthger, L. M., Buresch, K. C., Fetchko, T., Gardner, M., & Hanlon, R. T. (2010). Night vision by cuttlefish enables changeable camouflage. The Journal of Experimental Biology, 213(23), 3953-3960.

Atchison, D. A., Girgenti, C. C., Campbell, G. M., Dodds, J. P., Byrnes, T. M. & Zele, A. J. (2011). Influence of field size on pupil diameter under photopic and mesopic light levels. Clinical and Experimental Optometry, 94(6), 545-548.

Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: the role of voice production. Biological psychology, 87(1), 93-98.

Back to Table of Contents

TEST 42 BINOMINAL TEST

Question the test addresses

Do the proportion of individuals falling in each category differ from chance? Or does the proportion of individuals falling into each category differ from some pre-specified probabilities of falling into those categories?

When to use the test?

This test is used when you want to know if the observed frequencies of the two categories of a dichotomous variable differ from the frequencies that are expected under a binomial distribution with a specified probability parameter.

Practical Applications

Pneumococcal conjugate vaccine: Black et al (2000) study the efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children. Infants at 2,4,5 and 12 to 15 months of age were given the heptavalent pneumococcal conjugate vaccine or an alternative. Protective efficacy was estimated by calculating the ratio of the number of cases of invasive disease in the pneumococcal conjugate sample to the number of cases in the alternative vaccine group and subtracting this ratio from 1. Statistical validation of efficacy against invasive disease was evaluated with the binomial test of the null hypothesis that the vaccine has no efficacy for the seven serotypes. This was rejected with an overall two tailed p-value less than 0.05.

Lactate clearance and survival: Arnold et al (2009) investigate if early lactate clearance is associated with improved survival in emergency department patients with severe sepsis. They analyzed prospectively collected registries of consecutive emergency department patients (166 in total) diagnosed with severe sepsis at three urban hospitals. The difference in proportions of death between lactate clearance and non-clearance groups was assessed using the binomial test. The researchers observed mortality of 60% in the lactate non-clearance group versus 19% in the lactate clearance group (p-value <0.05).

Epidemiological case-control study: In an epidemiological case-control study of Vibrio vulnificus infections, Tacket et al (1984) study eleven cases with primary sepsis and eight cases with wound infection. The researchers report that among patients with primary sepsis, eight of eleven cases and one of eleven controls recalled having eaten raw oysters in the two weeks before the onset of illness (Binominal test p-value = 0.0078). For those

patients with wound infections, seven of eight and two of eight controls reported exposure of the skin to sea water or shellfish, respectively (Binominal test p-value = 0.0312).

How to calculate in R

The function binom.test{stats}can be used to perform this test. It takes the form binom.test(x, n, p = 0.5, alternative = "two.sided", conf.level = 0.95).

Note, x is the number of observed successes, n the number of trials and p is hypothesized probability of success. For a one sided test set alternative =”less” or alternative =”greater”.

Example: two sided test

```
> binom.test(x = 25, n= 30, p = 0.5, alternative = "two.sided", conf.level = 0.95)

        Exact binomial test

data:  25 and 30

number of successes = 25, number of trials = 30, p-value = 0.0003249

alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

 0.6527883 0.9435783

sample estimates:

probability of success

            0.8333333
```

Since the p-value is less than 0.05, reject the null hypothesis.

Example: one sided test

```
> binom.test(x = 25, n= 30, p = 0.5, alternative = "greater", conf.level = 0.95)

        Exact binomial test

data:  25 and 30

number of successes = 25, number of trials = 30, p-value = 0.0001625

alternative hypothesis: true probability of success is greater than 0.5

95 percent confidence interval:
```

0.6810288 1.0000000

sample estimates:

probability of success

0.8333333

Since the p-value is less than 0.05, reject the null hypothesis.

References

Black, S., Shinefield, H., Fireman, B., Lewis, E., Ray, P., Hansen, J. R., ... & Edwards, K. (2000). Efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children. The Pediatric infectious disease journal, 19(3), 187-195.

Arnold, R. C., Shapiro, N. I., Jones, A. E., Schorr, C., Pope, J., Casner, E., ... & Trzeciak, S. (2009). Multicenter study of early lactate clearance as a determinant of survival in patients with presumed sepsis. Shock, 32(1), 35.

Tacket, C. O., Brenner, F., & Blake, P. A. (1984). Clinical features and an epidemiological study of Vibrio vulnificus infections. Journal of Infectious Diseases, 149(4), 558-561.

Back to Table of Contents

TEST 43 ONE SAMPLE PROPORTIONS TEST

Question the test addresses

Is the observed proportion (probabilities of success)from a random experiment is equal to some pre-specified probability?

When to use the test?

This test is used you have a simple random sample where each observation can result in just two possible outcomes, a success and a failure.

Practical Applications

Detection of HIV-Specific T Cell Responses: Frahm et al (2007) report the design of a peptide test set with significantly increased coverage of HIV sequence diversity by including alternative amino acids at variable positions during the peptide synthesis step. The researchers assessed whether toggled peptides not only detected more, but also stronger in vitro responses. They found the number of the increased responses was significantly greater for the toggled peptides than the consensus when only one of the two peptide test sets scored positive (total T cells: p-value = 7.3 × 10−8, CD4 T cells: p-value = 3.3 × 10−6, using a 1-sample proportions test).

Urine cytology: Yoder et al (2007) followed up 250 patients with urine cytologic results, concurrent multitarget fluorescence in situ hybridization, and cystoscopic examination for recurrent urothelial carcinoma. Patient characteristics were analyzed to detect imbalance in the cohort according to age 60 or older, sex and specimen type using a one-sample proportions test. Of the 250 patients 39 were 60 or older (p-value <0.05 one-sample proportions test), 187 were male (p-value <0.05 one -sample proportions test) and 197 voided specimen types were observed (p-value <0.05 one - sample proportions test).

Computing in those over 50: Goodman and Syme (2003) conduct a questionnaire on computer use and ownership with 353 participants over the age of 50. They used the one-sample proportions test to analyze the response to the item 'how respondents who have used computers learnt how to do so'. They found the most common method was through a computing course (one-sample proportions test p-value <0.05).

How to calculate in R

The function prop.test{stats}can be used to perform this test. It takes the form prop.test (x, n, p = 0.5, alternative = "two.sided", conf.level = 0.95).

Note, x is the number of observed successes, n the number of trials and p is hypothesized probability of success. For a one sided test set alternative =”less” or alternative =”greater”.

Example:

Suppose you toss a coin 100 times and get 52 heads. We can use prop.test to assess whether or not the coin is fair. To do so enter the following:

```
> prop.test(52,100,p=0.5, , alternative = "two.sided", conf.level = 0.95)

        1-sample proportions test with continuity correction

data:  52 out of 100, null probability 0.5

X-squared = 0.09, df = 1, p-value = 0.7642

alternative hypothesis: true p is not equal to 0.5

95 percent confidence interval:

 0.4183183 0.6201278

sample estimates:

   p

0.52
```

Since the p-value is greater than 0.05, we cannot reject the null hypothesis that the population proportion is 0.5; Therefore we can consider the coin to be fair.

References

Frahm, N., Kaufmann, D. E., Yusim, K., Muldoon, M., Kesmir, C., Linde, C. H. ... & Korber, B. T. (2007). Increased sequence diversity coverage improves detection of HIV-specific T cell responses. The Journal of Immunology, 179(10), 6638-6650.

Goodman J, Syme, A Eisma R (2003) Older Adults' Use of Computers: A Survey. In: Proceedings of HCI 2003. (Vol 2), Bath, UK, pp. 25-38.

Yoder, B. J., Skacel, M., Hedgepeth, R., Babineau, D., Ulchaker, J. C., Liou, L. S., ... & Tubbs, R. R. (2007). Reflex UroVysion Testing of Bladder Cancer Surveillance Patients With Equivocal or Negative Urine Cytology A Prospective Study With Focus on the Natural History of Anticipatory Positive Findings. American journal of clinical pathology, 127(2), 295-301.

Back to Table of Contents

TEST 44 ONE SAMPLE POISSON TEST

Question the test addresses

Is the rate parameter of a Poisson distributed sample significantly different from a hypothesized value?

When to use the test?

This test is used when you have collected a random sample of count data which follow the Poisson distribution.

Practical Applications

Familial cell carcinoma: Kiemeney et al (1997) conduct an epidemiological study of familial bladder cancer among the Icelandic population. For 190 patients with bladder, ureter or renal pelvis transitional cell carcinoma, the first to third degree relatives were identified. The observed occurrence of transitional cell carcinoma of the urinary tract was compared to the expected occurrence by age, gender, and calendar specific incidence rates. The observed and expected frequencies were assessed using the one sample Poisson test. The researchers observed six cases of transitional cell cancer in first degree relatives with disease versus an expected frequency of 6.2 (95% confidence interval 0.35 to 2.10).

Pediatric Cardiac Surgery: Nieminen, Jokinen, and Sairanen (2007) studied all late deaths of patients operated on for congenital heart defect in Finland during the years 1953 to 1989. The researchers calculated the survival of patients, identified the causes of deaths from death certificates, and examined the modes of congenital heart defect-related deaths. They then compared the survival and the causes of non– congenital heart defect related deaths to those of the general population using the Poisson test. The observed number of accidental deaths was 28 in the patient population; the expected value was 44 (95% confidence interval 0.42 to 0.92). The researchers conclude patients died in accidents less often than the general population.

Excess Mortality in Obesity: A sample of 6,193 obese German patients were recruited in Düsseldorf and followed for 14 years. Bender et al (1998) grouped the cohort according to their Body Mass Index. Using mortality tables of the general population of the region, the one sample Poisson test was used to investigate the link between obesity and excess mortality. The researchers observed for men with a Body Mass Index of 40 a greater a p-value of less than 0.01. For women with a Body Mass Index of 40 or greater the researchers report a p-value of less than 0.01.

How to calculate in R

The function poisson.test{stats}can be used to perform this test. It takes the form poisson.test(observed,expected,alternative="two.sided",conf.level=0.95).

Note, observed are the number of observed events, expected the number expected from the Poisson distribution. For a one sided test set alternative =”less” or alternative =”greater”.

Example:

Kiemeney et al (1997) observed 6 cases of transitional cell cancer in first degree relatives of Icelandic probands with disease. The expected frequency was 6.2. To assess the null hypothesis of no difference enter:

> poisson.test(6,6.22,alternative="two.sided",conf.level=0.95)

Exact Poisson test

data: 6 time base: 6.22

number of events = 6, time base = 6.22, p-value = 1

alternative hypothesis: true event rate is not equal to 1

95 percent confidence interval:

0.3540023 2.0995939

sample estimates:

event rate

0.9646302

Since the p-value is greater than 0.05, do not reject the null hypothesis.

References

Bender, R., Trautner, C., Spraul, M., & Berger, M. (1998). Assessment of excess mortality in obesity. American Journal of Epidemiology, 147(1), 42-48.

Kiemeney, L. A., Moret, N. C., Witjes, J. A., Schoenberg, M. P., & Tulinius, H (1997). Familial transitional cell carcinoma among the population of Iceland. The Journal of urology, 157(5), 1649-1651.

Nieminen, H. P., Jokinen, E. V., & Sairanen, H. I. (2007). Causes of late deaths after pediatric cardiac surgery: a population-based study. Journal of

the American College of Cardiology, 50(13), 1263-1271.

Back to Table of Contents

TEST 45 PAIRWISE COMPARISON OF PROPORTIONS TEST

Question the test addresses

Is the difference between the pairwise proportions in three or more samples significant?

When to use the test?

This test is used when you want to know if the pairwise observed frequencies of three or more dichotomous samples on the same factor differ from each other. It is based on Pearson's Chi-Squared test with Yates' continuity correction alongside various corrections for multiple testing.

Practical Applications

Dark-eyed juncos: Wolf, Ketterson and Nolan (1988) study whether parental care by male dark-eyed juncos (junco hyemalis) increases either the quantity or quality of young that they produce. Mated pairs were divided into an experimental and a control treatment group and over a 4-year period, males were caught at the time their eggs hatched, and the subsequent growth and survival of the young of unaided females and control pairs were compared. Pairwise comparisons of the proportions were undertaken using Pearson's Chi-Squared test for survival by treatment, survival by age, and treatment by age. The researchers report that all pairwise interactions were significant (pairwise comparison of proportions test p-value <0.025).

Sound frequency and reef fish: Simpson et al (2005) compare the settlement of fishes to patch reefs where high frequency, low frequency or no sound was broadcast. The researchers found apogonids (cardinalfish) were attracted to reefs with either high or low frequency sound, but pomacentrids (damselfish) tended to be attracted to reefs with high frequency sound. The researchers conducted a pairwise test for sound and pomacentrids. They report the relationship between high frequency and low frequency sound reefs as significant (pairwise comparison of proportions test p-value < 0.05). The pairwise comparison of proportions test p-value between high frequency and no sound reefs was less than 0.01.

User-Friendliness of Formats: Dolnicar and Grun (2006) investigate consumer preferences for one of five answer formats. A total of 236 first

year marketing students at the University of Wollongong were asked to complete a survey on water recycling. The data was collected at the University of Wollongong among students attending a first year lecture in marketing. The students were asked to complete a survey on water recycling. The students could choose their favorite answer format out of five different formats. The five answer formats were binary (yes – no, dichotomous), 3-point scale, 7-point scale, continuous and percentage scale. Pairwise comparisons of the proportions are undertaken using Pearson's Chi-Squared test with Yates' continuity correction and Holm's method to correct for multiple testing. The p-value on the pairwise comparison between the 7-point scale and the 3 – point scale was 0.015, and between the 7-point scale and the binary scale was 0.579. Overall, Dolnicar and Grun conclude no single most popular answer format exists, and that the ordinal multi-category answer formats (binary, 3-point, 7-point) are generally preferred to formats where the answer is recorded on a nearly continuous scale.

How to calculate in R

The function pairwise.prop.test{stats}can be used to perform this test. It takes the form pairwise.prop.test(sample, p.adjust.method ="holm"). Note, p.adjust.method refers to the p-value adjustment due to the multiple comparisons. The adjustment methods include the Bonferroni correction ("bonferroni") in which the p-values are multiplied by the number of comparisons. Less conservative corrections include "holm", "hochberg", "hommel", "BH" (Benjamini & Hochberg adjustment), and "BY' (Benjamini & Yekutieli adjustment).

Example: with holm adjustment

Suppose you have collected the following data on six samples

	Treatment 1	Treatment 2
Sample 1	95	106
Sample 2	181	137
Sample 3	76	85
Sample 4	13	29
Sample 5	11	26
Sample 6	201	179

The data can be entered into R by typing:

> sample<-rbind(s1=c(95,106),s2=c(181,137), s3=c(76,85),s4=c(13,29),s5=c(11,26),s6=c(201,179))

> colnames(sample) <-c("treat1","trea2")

The test can then be carried out by typing:

> pairwise.prop.test(sample, p.adjust.method ="holm")

Pairwise comparisons using Pairwise comparison of proportions

data: sample

	s1	s2	s3	s4	s5
s2	0.437	-	-	-	-
s3	1.000	0.553	-	-	-
s4	0.658	0.039	0.658	-	-
s5	0.658	0.042	0.658	1.000	-
s6	1.000	1.000	1.000	0.146	0.146

In the case the p-value of sample 2 and sample 4 is 0.039 and significant at the 5% level. The p-value between sample 2 and sample 5 is also significant with a p-value of 0.042.

Example: with Benjamini & Yekutieli adjustment

To use the more conservative Benjamini & Yekutieli adjustment with the data from the above example type:

> pairwise.prop.test(sample, p.adjust.method ="BY")

Pairwise comparisons using Pairwise comparison of proportions

data: sample

	s1	s2	s3	s4	s5
s2	0.395	-	-	-	-
s3	1.000	0.429	-	-	-
s4	0.429	0.075	0.429	-	-
s5	0.429	0.075	0.429	1.000	-

s6 1.000 1.000 1.000 0.147 0.147

P value adjustment method: BY

In the case the p-value of sample 2 and sample 4 is 0.075 and not significant at the 5% level. The p-value between sample 2 and sample 5 is also no significant with a p-value of 0.075. However, both comparisons are significant at the 10% level.

References

Dolnicar, S., & Grun, B. (2006). The user-friendliness of alternative answer formats. Faculty of Commerce-Papers, 240.

Simpson, S. D., Meekan, M., Montgomery, J., McCauley, R., & Jeffs, A (2005). Homeward sound. Science (New York, NY), 308(5719), 221.

Wolf, L., Ketterson, E. D., & Nolan, V. (1988). Paternal influence on growth and survival of dark-eyed junco young: do parental males benefit?. Animal Behaviour, 36(6), 1601-1618.

Back to Table of Contents

TEST 46 TWO SAMPLE POISSON TEST

Question the test addresses

Is the rate parameter of a Poisson distributed sample significantly different from a hypothesized value?

When to use the test?

This test is used when you have collected random sample of count data which follow the Poisson distribution.

Practical Applications

Freshwater jellyfish: The impact of freshwater hydrozoan jellyfish Craspedacusta sowerbii on prey items Bosmina longirostris, amongst others was assessed by Smith and Alexander (2008). A two-sample Poisson test for means to determine if the abundance of pray species present in lake water was significantly reduced by the presence of freshwater hydrozoan jellyfish Craspedacusta sowerbii. In one experiment, the abundances of the most commonly observed species were significantly reduced, compared to the controls: Bosmina longirostris (p-value < 0.008). The researchers conclude the presence of freshwater hydrozoan jellyfish Craspedacusta sowerbii significantly increased prey mortality.

Plant Science: Using two parental clones (M1 and M2) of Trifolium ambiguum, Hay et al (2010) investigate seed development A two-sample Poisson test was used to compare the numbers of seeds from each of the two clones. One inflorescence was taken from one of the M1 plants on each of 14, 22, 28, 30, 33, 36, 40, 44, 47, 50, 54, 58, 61, and 64 days after pollination. In addition one inflorescence was taken from one of the M2 plants on each of 22, 28, 36, 40, 47, 50, 58, and 61 days after pollination. In one experiment the researchers observed M2 inflorescences produced significantly fewer seeds (Poisson test p-value < 0.001).

Windshield splatter: Biodiversity estimates between geographic locations collected by a moving vehicle are assessed by Pond et al (2009). The researchers design and test a system for phylogenetic profiling of metagenomic samples. The number of sequencing reads was used as a proxy for the relative biodiversity. In order to access the significance of differences in read counts corresponding to a particular taxon between trip A and trip B, a Poisson two-sample test was used. Taxa with p-values significant at 1% were considered as significant of differences between the two trips.

How to calculate in R

The function poisson.test{stats}can be used to perform this test. It takes the form

poisson.test(c(observed_sample1, observed_sample2), c(size_sample1 , size_sample2) , alternative="two.sided" , conf.level=0.95)

Note, observed_sample1 and observed_sample2 are the number of observed events in sample 1 and sample 2 respectively. For a one sided test set alternative =”less” or alternative =”greater”.

Example:

Suppose we observe rates of 2 out of 17887 for the first sample and 10 out of 20000 for the second sample we can assess whether the these samples differ by entering:

>poisson.test(c(10,2),c(20000,17877),alternative="two.sided",conf.level=0.95

Comparison of Poisson rates

data: c(10, 2) time base: c(20000, 17877)

count1 = 10, expected count1 = 6.336, p-value = 0.04213

alternative hypothesis: true rate ratio is not equal to 1

95 percent confidence interval:

0.9524221 41.9509150

sample estimates:

rate ratio

4.46925

Since the p-value is less than 0.05, reject the null hypothesis.

References

Hay, F. R., Smith, R. D., Ellis, R. H., & Butler, L. H. (2010). Developmenta changes in the germinability, desiccation tolerance, hardseededness, and longevity of individual seeds of Trifolium ambiguum. Annals of botany, 105(6), 1035-1052.

Pond, S. K., Wadhawan, S., Chiaromonte, F., Ananda, G., Chung, W. Y., Taylor, J., & Nekrutenko, A. (2009). Windshield splatter analysis with the Galaxy metagenomic pipeline. Genome research, 19(11), 2144-2153.

Smith, A. S., & Alexander, J. E. (2008). Potential effects of the freshwater jellyfish Craspedacusta sowerbii on zooplankton community abundance. Journal of plankton research, 30(12), 1323-1327.

Back to Table of Contents

TEST 47 MULTIPLE SAMPLE PROPORTIONS TEST

Question the test addresses

Is the difference between the observed proportion (probabilities of success) from two or more samples significantly different from zero?

When to use the test?

This test is used when you have multiple simple random samples where each observation can result in just two possible outcomes, a success and a failure.

Practical Applications

Knowledge about the human papillomavirus vaccine: Knowledge about efficacy and safety of human papillomavirus (HPV) vaccine is of ongoing concern to health professionals. Ragin et al (2009) evaluate perception of the vaccine in the adult population of Pittsburgh, Pennsylvania, USA and Hampton, Virginia. A total of 202 participants (55% white, 45% Black) participated in the survey. A two-sample proportions test of significance was performed to compare demographic variables. There was no significance difference between the two groups to the question "Have you heard of the Human Papillomavirus (HPV)?" (p-value >0.1).

Frontotemporal degeneration versus Alzheimer's: In a retrospective case-control study of participants at two Alzheimer's disease centers, Chow, Hynan, and Lipton (2009) studied whether Mini-Mental State Examination sub-scores reflect the disease progression projected by the clinical criteria of frontotemporal degeneration versus Alzheimer's disease. The two independent samples proportion test indicated a lower percentage of frontotemporal subjects (11 out of 29 participants) lost constructional praxis as tested by this Mini-Mental State Examination item than the Alzheimer's disease group (14 out of 18 participants (two sample proportion p-value = 0.018).

Gender gap amongst academics: Jordan, Clark, and Vann (2011) examine whether a gender gap exists in publication productivity of male and female associate professors of accounting at doctoral and nondoctoral granting institutions. As part of their study they investigate whether men and women differ by quality of academic training. They find the proportion of male faculty at nondoctoral institutions trained at tier one or two schools is 60% (18 out of 30) while the proportion of female faculty at nondoctoral institutions who received their doctorates from tier one or two universities is 70.8% (17 out of 24). A multiple sample proportions test indicates no

significant difference between the two groups (p-value >0.05).

How to calculate in R

The function prop.test{stats}can be used to perform this test. It takes the form prop.test (c(success_1, success_2,…,success_n), c(number_1, number_2,…,number_n), alternative = "two.sided", conf.level = 0.95).

Note, success_i is the number of observed successes in group i, number_i the total number participants in group i, and p is hypothesized probability of success. For a one sided test set alternative =”less” or alternative =”greater”.

Example: Gender gap amongst academics

Jordan, Clark, and Vann (2011) find the proportion of male faculty at nondoctoral institutions trained at tier one or two schools is 60% (18 out of 30) while the proportion of female faculty at nondoctoral institutions who received their doctorates from tier one or two universities is 70.8% (17 out of 24). We can assess this using the multiple sample proportions test as follows:

```
> prop.test(tier_1_or_2, total, alternative = "two.sided", conf.level = 0.95)

        2-sample test for equality of proportions with continuity correction

data:  tier_1_or_2 out of total

X-squared = 0.2933, df = 1, p-value = 0.5881

alternative hypothesis: two.sided

95 percent confidence interval:

 -0.3984195  0.1817529

sample estimates:

   prop 1    prop 2

0.6000000 0.7083333
```

Since the p-value is greater than 0.05, we cannot reject the null hypothesis.

References

Chow, T. W., Hynan, L. S., & Lipton, A. M. (2006). MMSE scores decline at a greater rate in frontotemporal degeneration than in AD. Dementia and geriatric cognitive disorders, 22(3), 194-199.

Jordan, C. E., Clark, S. J., & Vann, C. E. (2011). Do Gender Differences Exist I The Publication Productivity Of Accounting Faculty?. Journal of Applied Business Research (JABR), 24(3).

Ragin, C. C., Edwards, R. P., Jones, J., Thurman, N. E., Hagan, K. L., Jones, I A., ... & Taioli, E. (2009). Knowledge about human papillomavirus and the HPV vaccine–a survey of the general population. Infect Agent Cancer, 4(Suppl 1), 1-10.

Back to Table of Contents

TEST 48 CHI-SQUARED TEST FOR LINEAR TREND

Question the test addresses

Is the difference between the observed proportion (probabilities of success) from two or more samples with a linear trend significantly different from zero?

When to use the test?

This test is used when you have multiple simple random samples where each observation can result in just two possible outcomes, a success and a failure and you observe a trend in sample proportions. It is based on a test for zero slope in the linear regression of the proportions on the group scores.

Practical Applications

Fecal resistance: Okeke et al (2000) tested 758 fecal Escherichia coli isolates, recovered from Nigerian students in 1986, 1988, 1990, 1994, and 1998, for susceptibility to seven antimicrobial drugs. They observed prevalence's of strains resistant to tetracycline, ampicillin, chloramphenicol, and streptomycin were 9% to 35% in 1986 and 56% to 100% in 1998. The trend in resistance was formally analyzed by the chi-square test for trend. The researchers find the trend for tetracycline and streptomycin were statistically significant at the 10 % level (Chi-squared test for linear trend p-value <0.1). The researchers also observe the proportion of isolates resistant to three or more drugs increased steadily over the period of their study, from 30.2% in 1986 to 70.5% in 1998 (Chi-squared test for linear trend p-value <0.10). The authors conclude by observing their findings demonstrate that resistance gene reservoirs are increasing in healthy persons.

Dangers of swimming in Los Angeles: During the summer months of July and August 1988 in Los Angeles an outbreak of gastroenteritis affected 44 persons from 5 independent swimming groups who had used the same swimming pool. The cause was identified as Cryptosporidium by Sorvillo et al (1992) who apply statistical analysis to the incident as part of a public health investigation. The researchers use the Chi-squared test for linear trend to assess the relationship between time in the water and attack rate. They categorize time in the water as 1-3 hours, 4-6 hours and greater than 6 hours. They report the attack rate was highest for those spending more time in the water (p-value <0.001).

Adverse perinatal outcomes: The risk of adverse perinatal outcome was

related to maternal circulating concentrations of trophoblast-derived proteins at 8–14 wk gestation among women recruited to a multicenter, prospective cohort study undertaken by Smith et al (2002). Clotted blood sample were assayed for PAPP-A along with other proteins. Five dichotomous outcomes were defined: delivery of a small-for-gestational-age baby, moderately preterm delivery, extremely preterm delivery, preeclampsia, and stillbirth. Excluding stillbirths, the researchers observed a linear trend in proportions between birth weight and PAPP-A - the lowest decile of PAPP-A consistently had the highest proportion of adverse outcomes. They reported when data from the smallest decile were excluded, the test for trend remained statistically significant for birth weight less than fifth percentile (Chi-squared test for linear trend p– value < 0.0001) and delivery between 33–36 wk (p-value = 0.006) but was no longer statistically significant for the preeclampsia group (Chi-squared test for linear trend p-value = 0.22).

How to calculate in R

The function prop.trend.test{stats}can be used to perform this test. It takes the form prop.test (c(success_1, success_2,…,success_n), c(number_1, number_2,…,number_n)).

Note, success_i is the number of observed successes in group i, number i the total number participants in group i, and p is hypothesized probability of success. For a one sided test set alternative =”less” or alternative =”greater”.

Example: The dangers of swimming in Los Angeles

Sorvillo et al (1992) use the test to assess the relationship between time in the water and attack rate of Cryptosporidium amongst swimmers. They categorize time in the water as 1-3 hours, 4-6 hours and greater than 6 hours, with 5 out of 13, 5 out of 8, and 33 out of 37 swimmers reporting symptoms for each category respectively. We can assess this data using the Chi-squared test for linear trend test as follows:

```
> infected.swimmers  <- c( 5,5,33)

> all.swimmers <- c( 13,8,37)

> prop.trend.test(infected.swimmers, all.swimmers)

        Chi-squared Test for Trend in Proportions

data:  infected.swimmers out of all.swimmers ,
```

using scores: 1 2 3

X-squared = 13.5605, df = 1, p-value = 0.000231

Since the p-value is less than 0.05, we reject the null hypothesis of no linear trend.

References

Okeke, I. N., Fayinka, S. T., & Lamikanra, A. (2000). Antibiotic resistance in Escherichia coli from Nigerian students, 1986 1998. Emerging infectious diseases, 6(4), 393.

Smith, G. C., Stenhouse, E. J., Crossley, J. A., Aitken, D. A., Cameron, A. D., & Connor, J. M. (2002). Early pregnancy levels of pregnancy-associated plasma protein a and the risk of intrauterine growth restriction, premature birth, preeclampsia, and stillbirth. Journal of Clinical Endocrinology & Metabolism, 87(4), 1762-1767.

Sorvillo, F. J., Fujioka, K., Nahlen, B., Tormey, M. P., Kebabjian, R., & Mascola, L. (1992). Swimming-associated cryptosporidiosis. American Journal of Public Health, 82(5), 742-744.

Back to Table of Contents

TEST 49 PEARSON'S PAIRED CHI-SQUARED TEST

Question the test addresses

Are the paired observations on two variables in a contingency table independent of each other?

When to use the test?

A Chi-square test is designed to analyze categorical data. That means that the data has been counted and divided into categories. The test is used to discover if there is a relationship between two categorical variables, or to assess whether a sample on a categorical variable is different from a specific probability distribution. It is assumed you have collected an independent random sample of reasonable size.

Practical Applications

Joint pain: Eight hundred and forty-six patients with joint pain were recruited into a randomised double-blind trial by Parr et al (2012). The objective was to compare the efficacy, tolerability and effect on quality of life of daily dose of diclofenac sodium (DS) slow release and a combination of dextropropoxyphene and paracetamol (DP). The chi-square test was used to examine whether the two treatment groups were well matched for age and sex respectively (Chi-squared test p-value >0.05). The researchers also used the Chi-square test to assess limitations of movement. They find a significant advantage to patients on DS with 120 patients improving, 222 not changing and 7 deteriorating (chi-squared test p-value < 0.05).

Telemonitoring heart failure: Chaudhry, (2010) randomly assigned 1653 patients who had recently been hospitalized for heart failure to undergo either telemonitoring (826 patients) or usual care (827 patients). The researchers compared readmission for both groups using the chi-squared test. They found readmission occurred in 49.3% of patients in the telemonitoring group and 47.4% of patients in the usual-care group (chi-square test p-value = 0.45). The null hypothesis of no difference could not be rejected. Death occurred in 11.1% of the telemonitoring group and 11.4% of the usual care group (chi-square test p-value = 0.88). Again, the null hypothesis of no difference between the two groups could not be rejected.

Cardiovascular risk & bipolar disorder: The relationship between coronary heart disease and cardiovascular mortality risk in patients with bipolar disorder is investigated by Garcia-Portilla (2009). The study enrolled 194 patients with bipolar disorder. The researchers find the risk of Coronary

Heart Disease and Cardiovascular Mortality Risk significantly increasec with age in both males and females (Chi-squared test p-value <0.01).

How to calculate in R

The function chisq.test{stats}can be used to perform this test. It takes the form chisq.test(Table_data , correct = FALSE).

Note, set correct = TRUE when the number of observations is small, the function will then use a continuity correction when computing the test statistic.

Example: standard Chi-squared test

Suppose you have collected the following data on the voting patterns of 100 British citizens.

Gender	Labour	Conservative
Male	20	30
Female	30	20

This data can be entered into R using the following:

```
>Table_data<- as.table(rbind(c(20, 30), c(30,20)))

dimnames(Table_data) <- list(gender=c("Male","Female"), party=c("Labour", "Conservative"))
```

To conduct a chi-squared test enter:

```
> chisq.test(Table_data , correct = FALSE )

        Pearson's Chi-squared test

data:  Table_data

X-squared = 4, df = 1, p-value = 0.0455
```

Since the p-value is less than 0.05, reject the null hypothesis.

Example: Chi-squared test with continuity correction

Using the above data, type:

```
> chisq.test(Table_data , correct = TRUE)

        Pearson's Chi-squared test with Yates' continuity correction

data:  Table_data
```

X-squared = 3.24, df = 1, p-value = 0.07186

In this case the p-value is greater than 0.05, do not reject the null hypothesis at the 5% level of significance.

References

Chaudhry, S. I., Mattera, J. A., Curtis, J. P., Spertus, J. A., Herrin, J., Lin, Z., . & Krumholz, H. M. (2010). Telemonitoring in patients with heart failure. New England Journal of Medicine, 363(24), 2301-2309.

Garcia-Portilla, M. P., Saiz, P. A., Bascaran, M. T., Martíneza, S., Benabarre, A., Sierra, P., ... & Bobes, J. (2009). Cardiovascular risk in patients with bipolar disorder. Journal of affective disorders, 115(3), 302-308.

Parr, G., Darekar, B., Fletcher, A., & Bulpitt, C. J. (2012). Joint pain and quality of life; results of a randomised trial. British journal of clinical pharmacology, 27(2), 235-242.

Back to Table of Contents

TEST 50 FISHERS EXACT TEST

Question the test addresses

Are the paired observations on two variables in a contingency table independent of each other?

When to use the test?

The test is used to discover if there is a relationship between two categorical variables, or to assess whether a sample on a categorical variable is different from a specific probability distribution. It is assumed you have collected an independent random sample. It is often used when the number of observations is small.

Practical Applications

Asthma: Asthmatic subjects (16 women, 23 men), not taking systemic steroids and 15 age matched healthy controls (8 women, 7 men) were recruited into a study by Bullens et al (2006). Asthma severity was categorized as mild, moderate and severe. Asthmatic subjects were further subdivided into atopics (n = 21) and non-atopics, (n = 17). The researchers found no differences in FEV1% (Fishers' exact test p-value = 0.48), asthma severity classification (Fishers' exact test p-value = 0.49) or inhaled corticosteroids use (Fishers' exact test p-value = 0.28) between the allergic and the non-allergic asthmatics.

Night-time calf cramp: Blyton, Chuter and Burns (2012) explore the experience of night-time calf cramp in 80 adults who experienced night-time calf cramp at least once per week from the Hunter region in New South Wales, Australia. The researchers report those who suffered from day time muscle cramp were no more likely to experience night-time muscle cramp of muscles other than the calf (Fisher's exact test p-value = 0.68). They also observed subjects who experienced day time calf cramp were no more likely to experience more frequent night-time calf cramp (Fisher's exact test p-value = 0.50).

Extubation failure: Ko, Ramos, and Chaltela (2009) in an retrospective observational study, assess the ability of traditional weaning parameters to predict extubation failure in neurocritical (coma) patients. The researchers use the Four Scale (which evaluates brainstem function) obtained from the nursing notes, physicians' progress notes and direct calculation. Data on 62 patients undergoing extubation trial at neurological intensive care unit were assessed in the study. In patients, who failed extubation, 3 out of 11 had Four Scores of less than 12, in the group that was successfully

extubated, 11 out of 51 had Four Scores below 12 (Fishers exact test p-value = 0.6997). In 52 patients a spontaneous breathing trial was performed. The researchers found no significant difference between patients undergoing a spontaneous breathing trial and not undergoing it in terms of extubation failure (Fishers exact test p-value = 0.6708).

How to calculate in R

The function fisher.test{stats}can be used to perform this test. It takes the form fisher.test(Table_data, alternative = "two.sided", conf.level = 0.95).

Note to specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater").

Example: two-sided exact Fisher test

Suppose you have collected the following data on the voting patterns of 100 British citizens.

Gender Labour Conservative

Male 20 30

Female 30 20

This data can be entered into R using the following:

```
>Table_data<- as.table(rbind(c(20, 30), c(30,20)))

dimnames(Table_data) <- list(gender=c("Male","Female"), party=c("Labour", "Conservative"))
```

To conduct a chi-squared test enter:

```
> fisher.test(Table_data, alternative = "two.sided", conf.level = 0.95)

        Fisher's Exact Test for Count Data

data:  Table_data

p-value = 0.07134

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

 0.1846933 1.0640121

sample estimates:
```

odds ratio

0.4481632

Since the p-value is less than 0.05, reject the null hypothesis.

Example: one sided exact Fisher test

Using the above data, type:

> fisher.test(Table_data, alternative = "less", conf.level = 0.95)

Fisher's Exact Test for Count Data

data: Table_data

p-value = 0.03567

alternative hypothesis: true odds ratio is less than 1

95 percent confidence interval:

0.0000000 0.9391675

sample estimates:

odds ratio

0.4481632

The p-value is less 0.05, reject the null hypothesis at the 5% level of significance.

References

Blyton, F., Chuter, V., & Burns, J. (2012). Unknotting night-time muscle cramp: a survey of patient experience, help-seeking behaviour and perceived treatment effectiveness. Journal of Foot and Ankle Research, 5(1), 7.

Bullens, D. M., Truyen, E., Coteur, L., Dilissen, E., Hellings, P. W., Dupont, L J., & Ceuppens, J. L. (2006). IL-17 mRNA in sputum of asthmatic patients linking T cell driven inflammation and granulocytic influx. Respir Res, 7(1), 135.

Ko, R., Ramos, L., & Chalela, J. A. (2009). Conventional weaning parameters do not predict extubation failure in neurocritical care patients. Neurocritical care, 10(3), 269-273.

Back to Table of Contents

TEST 51 COCHRAN-MANTEL-HAENSZEL TEST

Question the test addresses

Is there a relationship between two categorical variables after adjusting for control variables?

When to use the test?

To test of the null hypothesis that two nominal variables are conditionally independent in each stratum, assuming that there is no three-way interaction. You have categorical data and want test whether the frequency distribution of values differs between groups on which you have taken repeated measurements. The initial data are represented as a series of K 2x2 contingency tables, where K is the number of measurement conditions. The rows usually correspond to the "Treatment group" values (e.g. "Placebo", "Drug ") and the columns to the "Recovery" values (e.g. "No change," "Improvement"). The null hypothesis is that the response is conditionally independent of the treatment. For example, you may want to know whether a treatment ("Drug " versus "Placebo") impacts the likelihood of recovery ("No change" or "Improvement"). If the treatments were administered at three different times of day, morning, afternoon, and night, and you want to control for this, you would use a 2x2x3 contingency table, where the third variable is the one you wish to control for.

Practical Applications

Hypoglycemia risk: Rosenstock et al (2005) assessed the risk for hypoglycemia in a meta-analysis for insulin glargine (total of 1,142 individuals) versus once- or twice-daily neutral protamine Hagedorn insulin (total of 1,162 individuals) in adults with type 2 diabetes. The analysis covered a total of 84 pooled study centers from four clinical studies. The Cochran-Mantel-Haenszel test was used to analyze categorical variables. Fasting plasma glucose levels were significantly lower at end point in the insulin glargine group than in the with neutral protamine hagedorn insulin group (p-value = 0.0233). The researchers conclude insulin glargine given once daily reduces the risk of hypoglycemia compared with neutral protamine hagedorn insulin.

Hematopoietic stem cell transplantation: Van Burik et al (2004) hypothesized chemoprophylaxis with echinocandin micafungin would be an effective agent for antifungal prophylaxis during neutropenia in patients undergoing hematopoietic stem cell transplantation. A total of 882 patients were recruited onto a double-blind randomized trial assigned to

50 mg of micafungin (1 mg/kg for patients weighing <50 kg) and 400 mg of fluconazole (8 mg/kg for patients weighing <50 kg) administered once per day. Success was defined as the absence of invasive fungal infection through the end of therapy and 4-weeks post treatment. The authors report the treatment success was greater in the micafungin group than in the fluconazole group (Cochran-Mantel-Haenszel test p-value =0 .026).

Comparative Effectiveness Research: Bourgeois et al (2012) carry out an observational study of clinical trials in the US between 2007 and 2010 addressing priority research topics defined by the Institute of Medicine. Searching various databases the researchers calculated the proportion of studies that were comparative effectiveness (CE) studies and compared study characteristics for CE and non-CE studies. After controlling for primary funding source, it was observed CE studies were less likely to report positive findings (Cochran-Mantel-Haenszel test p-value <0.007). Among CE studies involving a drug therapy, findings were positive for 30.0% (n = 3) of CE studies compared with 81.6% (n = 40) of non-CE studies (Cochran-Mantel-Haenszel test p-value <0.001).

How to calculate in R

The function mantelhaen.test{stats}can be used to perform this test. It takes the form mantelhaen.test(Data, alternative = "two.sided", correct = FALSE, exact = FALSE, conf.level = 0.95).

Note to specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater"). For continuity correction set correct = TRUE. If you set exact =TRUE the exact conditional test will be calculated.

Example: 2x2x2 contingency table (K=2)

Suppose you have collected data on the response to men and women for a new drug as follows.

Males on drug who report improvements = 12, males on drug with no change = 16.

Males on placebo who report improvements = 7, males on placebo with no change = 19.

Females on drug who report improvements = 16, Females on drug with no change = 11.

Females on placebo who report improvements = 5, Females on placebo

with no change = 20.

This data can be entered into R by typing the following:

```
Data <-array(c(12, 16, 7, 19,16, 11, 5, 20), dim = c(2, 2, 2),

dimnames = list(Treatment = c("Drug", " Placebo"),

Response = c("Improved", "No Change"),

Sex = c("Male", "Female")))
```

To conduct the Cochran-Mantel-Haenszel test type:

```
> mantelhaen.test(Data, alternative = "two.sided", correct = FALSE, exact =
FALSE, conf.level = 0.95)

        Mantel-Haenszel chi-squared test without continuity correction

data:  Data

Mantel-Haenszel X-squared = 8.3052, df = 1, p-value = 0.003953

alternative hypothesis: true common odds ratio is not equal to 1

95 percent confidence interval:

 1.445613 7.593375

sample estimates:

common odds ratio

        3.313168
```

The significant p-value 0f 0.003953 indicates that the association between treatment and response remains strong after adjusting for gender.

References

Bourgeois, F. T., Murthy, S., & Mandl, K. D. (2012). Comparative effectiveness research: an empirical study of trials registered in ClinicalTrials. gov. PLoS One, 7(1), e28820.

Rosenstock, J., Dailey, G., Massi-Benedetti, M., Fritsche, A., Lin, Z., & Salzman, A. (2005). Reduced Hypoglycemia Risk With Insulin Glargine A meta-analysis comparing insulin glargine with human NPH insulin in type 2 diabetes. Diabetes care, 28(4), 950-955.

Van Burik, J. A. H., Ratanatharathorn, V., Stepan, D. E., Miller, C. B., Lipton

J. H., Vesole, D. H., ... & Walsh, T. J. (2004). Micafungin versus fluconazole for prophylaxis against invasive fungal infections during neutropenia in patients undergoing hematopoietic stem cell transplantation. Clinical infectious diseases, 39(10), 1407-1416.

Back to Table of Contents

TEST 52 MCNEMAR'S TEST

Question the test addresses

Is there a difference between paired proportions?

When to use the test?

McNemar's test is basically a paired version of Chi-square test. It is mainly used when the sample consist of paired observations where the same subjects are measured twice. For example you want to assess whether the number of attendees who liked your latest play were significantly changed between before and after the screening. It is also used in circumstances where subjects are matched on some variable, or responses on two measures are used (e.g., favorability to shorter school holidays compared to favorability for the use of school vouchers). In essence, it tests for symmetry of rows and columns in a two-dimensional contingency table.

Practical Applications

Breath biomarkers and tuberculosis: Phillips et al (2010) hypothesize that volatile organic compounds (VOCs) in breath may contain biomarkers of active pulmonary tuberculosis. The sample consisted of breath VOCs of 226 symptomatic high-risk patients in UK, Philippines, and USA. Diagnosis of disease was based on sputum culture, smear microscopy and chest radiography. Mcnemar's test was used to assess concordance between these diagnostic tests. For sputum culture versus chest radiography, sputum culture versus smear and chest radiography versus smear microscopy, Mcnemar's test p-value was less than 0.01.The authors observe these statistically significant outcomes indicate low agreement between the diagnostic methods

Left cardiac sympathetic denervation: Clinical status and therapy before and after left cardiac sympathetic denervation were analyzed by Schwartz et al (1991). The sample consisted of eighty five patients worldwide who had been treated with left cardiac sympathetic denervation between March 1969 and October 1990. As part of the study treatment classes were dichotomized as "with beta-blockers "(alone or with other drugs) and "without, beta-blockers"(no therapy or miscellaneous). The researchers report McNemar's test for dichotomous outcome in matched samples p-value >0.05, and the null hypothesis of no difference cannot be rejected.

Adolescent obesity and depressive symptoms: The relationship between severe obesity and depressive symptoms over three years in fifty one adolescents in grades 7–12 was investigated by Goodman and Must (2011).

Obese participants were paired with an age, sex, and race normal weight subjects. Depressive symptoms (using the CESD scale) were assessed at baseline, 2 and 3. No relationship was observed at the 5% level of significance between weight status and CESD scores at baseline (p-value = 0.01) or 2 years (p-value =0.08). However, a positive association between weight status and CESD scores was present at 3 years (p-value p=0.02). The researchers conclude obesity-related programs should not assume severely obese adolescents are also suffering from a high degree of psychological distress.

How to calculate in R

The function mcnemar.test{stats}can be used to perform this test. It takes the form mcnemar.test(data).

Example: concordance of diagnostic tests

Phillips et al (2010) study concordance the diagnostic test of sputum culture versus chest radiography.We can enter the data given in their paper into R by typing:

```
data<-matrix(c(59, 4, 128, 20),
    nrow = 2,
    dimnames = list("chest radiography" = c("positive", "negative"),
                    "sputum culture" = c("positive", "negative")))
```

To conduct the McNemar's test type:

```
> mcnemar.test(data)

    McNemar's Chi-squared test with continuity correction

data:  data

McNemar's chi-squared = 114.6136, df = 1, p-value < 2.2e-16
```

Since the p-value is less than 0.05, we reject the null hypothesis.

References

Goodman, E., & Must, A. (2011). Depressive Symptoms in Severely Obese Compared With Normal Weight Adolescents: Results From a Community Based Longitudinal Study. Journal of Adolescent Health, 49(1), 64-69.

Phillips, M., Basa-Dalay, V., Bothamley, G., Cataneo, R. N., Lam, P. K. Natividad, M. P. R., ... & Wai, J. (2010). Breath biomarkers of active

pulmonary tuberculosis. Tuberculosis, 90(2), 145-151.

Schwartz, P. J., Locati, E. H., Moss, A. J., Crampton, R. S., Trazzi, R., & Ruberti, U. (1991). Left cardiac sympathetic denervation in the therapy of congenital long QT syndrome. A worldwide report. Circulation, 84(2), 503-511.

Back to Table of Contents

TEST 53 EQUAL MEANS IN A ONE-WAY LAYOUT WITH EQUAL VARIANCES

Question the test addresses

Do three or more samples come from populations with the same mean?

When to use the test?

The test is used in a situation where you have three or more independent samples on a treatment factor and you want to test for differences among the sample means. The population from which the samples were obtained is assumed to be normally distributed. The variances across the samples are assumed to be equal.

Practical Applications

Soil nitrogen levels on pest legume: Guenther and Roberts (2012) study the effect of varying soil nitrogen levels on the Lespedeza cuneata pest legume. Measurements of stem height and root, shoot, and total biomass of Lespedeza cuneata were taken for three weeks using four soil nitrogen treatments. Treatment one had no added nitrogen, treatment two had 50 parts per million (ppm) ammonium nitrate, treatment three had 100 ppm, and treatment four had 200 ppm. Testing days were 7, 12, 14,19 and 21 days after planting. The one way ANOVA test p-values were 0.402, 0.58, 0.737, 0.526 and 0.309 respectively. The authors conclude there was no significant variation in shoot height among nitrogen treatments on any measurement day.

Iodine concentration in milk: Bath, Button and Rayman (2012) compare the iodine concentration of retail organic and conventional milk. Ninety-two samples of organic and 80 samples of conventional milk, purchased at retail outlets in 16 areas of the United Kingdom were collected and analyzed. One-way ANOVA was used for comparison of iodine concentration between area of purchase and region of origin of the milk. The researchers found no difference in iodine concentration between the 16 areas of purchase of either supermarket own-brand organic or conventional milk samples (p-value = 0.75 and p-value = 0.49 respectively) or between the four regions of south east England, south west England, Wales and Northern Ireland (p-value = 0.36 and p-value = 0.66 respectively).

Stigmatization of obesity: Latner, Stunkard and Wilson (2012) assess stigmatization of obesity relative to the stigmatization of various

disabilities among young people. A total of 356 young people were recruited onto the study. Participants were asked to rank six drawings of adults with obesity, various disabilities, or no disability in order of how well they liked each person. The researchers divided participants into three categories based on their current Body Mass Index (BMI) and thei highest-ever BMI: 25 (overweight), 18.5 to 24.9 (normal weight), and less than 18.5 kg/m2 (underweight). One-way ANOVA revealed no differences ir participants' liking of any of the six drawings among the three weight categories (p-value >0.05).

How to calculate in R

The function oneway.test{stats}can be used to perform this test. It takes the form oneway.test(Value~ Sample_Group, data = data, var.equal = TRUE).

Example: using oneway.test

Suppose you have collected the following experimental data on three samples:

group	Value
1	2.9
1	3.5
1	2.8
1	2.6
1	3.7
2	3.9
2	2.5
2	4.3
2	2.7
3	2.9
3	2.4
3	3.8
3	1.2
3	2

We can enter this data into R by typing:

```
Value <- c(2.9, 3.5, 2.8, 2.6, 3.7, 3.9, 2.5, 4.3, 2.7, 2.9, 2.4, 3.8, 1.2, 2.0)

Sample_Group <- factor(c(rep(1,5),rep(2,4),rep(3,5)))

data <- data.frame(Sample_Group, Value)
```

To use the oneway.test test type:

```
> oneway.test(Value~ Sample_Group, data = data, var.equal = TRUE)

        One-way analysis of means

data:  Value and Sample_Group

F = 1.5248, num df = 2, denom df = 11, p-value = 0.2603
```

Since the p-value is greater than 0.05, do not reject the null hypothesis.

References

Bath, S. C., Button, S., & Rayman, M. P. (2012). Iodine concentration of organic and conventional milk: implications for iodine intake. British Journal of Nutrition, 107(07), 935-940.

Guenther, E. M., & Roberts, J. M. (2012). Soil Nitrogen Influences Early Root Allocation of Lespedeza cuneata. Tillers, 5, 21-23.

Latner, Janet D., Albert J. Stunkard, and G. Terence Wilson. "Stigmatized students: age, sex, and ethnicity effects in the stigmatization of obesity." Obesity Research 13.7 (2012): 1226-1231.

Back to Table of Contents

TEST 54 WELCH-TEST FOR MORE THAN TWO SAMPLES

Question the test addresses

Do your three or more samples come from populations with the same mean?

When to use the test?

The test is used in a situation where you have three or more independent samples on a treatment factor and you want to test for differences among the sample means. The population from which the sample was obtained is assumed to be normally distributed. The variances across the samples are assumed to be equal.

Practical Applications

Copyright permission request: For 744 copyright holder's Akmon (2010) assess the response time from patent office staff's initial permission request (to put the copyright online) until an answer is obtained from the right holder. Copyright holders were categorized as individual, non-profit, commercial, government, educational, association and unknown. The mean response time from staff's initial permissions request until an answer was obtained was 41 days. The Welch-test for more than two samples was used to determine any differences in mean response times between the six different types of copyright holders. Welch's test suggests that there were significant differences in the mean response time between the groups (p-value<0.001).

Math skills assessment: A math skills assessment was administered to students from three universities in five disciplines by Price et al (2012). The disciplines the students were studying were production (184 students), business statistics (230 students), quantitative analysis (181 students), statistics II (127 students) and microeconomics (104 students). Welch test for more than two samples was used to assess whether there was a significant difference in the mean percent correct responses of the five disciplines. The Welch-test for more than two samples generated a p-value of <0.0001. The researchers reject the null hypothesis of no significant difference between the mean performances of students by discipline.

Cultural variability in learning style: Sywelem et al (2012) examine how cultural variability is reflected in the learning style of students in Egypt, Saudi Arabia and United States. A total of 316 students were asked to complete the Steinbach Learning Style Survey; 118 were American students 94 were Saudi students and 104 were Egyptian students. The researchers

assess differences in mean scores using the Welch test for more than two samples. They report a statistically significant difference in means among the American, Egyptian and Saudi students (Welch test for more than two samples p-value < 0.01).

How to calculate in R

The function oneway.test{stats}can be used to perform this test. It takes the form oneway.test(Value~ Sample_Group, data = data, var.equal = FALSE

).

Example: using oneway.test

Suppose you have collected the following experimental data on three samples:

group	Value
1	2.9
1	3.5
1	2.8
1	2.6
1	3.7
2	3.9
2	2.5
2	4.3
2	2.7
3	2.9
3	2.4
3	3.8
3	1.2
3	2

We can enter this data into R by typing:

```
Value <- c(2.9, 3.5, 2.8, 2.6, 3.7, 3.9, 2.5, 4.3, 2.7, 2.9, 2.4, 3.8, 1.2, 2.0)

Sample_Group <- factor(c(rep(1,5),rep(2,4),rep(3,5)))

data <- data.frame(Sample_Group, Value)
```

To use the oneway.test test type:

```
> oneway.test(Value~ Sample_Group, data = data, var.equal = FALSE)

        One-way analysis of means (not assuming equal variances)

data:  Value and Sample_Group

F = 1.0565, num df = 2.000, denom df = 6.087, p-value = 0.4038
```

Since the p-value is greater than 0.05, do not reject the null hypothesis.

References

Akmon, D. (2010). Only with your permission: how rights holders respond (or don't respond) to requests to display archival materials online. Archival Science, 10(1), 45-64.

Price, B. A., Randall, C. H., Frederick, J., Gáll, J., & Jones, T. W. (2012) Different Cultures, Different Students, Same Test: Comparing Math Skills o Hungarian and American College Students. Journal of Education and Learning, 1(2), p128.

Sywelem, M., Al-Harbi, Q., Fathema, N., & Witte, J. E. (2012). Learning Styl Preferences of Student Teachers: A Cross-Cultural Perspective. Institute for Learning Styles Journal• Volume, 1, 10.

[Back to Table of Contents]

TEST 55 KRUSKAL WALLIS RANK SUM TEST

Question the test addresses

Do your three or more samples come from populations with the same mean?

When to use the test?

The test is used in a situation where you have three or more independent samples on a treatment factor and you want to test for differences among the sample means. Sample observations in each group are assumed to come from populations with the same shape of distribution.

Practical Applications

Oxytocin and trusting behavior: Kosfeld et al (2005) hypothesize that oxytocin increases the trusting behavior of Investors. As part of their analysis a trust game with real monetary stakes was created. A total 29 individuals were administered oxytocin before playing the game. A control group of 29, who did not receive oxytocin also played the game. No significant difference between the control group and the oxytocin group was observed (Kruskal-Wallis test p-value = 0.766).

Cortisol and traumatic memories: Aerni et al (2004) study whether cortisol administration can also reduce excessive retrieval of traumatic memories in patients with chronic posttraumatic stress disorder. The researchers use a single-case statistical analyses with Kruskal-Wallis nonparametric tests performed to assess treatment effects on daily symptom ratings over 3 months. Mr. C was a 55-year-old man who had a severe car accident several years before inclusion in the study. He was administered cortisol in the first month, followed by 2 months of placebo medication. Significant treatment effects were detected for the intensity of the feeling of reliving the traumatic event (Kruskal-Wallis test p-value <0.001), physiological distress (Kruskal-Wallis test p-value <0.001), and the frequency of nightmares (Kruskal-Wallis test p-value <0.05).

Seed survival: Seed survival and density, mortality, height, crown area, and basal diameters of seedlings and sprouts in tropical dry forest in lowland Bolivia were analyzed by Kennard et al (2002). Four treatments of varying disturbance intensity (high-intensity burn, low intensity burn, plant removal, and harvesting gap) were administered with results monitored over a period of 18 months following treatments. Distributions of seedling and sprout densities were not normally distributed and were therefore compared among treatments using Kruskall–Wallis test. The Kruskall–Wallis

test p-value at 3, 6, 9 and 12 months were all less than 0.01.

How to calculate in R

The function kruskal.test{stats}can be used to perform this test. It takes the form: kruskal.test(Value ~ Sample_Group, data=data).

Example: using the format kruskal.test(Value ~ Sample_Group, data=data)

Suppose you have collected the following experimental data on three samples:

group	Value
1	2.9
1	3.5
1	2.8
1	2.6
1	3.7
2	3.9
2	2.5
2	4.3
2	2.7
3	2.9
3	2.4
3	3.8
3	1.2
3	2

We can enter this data into R by typing:

```
Value <- c(2.9, 3.5, 2.8, 2.6, 3.7, 3.9, 2.5, 4.3, 2.7, 2.9, 2.4, 3.8, 1.2, 2.0)

Sample_Group <- factor(c(rep(1,5),rep(2,4),rep(3,5)))

data <- data.frame(Sample_Group, Value)
```

To conduct the Kruskall–Wallis test type:

```
> kruskal.test(Value ~ Sample_Group, data=data)

        Kruskal-Wallis rank sum test

data:  Value by Sample_Group

Kruskal-Wallis chi-squared = 2.2707, df = 2, p-value = 0.3213
```

Since the p-value is greater than 0.05, do not reject the null hypothesis.

Example: using the format kruskal.test(sample, g)

It is also possible to use the format kruskal.test(sample, g)to conduct the test, where sample refers to the sample data and g represents the sample groups or levels. Using the data from the previous example, we would enter it as follows:

```
sample_1 <- c(2.9, 3.5, 2.8, 2.6, 3.7)

sample_2 <- c(3.9, 2.5, 4.3, 2.7)

sample_3<- c(2.9, 2.4, 3.8, 1.2, 2.0)

kruskal.test(list(sample_1, sample_2, sample_3))

sample <- c(sample_1, sample_2, sample_3)

g <- factor(rep(1:3, c(5, 4, 5)),

          labels = c("sample_1",

                     " sample_2",

                     " sample_3"))
```

To conduct the Kruskall–Wallis test type:

```
> kruskal.test(sample, g)

        Kruskal-Wallis rank sum test

data:  sample and g

Kruskal-Wallis chi-squared = 2.2707, df = 2, p-value = 0.3213
```

Since the p-value is greater than 0.05, do not reject the null hypothesis.

References

Aerni, A., Traber, R., Hock, C., Roozendaal, B., Schelling, G.

Papassotiropoulos, A., ... & Dominique, J. F. (2004). Low-dose cortisol for symptoms of posttraumatic stress disorder. American Journal of Psychiatry, 161(8), 1488-1490.

Kennard, D. K., Gould, K., Putz, F. E., Fredericksen, T. S., & Morales, F (2002). Effect of disturbance intensity on regeneration mechanisms in a tropical dry forest. Forest Ecology and Management, 162(2), 197-208.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. Nature, 435(7042), 673-676.

Back to Table of Contents

TEST 56 FRIEDMAN'S TEST

Question the test addresses

Are the distributions from various groups the same across repeated measures?

When to use the test?

It is used for testing the difference between several related samples where observations are repeated on the same subjects. A sample (often referred to as a group) is measured on three or more different occasions where the dependent variable being measured is ordinal, interval or ratio; or for continuous data that has violated the normality assumption and/ or equal variances (of the residuals) necessary to run the one-way ANOVA with repeated measures.

Practical Applications

Horses concept of people: Sankey et al (2011) investigate whether horses are sensitive to the attention state of humans and whether they respond differently to a familiar order when given by a familiar or unknown person. A total of sixteen horses underwent a training program to learn to remain immobile in response to a vocal command. The experimenter (known and unknown) giving the command behaved differently according to four experimental conditions (looking at condition, eyes closed condition, distracted condition, back turned condition). The Friedman and Wilcoxon signed-ranks test was used to compare the duration of immobility across experimental conditions. Horse behavior differed significantly between conditions (Friedman's test p-value = 0.04).

Intrapulmonary arteriovenous pathways and exercise: Lovering et al (2008) study whether breathing 100% oxygen affected intrapulmonary arteriovenous pathways during exercise. Fifteen healthy female subjects aged 19–52 years volunteered to participate in the study. The bubble score as a function of exposure to hyperoxia during exercise was recorded at start, 30, 60 and 120 seconds for each participate. Analysis of bubbles scores was made using a Friedman's test. The researchers observed bubble scores were significantly reduced with 120 seconds of exposure to 100% oxygen (Friedman's test p-value <0.05).

Texas Hold'em poker emotional characteristics: Schlicht et al (2010) investigates whether an opponent's face influences players' wagering decisions in a zero-sum game with hidden information. Fourteen adults participated for monetary compensation. They made risky choices in a

Texas Hold'em style poker game while being presented various opponents faces representing both positive (trust) and negative (threatening) emotional states. A Friedman's test found a significant main effect of trustworthiness on reaction time (p-value =0.03), trustworthiness on correct decisions (p-value = 0.02), A Friedman's test found a significant main effect of trustworthiness on correct decisions (p-value = 0.02), trustworthiness on calling behavior (p-value = 0.01) and trustworthiness on reaction time (p-value = 0.03). The researchers conclude faces relaying positive emotional characteristics impact peoples' decisions; People took significantly longer and made more mistakes against emotionally positive opponents.

How to calculate in R

The function friedman.test{stats}can be used to perform this test. It takes the form friedman.test(Data).

Example: diet and perceived energy

Suppose you wish to examine whether various diets have an effect on perceived energy level. To test this, you recruit 12 healthy individuals who each follow a specific diet for two weeks (healthy balanced, low fat and low carbohydrate) . At the end of the two week period, subjects are asked to record how healthy they feel on a scale of 1 to 10, with 1 being extremely energized and 10 indicating low energy. The resulting scores are given in the below. The first column refers to the individual, the second column balanced diet, third column low fat diet and final column a low carb diet

1	8	8	7
2	7	6	6
3	6	8	6
4	8	9	7
5	5	8	5
6	9	7	7
7	7	7	7
8	8	7	7
9	8	6	8
10	7	6	6

11 7 8 6

12 9 9 6

The data can be entered in R by typing the following:

```
Diet_data<-
matrix(c(8, 8, 7,
         7, 6, 6,
         6, 8, 6,
         8, 9, 7,
         5, 8, 5,
         9, 7, 7,
         7, 7, 7,
         8, 7, 7,
         8, 6, 8,
         7, 6, 6,
         7, 8, 6,
         9, 9, 6
),
       nrow = 12,
       byrow = TRUE,
       dimnames = list(1 : 12,
                       c("Healthy Balanced", "Low Fat", "Low Carb")))
```

Friedman's test can be run by typing:

```
> friedman.test(Diet_data)
        Friedman rank sum test
data:  Diet_data
Friedman chi-squared = 7.6, df = 2, p-value = 0.02237
```

Since the p-value is less than 0.05, we reject the null hypothesis.

References

Lovering, A. T., Stickland, M. K., Amann, M., Murphy, J. C., O'Brien, M. J Hokanson, J. S., & Eldridge, M. W. (2008). Hyperoxia prevents exercise-induced intrapulmonary arteriovenous shunt in healthy humans. The Journal of physiology, 586(18), 4559-4565.

Schlicht, E. J., Shimojo, S., Camerer, C. F., Battaglia, P., & Nakayama, K (2010). Human wagering behavior depends on opponents' faces. PloS one, 5(7), e11663.

Sankey, C., Henry, S., André, N., Richard-Yris, M. A., & Hausberger, M (2011). Do horses have a concept of person?. PloS one, 6(3), e18331.

Back to Table of Contents

TEST 57 QUADE TEST

Question the test addresses

Are the distributions from various groups the same across repeated measures?

When to use the test?

It is used for testing the difference between several related samples where observations are repeated on the same subjects. A sample (often referred to as a group) is measured on three or more different occasions where the dependent variable being measured is ordinal, interval or ratio; or for continuous data that has violated the normality assumption and/ or equal variances (of the residuals) necessary to run the one-way ANOVA with repeated measures. As a simple rule of thumb the Quade test is generally more powerful for a small number of treatments whilst the Friedman test is generally more powerful when the number of treatments is five or more.

Practical Applications

Resource conservation: Wright and Hudson (2013) examine whether conservation of a natural resource, such as a deep portion of an aquifer, could be encouraged, and whether coordination between individuals could be induced. Four treatments were constructed. Participants were faced with a complex bidding process through which units were selected for conservation, and some participants were offered an agglomeration bonus for conserving units that shared a border. The Quade test was used to compare average bids amongst various treatments by non-use value. For a non-use value = 1, a p-value = 0.053 is reported, for non-use value = 3, a p-value of 0.035 is reported; And for non-use value = 5, a p-value of 0.386 is reported. The authors conclude that at least one of the treatments yields larger bids relative to the others.

Social status in spotted Hyenas: The maternal effects on offspring social status in spotted hyenas are investigated by East et al (2009). One of their metrics assessed the closeness between the rank of the adopted offspring at adulthood and the rank of either its genetic mother or surrogate mother at different stages in the adopted individual's development. The mean differences in terms of absolute values between the ranks held by the adopted offspring and the genetic mother at offspring adulthood of 8.5 ± 1.8 rank positions and between the offspring at adulthood and the genetic mother at offspring birth of 9.7 ± 2.0 rank positions were found to be significantly larger than the difference of 1.5 ± 0.4 rank positions between

adopted offspring and surrogate mother when the offspring attained adulthood (Quade test p-value = 0.0014). The researchers observe these results are consistent with the predictions of the behavioral support pathway but not with those of either the direct genetic transfer or endocrine pathways.

Abundance of saproxylic beetles: Hjältén et al (2012) conducted a large-scale field experiment to evaluate the relative importance of manipulated microhabitats, i.e., dead wood substrates of spruce (snags, and logs that were burned, inoculated with wood fungi or shaded) and macrohabitats, i.e., stand types (clear-cuts, mature managed forests, and forest reserves) for species richness, abundance and assemblage composition of all saproxylic and red-listed saproxylic beetles. Beetles were collected in 30 forest stands during the years 2001, 2003, 2004 and 2006. The researchers report the volume of spruce dead wood in decomposition class DC1 (defined as dead wood with bark intact or starting to loosen, 50% bark remaining, wood hard) differed among forest types (Quade test p-value = 0.010). They also found the estimated abundance of red-listed beetles in natural dead wood differed between forest types (Quade test p-value = 0.013).

How to calculate in R

The function quade.test{stats}can be used to perform this test. It takes the form quade.test(data).

Example: diet and perceived energy

Suppose you wish to examine whether various diets have an effect on perceived energy level. To test this, you recruit 12 healthy individuals who each follow a specific diet for two weeks (healthy balanced, low fat and low carbohydrate). At the end of the two week period, subjects are asked to record how healthy they feel on a scale of 1 to 10, with 1 being extremely energized and 10 indicating low energy. The resulting scores are given below. The first column refers to the individual, the second column balanced diet, third column low fat diet and final column a low carb diet

1	8	8	7
2	7	6	6
3	6	8	6
4	8	9	7

5	5	8	5
6	9	7	7
7	7	7	7
8	8	7	7
9	8	6	8
10	7	6	6
11	7	8	6
12	9	9	6

The data can be entered in R by typing the following:

```
Diet_data<-
matrix(c(8, 8, 7,
         7, 6, 6,
         6, 8, 6,
         8, 9, 7,
         5, 8, 5,
         9, 7, 7,
         7, 7, 7,
         8, 7, 7,
         8, 6, 8,
         7, 6, 6,
         7, 8, 6,
         9, 9, 6
),
       nrow = 12,
       byrow = TRUE,
       dimnames = list(1 : 12,
                       c("Healthy Balanced", "Low Fat", "Low Carb")))
```

The Quade test can be run by typing:

```
> quade.test(Diet_data)

        Quade test

data:  Diet_data

Quade F = 3.9057, num df = 2, denom df = 22, p-value = 0.03535
```

Since the p-value is less than 0.05, we reject the null hypothesis.

References

East, M. L., Höner, O. P., Wachter, B., Wilhelm, K., Burke, T., & Hofer, H. (2009). Maternal effects on offspring social status in spotted hyenas. Behavioral Ecology, 20(3), 478-483.

Hjältén, J., Stenbacka, F., Pettersson, R. B., Gibb, H., Johansson, T., Danell K., ... & Hilszczański, J. (2012). Micro and Macro-Habitat Associations ir Saproxylic Beetles: Implications for Biodiversity Management. PloS one 7(7), e41100.

Wright, A. P., & Hudson, D. (2013). Applying a Voluntary Incentive Mechanism to the Problem of Groundwater Conservation: An Experimenta Approach. In 2013 Annual Meeting, February 2-5, 2013, Orlando, Florida (No. 143030). Southern Agricultural Economics Association.

Back to Table of Contents

TEST 58 D' AGOSTINO TEST OF SKEWNESS

Question the test addresses

Is the sample skewed?

When to use the test?

To test for a lack of symmetry (skewness) in a sample. Under the hypothesis of normality, data should be symmetrical (i.e. skewness should be equal to zero). The test is useful for detecting nonnormality caused by asymmetry. If a distribution has normal kurtosis but is skewed, the test for skewness may be more powerful than the Shapiro-Wilk test, especially if the skewness is mild.

Practical Applications

Belgrade Stock returns: Djorić and Nikolić-Djorić (2011) investigate the distributions of daily log returns of the Belgrade Stock Exchange index BELEX15. The BELEX15 index is composed of 15 of the most liquid Serbia shares. The sample period covers 1067 trading days from 4 October 2005 to 25 December 2009. Visual inspection of returns indicates the variances may change over time around some level, with large (small) changes tending to be followed by large (small) changes of either sign (volatility tends to cluster). In order to investigate the asymmetry of the data the researchers perform the D'Agostino test of skewness (p-value = 0.1238). The null hypothesis of symmetry was not rejected using this test.

Spatial-spectral algorithms: Spatial-spectral algorithms were developed by Webster et al (2011) for applying automated pattern recognition morphometric image analysis to quantify histologic tumor and nontumor tissue areas in biospecimen tissue sections. The researchers found lymphoma and melanoma tumor area content distributions exhibited negative skewness (p-value <0.05, D'Agostino test) and the distribution of tumor area percentages in osteosarcoma patients did not reject the null hypothesis due to deviation from symmetry (p-value = 0.3, D'Agostino test).

Visual field decay: Caprioli et al (2011) measure the rate of visual field (VF) decay in in 389 glaucoma patients. Based on an exponential model, global rates of VF decay for each eye were observed to be skewed to the right (D'Agostino's test p-value < 0.0001). The researchers conclude this is consistent with an overall worsening of VFs over the course of follow-up.

How to calculate in R

The function agostino.test{moments} can be used to perform this test. It takes the form agostino.test(sample, alternative = "two.sided" or "less"or "greater")).

Example:

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> agostino.test(sample, alternative = "two.sided")

D'Agostino skewness test

data: sample

skew = 0.3527, z = 0.5595, p-value = 0.5758

alternative hypothesis: data have a skewness

The two sided p-value at 0.5758 is greater than 0.05, therefore do not reject the null hypothesis, the data are not skewed.

References

Caprioli, J., Mock, D., Bitrian, E., Afifi, A. A., Yu, F., Nouri-Mahdavi, K., & Coleman, A. L. (2011). A method to measure and predict rates of regiona visual field decay in glaucoma. Investigative ophthalmology & visual science, 52(7).

Djorić, D., & Nikolić-Djorić, E. (2011). Return distribution and value at risk estimation for BELEX15. The Yugoslav Journal of Operations Research 21(1).

Webster, J. D., Simpson, E. R., Michalowski, A. M., Hoover, S. B., & Simpson R. M. (2011). Quantifying histological features of cancer biospecimens for biobanking quality assurance using automated morphometric pattern recognition image analysis algorithms. Journal of biomolecular techniques: JBT, 22(3), 108.

Back to Table of Contents

TEST 59 ANSCOMBE-GLYNN TEST OF KURTOSIS

Question the test addresses

Does the sample exhibit more (or less) kurtosis relative to the normal distribution?

When to use the test?

The test is useful for detecting nonnormality caused by tail heaviness. If a distribution is symmetric but heavy-tailed (positive kurtosis), the test for kurtosis may be more powerful than the Shapiro-Wilk test, especially if the heavy-tailedness is not extreme.

Practical Applications

Distribution of Earth Orientation Parameters: The Universal Time UT1-UTC together with the pole coordinates (x, y) and celestial pole offsets (dX, dY) are known as Earth Orientation Parameters (EOPs). The EOPs are used to perform transformation between the terrestrial reference frame and the celestial reference frame; and is of great importance for the purpose of navigation and tracking objects in space. Niedzielski, Sen and Kosek (2011) examine the empirical probability distributions of the EOP time series over the time interval from 01.01.1962 to 31.12.2008. The test by Anscombe and Glynn (1983) was used to assess kurtosis (p-value <0.01 for UT1-UTC, x,y and dX, dY). The researchers conclude is apparent that the empirical distributions contain significant kurtosis.

Kurtosis in osteosarcomas: Webster et al (2011) study the distribution of routinely processed tissue sections of osteosarcomas from 43 patients. Useable measurements were acquired successfully for 76/77 (98.7%) of the osteosarcomas. The researchers observe the distribution of tumor area percentages in osteosarcoma is nonnormal (Shapiro-Wilk test p-value <0.05); however, this could not be explained, at the 10% level of significance by deviation from symmetry (D'Agostino skewness test p-value =0.3). However, it could be explained by excessive kurtosis at the 10% level of significance, but not at the 5% level (Anscombe-Glynn test p-value = 0.06).

US output growth-rate: Fagiolo et al (2008) investigate the distribution of US output growth-rate time series. They use quarterly real Gross Domestic Product (GDP) from 1947Q1 to 2005Q3 (234 observations); monthly industrial production (IP) from January 1921 to October 2005 (1017 observations); and they also look at industrial production (IPS) in the sub-period 1947 to 2005 (702 observations). The Anscombe-Glynn test of

kurtosis is applied to each series (GDP p-value = 0.0036, IP p-value <0.001 IPS p-value <0.001). The researchers conclude the growth-rate distributions are markedly nonnormal due to excess kurtosis.

How to calculate in R

The function anscombe.test{moments} can be used to perform this test. It takes the form anscombe.test (sample, alternative = "two.sided" or "less"or "greater")).

Example:

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> anscombe.test (sample, alternative = "two.sided")

Anscombe-Glynn kurtosis test

data: sample

kurt = 2.5504, z = -0.1187, p-value = 0.9055

alternative hypothesis: kurtosis is not equal to 3

The two sided p-value at 0.9055 is greater than 0.05, therefore do not reject the null hypothesis, the data do not exhibit excess kurtosis relative to the normal distribution.

References

Fagiolo, G., Napoletano, M., & Roventini, A. (2008). Are output growth-rate distributions fat-tailed? some evidence from OECD countries. Journal of Applied Econometrics, 23(5), 639-669.

Niedzielski, T., Sen, A. K., & Kosek, W. (2009). On the probability distribution of Earth Orientation Parameters data. Artificial Satellites, 44(1), 33-41.

Webster, J. D., Simpson, E. R., Michalowski, A. M., Hoover, S. B., & Simpson R. M. (2011). Quantifying histological features of cancer biospecimens for biobanking quality assurance using automated morphometric pattern recognition image analysis algorithms. Journal of biomolecular techniques:

JBT, 22(3), 108.

Back to Table of Contents

TEST 60 BONETT-SEIER TEST OF KURTOSIS

Question the test addresses

Does the sample exhibit more (or less) kurtosis calculated by Geary's measure, relative to the normal distribution?

When to use the test?

To test for heavy tails (kurtosis) in a sample. This test uses Geary's measure of kurtosis for normally distributed data. Under the null hypothesis of normality the data should have Geary's kurtosis equal to 0.7979.

Practical Applications

Craniovertebral angle of male Siamese fighting fish: Takeuchi (2010) study the relationship between lateralized eye use during aggressive displays of male Siamese fighting fish (Betta splendens), toward their own mirror image and morphological asymmetry. A total of 25 fish were used in the experiment. Fish were introduced one at a time to an octagonal shaped experimental tank lined with mirrors. Observations were made by the researchers on the aggressive displays by the fish along the mirrored wall. Following the behavioral experiment the researchers constructed an asymmetry index for fish head incline and an asymmetry index of opercula. The mean craniovertebral angle was approximately zero degrees with excess kurtosis at the 5% level of significance (Bonett–Seier test p-value = 0.040).

Ohio hemlock ravine forest ecosystems: Martin and Goebel (2011) document the community composition in southeastern Ohio hemlock ravine forest ecosystems. Sites were sampled within Lake Katharine State Nature Preserve in Jackson County. At each of the eight study sites, three transects were established parallel to the stream at 10, 30, and 50 meters from the stream bank. In each transect, the researchers used a series of five circular plots (5.62-meter radius) for a total of 15 plots per study site. Within each circular plot, physiographical data was recorded, slope percent (using a clinometer), slope shape, slope position, and aspect. All species and diameter at breast height of the woody vegetation were also recorded. Indices of species richness were calculated using Shannon’s diversity and Pielou’s evenness. Neither index exhibited excess levels of Geary's measure of kurtosis (Bonett-Seier test p-value >0.05).

Leaf area fluctuating asymmetry of the common oak: Wuytack (2012) study leaf characteristics of the common oak to monitor ambient ammonia

concentrations. A passive biomonitoring study with common oak at 34 sampling locations in the near vicinity of livestock farms, located in Flanders (northern Belgium) was undertaken. Leaf area fluctuating asymmetry was one of the primary metrics of the study. During the first and second in-leaf season the Bonett-Seier test indicated a leptokurtic distribution (p-value <0.001 for both first and second in leaf season).

How to calculate in R

The function bonett.test{moments} can be used to perform this test. It takes the form bonett.test (sample, alternative = "two.sided" or "less"or "greater")).

Example:

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> bonett.test (sample, alternative = "two.sided")

Bonett-Seier test for Geary kurtosis

data: sample

tau = 0.8400, z = -0.6612, p-value = 0.5085

alternative hypothesis: kurtosis is not equal to sqrt(2/pi)

The two sided p-value at 0.5085 is greater than 0.05, therefore do not reject the null hypothesis, the data do not exhibit excess Geary's measure of kurtosis relative to the normal distribution.

References

Martin, K. L., & Goebel, P. C. (2011). Preparing for hemlock woolly adelgid in Ohio: communities associated with hemlock-dominated ravines of Ohio's Unglaciated Allegheny Plateau. In Proceedings, 17th central hardwood forest conference. General Technical Report NRS-P-78. US Department o Agriculture, Forest Service, Northern Research Station, Newtown Square Pennsylvania (pp. 436-446).

Takeuchi, Y., Hori, M., Myint, O., & Kohda, M. (2010). Lateral bias of agonistic responses to mirror images and morphological asymmetry in the

Siamese fighting fish (Betta splendens). Behavioural brain research, 208(1), 106-111.

Wuytack T (2012) Biomonitoring ambient air quality using leaf characteristics of trees. PhD thesis, University of Antwerp & Ghent University, Belgium.

Back to Table of Contents

TEST 61 SHAPIRO-WILK TEST

Question the test addresses

Is the sample from a normal distribution?

When to use the test?

To investigate whether the observed sample is from a normal distribution. It is used for assessing whether the sample data are randomly obtained from a normally distributed population. It does not require that the mean or variance of the hypothesized normal distribution be specified in advance.

Practical Applications

Maize Yield: Olorede et al (2013) investigate effects of different levels of fertilizer on yield and performance of maize in Nigeria. The researchers use a completely randomized design replicated three times. The residuals from their analysis of variance models were tested against normality using the Shapiro-Wilk test. The researchers report a p-value of 0.6471 for the standardized residuals for "Leave Area", a p-value of 0.5424 for the standardized residuals for "height of Maize", and 0.9836 for the standardized residuals for "Cob and Grain Weight of Maize".

Intrauterine growth: Placental telomere length measurement during ongoing pregnancies complicated by intrauterine growth restriction are reported by Toutain (2013). As part of their study distributions of the quantitative fluorescence In situ hybridization as a function of pregnancy term of placental biopsy were tested for normality with the Shapiro-Wilk test. The researchers report placental telomere fluorescence intensities followed a normal distribution (Shapiro-Wilk test p-value > 0.05).

Load forecasting: Hodge et al (2013) analyzed and characterized the load forecasting errors from two timescales and geographic locations. The data came from two independent system operators in the United States: the California Independent System Operator (CAISO) and the New Yor Independent System Operator (NYISO). Day-ahead and two-day-ahead loa forecasts for each hour of the day, as well as matching actual load data, were obtained for 2010. The forecast error distributions did not follow a normal distribution (Shapiro-Wilk p-value <0.00001 for the CAISO day ahead forecast error, the CAISO two-day-ahead forecast error and the NYISO day-ahead forecast error). The hyperbolic distribution was proposed as a more accurate means of modeling the distribution.

How to calculate in R

The function shapiro.test{stats}can be used to perform this test. It takes the form shapiro.test (sample) . As an alternative the function shapiroTest{fBasics}can also be used, it takes the form shapiroTest(sample).

Example: testing against a normal distribution

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> shapiro.test (sample)

Shapiro-Wilk normality test

data: sample

W = 0.9712, p-value = 0.6767

Since the p-value is greater than 0.05, do not reject the null hypothesis that the data are from the normal distribution. Alternatively, we can try:

> shapiroTest(sample)

Title:

Shapiro - Wilk Normality Test

Test Results:

STATISTIC:

W: 0.9712

P VALUE:

0.6767

The results are identical, and we cannot reject the null hypothesis.

References

Olorede, K. O., Mohammed, I. W., & Adeleke, L. B. (2013). Economi Selection of Efficient Level of NPK 16: 16: 16 Fertilizer for Improved Yield Performance of a Maize Variety in the South Guinea Savannah Zone o

Nigeria. Mathematical Theory and Modeling, 3(1), 27-39.

Hodge, B. M., Lew, D., & Milligan, M. (2013). Short-Term Load Forecasting Error Distributions and Implications for Renewable Integration Studies.

Toutain, J., Prochazkova-Carlotti, M., Cappellen, D., Jarne, A., Chevret, E. Ferrer, J., ... & Saura, R. (2013). Reduced Placental Telomere Length during Pregnancies Complicated by Intrauterine Growth Restriction. PloS one, 8(1) e54013.

Back to Table of Contents

TEST 62 KOLMOGOROV-SMIRNOV TEST OF NORMALITY

Question the test addresses

Is the sample from a normal distribution?

When to use the test?

To investigate whether the observed sample is from a normal distribution. It is used for assessing whether the sample data are randomly obtained from a normally distributed population

Practical Applications

Episodic memory performance: Kinugawa et al (2013) examined episodic memory performance in healthy young (N = 17, age: 21–45), middle-aged (N = 16, age: 48–62) and senior participants (N = 8, age: 71–83) along with measurements of trait and state anxiety. All variables were analyzed with the Kolmogorov–Smirnov normality test to assess whether the data varies significantly from the pattern expected if the data was drawn from a population with a normal distribution. The researchers found the three age groups response to the number of correctly remembered stimulus-position associations did not reject the null hypothesis of being normally distributed (Kolmogorov–Smirnov normality test p-value > 0.05).

Support vector regression: Premanode et al (2013) develop an approach to prediction of foreign exchange time series using support vector regression. Daily trading data for the EUR-USD (euro - US dollar) exchange rate wa collected over the period January 2, 2001 to June 1, 2012. The Kolmogorov-Smirnov test of normality returned a p-value < 0.01 and the null hypothesis was rejected. The researchers propose an Empirical Mode Decomposition (EMD) de-noising model to model exchange rates. The approach uses a sifting process and curve spline technique to decompose a foreign exchange signal into a new oscillatory signal known as an intrinsic mode function (IMF). For each decomposition a number of IMF's are generated The researcher report for IMF number 7, a Kolmogorov-Smirnov test o normality p-value of 0.0593, and the null hypothesis of normality cannot be rejected.

Eye tracking: Kaspar (2013) carry out eye-tracking studies to investigate the influence of the current emotional context on viewing behavior under natural conditions. Participants viewed complex scenes embedded in sequences of emotion-laden images. The researchers find eye-movement

parameters to be normally distributed in all conditions (Kolmogorov-Smirnov normality test p-value ≥0.561 for all samples). Eye-movement parameters on target images embedded into the different emotional contexts were also analyzed. The authors report eye-movement parameters on targets were also normally distributed in all context conditions (Kolmogorov-Smirnov normality test p-value ≥0.238).

How to calculate in R

The function ksnormTest{fBasics}can be used to perform this test. It takes the form ksnormTest(sample). As an alternative the function ks.test{stats} can also be used. It for testing against normality, it takes the form ks.test(sample,"pnorm")

Example: testing against a normal distribution

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> ksnormTest(sample)

Title:

One-sample Kolmogorov-Smirnov test

Test Results:

STATISTIC:

D: 0.1549

P VALUE:

Alternative Two-Sided: 0.5351

Alternative Less: 0.9256

Alternative Greater: 0.2727

The two sided p-value at 0,5351 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution.

As an alternative you can enter:

> ks.test(sample,"pnorm")

One-sample Kolmogorov-Smirnov test

data: sample

D = 0.1549, p-value = 0.5351

alternative hypothesis: two-sided

Again the two sided p-value at 0,5351 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution.

References

Kaspar, K., Hloucal, T. M., Kriz, J., Canzler, S., Gameiro, R. R., Krapp, V., & König, P. (2013). Emotions' Impact on Viewing Behavior under Natural Conditions. PloS one, 8(1), e52737.

Kinugawa, K., Schumm, S., Pollina, M., Depre, M., Jungbluth, C., Doulazm M., ... & Dere, E. (2013). Aging-related episodic memory decline: are emotions the key?. Frontiers in behavioral neuroscience, 7.

Premanode, B., Vonprasert, J., & Toumazou, C. (2013). Prediction of exchange rates using averaging intrinsic mode function and multiclass support vector regression. Artificial Intelligence Research, 2(2), p47.

Back to Table of Contents

TEST 63 JARQUE-BERA TEST

Question the test addresses

Is the sample from a normal distribution?

When to use the test?

To test of the null hypothesis that the sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.

Practical Applications

Robots, humans and the disposition effect: The disposition effect, the fact that investors seem to hold on to their losing stocks to a greater extent than they hold on to their winning stocks, is explored by Da Costa et al (2013). Three groups are exposed to a simulated stock market – experienced investors, inexperienced and robots which make random trade decisions. The Jarque Bera test was used to assess the normality of the disposition effect for each of the three groups. The researchers rejected the null hypothesis of normality for the robots (p-value < 0.05), although it was not rejected for experienced investors and inexperienced investors (p-value > 0.05 for both groups).

Indian foreign investment inflows: Bhattacharya (2013) studies the relationship between foreign investment inflows and the primary, secondary and tertiary sector of the Indian economy over the period 1996 to 2009. The researcher builds an econometric model (Vector Auto-regression model) and uses the Jarque Bera test to assess the normality of the model residuals. The null hypothesis cannot be rejected (p-value=0.51) so the author concludes the model residual series is normally distributed.

Long memory properties in developed stock markets: The existence of long memory properties in developed stock markets is analyzed by Bhattacharya and Bhattacharya (2013). The daily closing values of the individual indices over the period January 2005 to July 2011 were collected. Daily logarithmic index returns were calculated for ten stock market indices in the Netherlands, Australia, Germany, USA, France, UK, Hong Kong, Japan, New Zealand and Singapore. As part of the analysis the researchers use the Jarque-Bera test. The null hypothesis is rejected to for all ten stock market indices (p-value <0.05). The researchers conclude logarithmic return series of the stock market indices cannot be regarded as normally distributed.

How to calculate in R

The function jarqueberaTest{fBasics}or jarque.bera.test{tseries} can be used to perform this test. It takes the form jarqueberaTest(sample) or jarque.bera.test (sample).

Example: testing against a normal distribution

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> jarqueberaTest(sample)

Title:

Jarque - Bera Normalality Test

Test Results:

STATISTIC:

X-squared: 0.7289

P VALUE:

Asymptotic p Value: 0.6946

Or we can use jarque.bera.test:

> jarque.bera.test (sample)

Jarque Bera Test

data: sample

X-squared = 0.7289, df = 2, p-value = 0.6946

In both cases the two sided p-value at 0.6946 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution.

References

Bhattacharya, M. (2013). Foreign Investment inflows and Sectoral growth pattern in India-An Empirical study. Serbian Journal of Management, 8(1).

Bhattacharya, S. N., & Bhattacharya, M. (2013). Long memory in return

structures from developed markets. Cuadernos de Gestión.

Da Costa Jr, N., Goulart, M., Cupertino, C., Macedo Jr, J., & Da Silva, S (2013). The disposition effect and investor experience. Journal of Banking & Finance.

Back to Table of Contents

TEST 64 D' AGOSTINO TEST

Question the test addresses

Is the sample from a normal distribution?

When to use the test?

To test for nonnormality due to a lack of symmetry (skewness). If a distribution has normal kurtosis but is skewed, the test for skewness may be more powerful than the Shapiro-Wilk test, especially if the skewness is mild.

Practical Applications

Effects of T'ai Chi on balance: Twenty two subjects with mild balance disorders were recruited by Hain et al (1999). There were 5 men and 17 women. They were divided into 3 age groups (20-60 years, 61-75 years, and 76 years and older) containing 6, 7, and 9 subjects, respectively. Each subject participated in a T'ai Chi course, which consisted of 8 one-hour sessions held over 2 months, with 1 meeting per week. Students were asked to practice at home every day for at least 30 minutes, and they were given a practice videotape and written materials that illustrated the exercises. Before and after intervention data was collected via objective tests of balance and subjective tests. The data was tested for normality using the D' Agostino test (p-value >0.05). The researchers conclude the data is distributed normally.

Distribution of fish movement: Using a mark–recapture technique in a small temperate stream, Skalski and Gilliam (2000) explore the movement of four fish species From 15 March 1996 through 15 August 1996 period in Durant Creek, Wake County, North Carolina. The four fish species were bluehead chub, creek chub, rosyside dace and redbreast sunfish. The researchers tested the hypothesis that movement distributions were normal using D'Agostino's test for normality (p-value <0.01 for all four species).The researchers observe the movement distributions had higher peaks and longer tails (Leptokurtosis) than a normal distribution. Bluehead chub, creek chub, and redbreast sunfish movement distributions were significantly more leptokurtic than royside dace.

Cerrado tree and severe fire: Silva et al (2009) assessed the effects of a severe fire on the population structure and spatial distribution of Zanthoxylum rhoifolium, a widespread cerrado tree in Brazil. In total 149 individuals of Zanthoxylum rhoifolium were identified before the fire and 112 after the fire, of which 77 were direct resprouts from burnt saplings.

The researchers tested whether the distribution of the number individuals per plot before and after the fire fit a normal distribution (D'Agostino test p-value < 0.01 and p-value < 0.01 before and after the fire respectively). The researchers conclude the distribution of number of individuals per plot did not fit a normal distribution. The distribution of the height and diameter values before and after the fire were also tested (D'Agostino test p-value < 0.01 and p-value < 0.01 for height and diameter before the fire, and p-value < 0.01 and p-value < 0.01 for height and diameter after the fire respectively). The researchers conclude the distribution of individuals per plot did not fit a normal distribution.

How to calculate in R

The function dagoTest{fBasics}can be used to perform this test. It takes the form dagoTest(sample).

Example: testing against a normal distribution

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> dagoTest(sample)

Title:

D'Agostino Normality Test

Test Results:

STATISTIC:

Chi2 | Omnibus: 0.7348

Z3 | Skewness: 0.8489

Z4 | Kurtosis: -0.1187

P VALUE:

Omnibus Test: 0.6925

Skewness Test: 0.3959

Kurtosis Test: 0.9055

The two sided p-value (Omnibus test) at 0.6925 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution. The test also reports p-values for skewness and kurtosis. In both cases the p-values are greater than 0,05 and so the null hypotheses of skewness and kurtosis cannot be rejected.

References

Hain, T. C., Fuller, L., Weil, L., & Kotsias, J. (1999). Effects of T'ai Chi or balance. Archives of Otolaryngology—Head & Neck Surgery, 125(11), 1191.

Skalski, G. T., & Gilliam, J. F. (2000). Modeling diffusive spread in a heterogeneous population: a movement study with stream fish. Ecology, 81(6), 1685-1700.

Silva, I. A., Valenti, M. W., & Silva-Matos, D. M. (2009). Fire effects on the population structure of Zanthoxylum rhoifolium Lam (Rutaceae) in a Brazilian savanna. Brazilian Journal of Biology, 69(3), 813-818.

Back to Table of Contents

TEST 65 ANDERSON-DARLING TEST OF NORMALITY

Question the test addresses

Is the sample from a normal distribution?

When to use the test?

To test of the null hypothesis that the sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.

Practical Applications

Quality of the mechanized coffee harvesting: The quality of the mechanized harvesting of coffee in the municipality of Patos de Minas, MG, Brazil, was assessed by Cassia et al (2013). The researchers assessed five dimensions of the mechanized harvesting process - the harvested coffee load, stripping efficiency, gathering efficiency, harvested coffee and leaf loss on plants as a result of mechanized harvesting. The Anderson-Darling test was used to assess the normality of these five variables. The researchers observed coffee load and leaf loss were normally distributed (Anderson-Darling test p-value > 0.05). This was not the case for stripping efficiency, gathering efficiency or harvested coffee (Anderson-Darling test p-value <0.05).

Magnetic activity and shift in frequency: Baldner, Bogart and Basu (2011) examine changes in frequency for a large sample of active regions analyzed with data from the Michelson Doppler Imager onboard the SoHO spacecraft, spanning most of solar cycle 23. The relation between magnetic activity and shift in frequency is modeled using linear regression. The Anderson-Darling test is applied to the residuals from the linear best fits model (p-value <0.01). The researchers comment this failure implies either that the errors are non-normal, or that the relation between magnetic activity and shift in frequency is not entirely linear.

Resonance Raman spectroscopy: Scarmo et al (2011) examine the feasibility of using Resonance Raman spectroscopy (RRS) as a method of measuring carotenoid status in skin as a biomarker of fruit/vegetable intake in preschool children. A total of 381 participants were recruited onto the study. The mean RRS score was 20.48, with a standard deviation of 6.68. The Anderson–Darling test for normality was significant (p-value <0.01). However, the researchers suggest the data were approximately normally distributed, with a slight right-skew (skewness = 1.06).

How to calculate in R

The function ad.test{nortest}can be used to perform this test. It takes the form ad.test(sample).

Example: testing against a normal distribution

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> ad.test(sample)

Anderson-Darling normality test

data: sample

A = 0.2058, p-value = 0.8545

The two sided p-value at 0.8545 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution.

References

Baldner, C. S., Bogart, R. S., & Basu, S. (2011). Evidence for solar frequency dependence on sunspot type. The Astrophysical Journal Letters, 733(1), L5.

Cassia, Marcelo Tufaile, et al. "Quality of mechanized coffee harvesting in circular planting system." Ciência Rural 43.1 (2013): 28-34.

Scarmo, S., Henebery, K., Peracchio, H., Cartmel, B., Ermakov, H. L. I. Gellermann, W., ... & Mayne, S. T. (2012). Skin carotenoid status measured by resonance Raman spectroscopy as a biomarker of fruit and vegetable intake in preschool children. European journal of clinical nutrition.

Back to Table of Contents

TEST 66 CRAMER-VON MISES TEST

Question the test addresses

Is the sample from a normal distribution?

When to use the test?

To test of the null hypothesis that the sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.

Practical Applications

Efficiency of China stock market: Liu (2011) investigates market efficiency of the China stock market and Hong Kong stock market from 2002 through 2009. Daily and weekly return data are collected for the Shanghai Stock Index A, Shenzhen Stock Index A, Shanghai Stock Index B, Shenzhen Stoc Index B and Hang Seng index. The normality of the return series are testec using the Cramer-von Misses normality test. For both weekly and daily data across all indices the p-value < 0.001 and the null hypothesis of normality is strongly rejected.

Noncontact ultrasound therapy: In a retrospective study Bell and Cavorsi (2008) investigate the impact of adjunctive noncontact ultrasound therapy on the healing of wounds that fail to progress to healing with conventional wound care alone. The researchers carried out a retrospective review of charts for patients who had received outpatient wound care at the Center for Advanced Wound Care, St Joseph's Medical Center, Reading, Pennsylvania, from January 2005 to December 2006 and who were treated with noncontact ultrasound therapy as an adjunct to conventional wound care. The primary endpoint was the percentage of change in wound area. The Cramer-von Mises normality test (p-value <0.005) indicated significant departure from the normality assumption. Additionally, a visual inspection of the histogram of values for the percentage of reduction in wound area confirmed a significant skewed distribution.

Behavior under tensile fatigue loading: Perrin et al (2005) study fatigue behavior and variability under tensile fatigue loading. They develop two mathematical models that predict the fatigue life under alternative stress loading. In the first model, fatigue life at stress amplitude is represented by a lognormal random variable whose mean and standard deviation depend on stress amplitude. The second model, which does not have a closed form solution, yields an iso-probability number of cycles to failure –stress loading probability curve. The goodness of fit tests for each of the fatigue

models is assessed using the Cramer-von Mises normality test (p-value <0.05 for both models).

How to calculate in R

The function cvm.test{nortest} or cvmTest{fBasics}can be used to perform this test. It takes the form cvm.test(sample) or cvmTest(sample).

Example: testing against a normal distribution

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> cvm.test(sample)

Cramer-von Mises normality test

data: sample

W = 0.0243, p-value = 0.9128

The p-value at 0.9128 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution. Alternatively using cvmTest

> cvmTest(sample)

Title:

Cramer - von Mises Normality Test

Test Results:

STATISTIC:

W: 0.0243

P VALUE:

0.9128

The p-value at 0.9128 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution.

References

Bell, A. L., & Cavorsi, J. (2008). Noncontact ultrasound therapy fo adjunctive treatment of nonhealing wounds: retrospective analysis. Physical Therapy, 88(12), 1517-1524.

Liu, T. (2011). Market Efficiency in China Stock Market and Hong Kong Stock Market. International Research Journal of Finance and Economics, 76, 128-137.

Perrin, F., Sudret, B., Pendola, M., & Lemaire, M. (2005, November) Comparison of two statistical treatments of fatigue test data. In Conf. Fatigue Design, Senlis.

Back to Table of Contents

TEST 67 LILLIEFORS TEST

Question the test addresses

Is the sample from a normal distribution?

When to use the test?

To test of the null hypothesis that the sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.

Practical Applications

Fish yield prediction: Multiple regression analysis and back propagation of the neural networks were used by Laë, Lek and Moreau (1999) to develop stochastic models of fish yield prediction using habitat features for 59 lakes distributed all over Africa and Madagascar. Residuals from the regression model had an average of 1.2 and a standard deviation of 16 with the minimum value of 55.7, and the maximum 39. In order to test the normality of model residuals, the Lilliefors test was applied (p-value 0.15). Residuals from the neural network model had an average of 0.9 and a standard deviation of 29 with the minimum value of 92, and the maximum 100 (Lilliefors test of normality p-value < 0.001).

Mental retardation and self-determined behavior: Wehmeyer et al (1996) asked participants with mental retardation to complete various instruments that measured self-determined behavior. The sample included 407 individuals with mental retardation from ten states in the US. Participants answered seven questions (e.g., self-care, learning, mobility, self-direction, receptive and expressive language, capacity for independent living, and economic self-sufficiency). Participants responded none (0), a little (1), or a lot (2) to each questions. The sample averaged 5.3 points with the median score was 5.0, indicating that the sample was composed primarily of individuals with milder cognitive impairments. A Lilliefors test of normality did not reach significance (p-value> 0.05), indicating the scores were approximately normally distributed.

Medical outcomes study and Human Immunodeficiency Virus: Delate and Coons (2001) examine the ability of the Medical Outcomes Study-Human Immunodeficiency Virus Health Survey and the EuroQol Group's EQ-5D questionnaire to discriminate between subjects in predefined disease-severity groups on the basis of clinical-indicator status (i.e., CD4 cell counts, HIV type 1 RNA copies). Data was obtained from the medical records and instruments completed by 242 Human Immunodeficiency Virus -infected

patients. The distributions of the study variable values were assessed by means of the Lilliefors test of normality and were found to differ significantly from normality (p-value <0.01 in all cases).

How to calculate in R

The function lillie.test{nortest} or lillieTest{fBasics}can be used to perform this test. It takes the form cvm.test(sample) or lillieTest (sample).

Example: testing against a normal distribution

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> lillieTest(sample)

Title:

Lilliefors (KS) Normality Test

Test Results:

STATISTIC:

D: 0.0923

P VALUE:

0.8429

The p-value at 0.8429 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution. Alternatively using lillie.test

> lillie.test(sample)

Lilliefors (Kolmogorov-Smirnov) normality test

data: sample

D = 0.0923, p-value = 0.8429

The p-value at 0.8429 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution.

References

Delate, T., & Coons, S. J. (2001). The use of 2 health-related quality-of-life measures in a sample of persons infected with human immunodeficiency virus. Clinical Infectious Diseases, 32(3), e47-e52.

Laë, R., Lek, S., & Moreau, J. (1999). Predicting fish yield of African lakes using neural networks. Ecological modelling, 120(2), 325-335.

Wehmeyer, M. L., Kelchner, K., & Richards, S. (1996). Essential characteristics of self-determined behavior of individuals with mental retardation. AJMR-American Journal on Mental Retardation, 100(6), 632-642.

Back to Table of Contents

TEST 68 SHAPIRO-FRANCIA TEST

Question the test addresses

Is the sample from a normal distribution?

When to use the test?

To test of the null hypothesis that the sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.

Practical Applications

Movement in normal knees: Vedi et al (1999) study in vivo of meniscal movement in normal knees under load. Using an open MR scanner, they image physiological positions of 16 footballers were scanned moving from full extension to 90 degree flexion in the sagittal and coronal planes. Excursion of the meniscal horns, radial displacement and meniscal height were measured. The difference between meniscal movements in the erect and sitting positions was assessed using the Shapiro-Francia test for normality which showed a normal distribution ($p>0.05$).

Rheumatoid Arthritis Larsen scores: Nine hundred sixty-four patients fulfilling the American College of Rheumatology criteria for the classification of Rheumatoid Arthritis were recruited from the Royal Hallamshire Hospital, Sheffield. Modified Larsen scores of radiographic damage were calculated and analyzed by Marinou et al (2007). The Shapiro-Francia test for normality was applied to the data and showed strong evidence against the assumption of normality for the modified Larsen score distribution (p-value < 0.05).

Breastfeeding at baby friendly hospitals: Merewood et al (2005) analyze breastfeeding data from 32 baby-friendly hospitals in 2001 across the United States to determine whether breastfeeding rates in such hospitals differed from national, regional, and state rates. The authors report the mean breastfeeding initiation rate for the 28 Baby-Friendly hospitals in 2001 was 83.8%, compared with a US breastfeeding initiation rate of 69.5% in 2001. The mean rate of exclusive breastfeeding during the hospital stay was 78.4%, compared with a national mean of 46.3%. The Shapiro-Francia test for normality was used to assess whether the distribution of newborn breastfeeding initiation and exclusivity rates differed significantly from the normal distribution (p-value >0.05).

How to calculate in R

The function sf.test{nortest} or sfTest{fBasics}can be used to perform this test. It takes the form sf.test(sample) or sfTest (sample).

Example: testing against a normal distribution

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> sfTest (sample)

Title:

Shapiro - Francia Normality Test

Test Results:

STATISTIC:

W: 0.9759

P VALUE:

0.7035

The p-value at 0.7035 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution. Alternatively using sf.test

> sf.test(sample)

Shapiro-Francia normality test

data: sample

W = 0.9759, p-value = 0.7035

The p-value at 0.7035 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution.

References

Marinou, I., Healy, J., Mewar, D., Moore, D. J., Dickson, M. C., Binks, M. H., . & Wilson, A. G. (2007). Association of interleukin-6 and interleukin-10 genotypes with radiographic damage in rheumatoid arthritis is dependent on autoantibody status. Arthritis & Rheumatism, 56(8), 2549-2556.

Merewood, A., Mehta, S. D., Chamberlain, L. B., Philipp, B. L., & Bauchne H. (2005). Breastfeeding rates in US Baby-Friendly hospitals: results of a national survey. Pediatrics, 116(3), 628-634.

Vedi, V., Spouse, E., Williams, A., Tennant, S. J., Hunt, D. M., & Gedroyc, W M. W. (1999). Meniscal movement An in-vivo study using dynamic MRI Journal of Bone & Joint Surgery, British Volume, 81(1), 37-41.

Back to Table of Contents

TEST 69 MARDIA'S TEST OF MULTIVARIATE NORMALITY

Question the test addresses

Is my sample of k factors drawn from the multivariate normal distribution?

When to use the test?

Used to test if the null hypothesis of multivariate normality is a reasonable assumption regarding the population distributions of a random sample of k factors. Specifically, if a sample was randomly drawn from a multivariate normal distribution there should be no significant skew, and kurtosis should be that associated with the normal distribution. In this test the skewness and kurtosis are functions of the squared Mahalanobis distances. A large value of multivariate kurtosis, in comparison to the expected value under normality, indicates that one or more observations have a large Mahalanobis distance and are thus located far from the centroid of the data set. This property is useful in multivariate outlier detection.

Practical Applications

Women managers and stress: Long, Kahn and Schutz (1992) developed a model of managerial women's stress. A survey was administered to a total of 249 Canadian women managers. Areas covered in the survey include, personal and job demographics, Sex-role attitudes, agentic traits, aspects of the work environment, work performance, job satisfaction, attitudes toward women distress: anxiety, depression, and somatic symptoms. Mardia's test was used to assess the multivariate normality of the sample. The researchers report a measure of multivariate kurtosis of 1.02 (p –value =0.5), and conclude the data appear not to deviate from an assumed distribution of multivariate normal.

Engineering seismology: Iervolino (2008) study 190 horizontal components from 95 recordings of Italian earthquakes. The researchers focus on the parameters, the peak ground acceleration, peak velocity, Arias Intensity and the Cosenza and the Manfredi index. Mardia's test of multivariate normality was used to assess the joint normality of the logs of the parameters. It resulted in skew =20.03 (p-value < 0.001), kurtosis = -0.61 (p-value <0.01). The null hypothesis of multivariate normality was rejected.

Elders and depression: Gellis (2010) investigate responses to the Center for Epidemiologic Studies depression scale from a cross-sectional survey of elders. The scale is a 20 item index care self-report depression instrument. A total of 618 participants were recruited in order to determine the validity of a shorter version of the depression metric. Analysis consisted of

confirmatory factor and rating scale analysis. Multivariate normality was evaluated using Mardia's test (p-value >0.05).

How to calculate in R

The function mardia{psych} perform this test. It takes the form mardia (multivariate.dataset). The parameter multivariate.dataset refers to a dataframe of you multivariate sample.

Example: the daily difference in European stock prices

Let us try out the test on daily difference in closing prices of major European stock indices. We use the data frame EuStockMarkets which contains daily closing prices for DAX, SMI, CAC , FTSE over the period 199 1998. Since we are interested in daily difference enter:

diff =diff(EuStockMarkets,1)#calculate daily difference

To apply the test type:

> mardia(diff)

Call: mardia(x = diff)

Mardia tests of multivariate skew and kurtosis

Use describe(x) the to get univariate tests

n.obs = 1859 num.vars = 4

b1p = 0.91 skew = 281.49 with probability = 0

small sample skew = 282.12 with probability = 0

b2p = 61.99 kurtosis = 118.22 with probability = 0

The skew (large sample is 281.49) with a p-value = 0. The value of kurtosis is 118.22, with a p-value = 0. Clearly, in this case we can reject the assumption of multivariate normality. Note, use the small sample p-value when you have 30 or less observations.

References

Gellis, Z. D. (2010). Assessment of a brief CES-D measure for depression ir homebound medically ill older adults. Journal of gerontological social work, 53(4), 289-303.

Iervolino, I., Giorgio, M., Galasso, C., & Manfredi, G. (2008, October) Prediction relationships for a vector-valued ground motion intensity

measure accounting for cumulative damage potential. In 14 th World Conference on Earthquake Engineering (pp. 12-17).

Long, B. C., Kahn, S. E., & Schutz, R. W. (1992). Causal model of stress and coping: Women in management. Journal of Counseling Psychology, 39(2), 227.

Back to Table of Contents

TEST 70 KOLOMOGOROV – SMIRNOV TEST FOR GOODNESS OF FIT

Question the test addresses

Is there a significant difference between the observed distribution in a sample and a specified population distribution?

When to use the test?

To compare a random sample with a known reference probability distribution. The test requires no prior assumption about the distribution of data. The test statistic is most sensitive to the region near the mode of the sample distributions, and less sensitive to their tails.

Practical Applications

Automatic detection of influenza epidemics: Closas, Coma and Méndez (2012) develop a statistical method to detect influenza epidemic activity. Non-epidemic incidence rates are modeled against the exponential distribution through a sequential detection algorithm. Detection of weekly incidence rates is assessed by the Kolmogorov-Smirnov test on the absolute difference between the empirical and the cumulative density function of an exponential distribution. The researchers report the Kolmogorov-Smirnov test detected the following weeks as epidemic for each influenza season: 50 − 10 (2008-2009 season), 38 − 50 (2009-2010 season), weeks 50 − 9 (2010-2011 season) and weeks 3 to 12 for the 2011-2012 season. The researchers conclude the proposed test could be applied to other data sets to quickly detect influenza outbreaks.

Is the universe really weakly random? Næss (2012) pick at random 10 000 disks with a radius of 1.5 degrees from the WMAP 7 year W-band map, with the region within 30 degrees from the galactic equator excluded. Each disk contains on average 540 pixels, which are whitened using the author's model. After whitening, the values should follow the standard normal distribution. The author test this assumption using the Kolmogorov-Smirnov test (p-value >0.05). They cannot reject the null hypothesis.

Ovarian-cancer specimens: Merritt et al (2008) examined 111 ovarian-cancer specimens using quantitative reverse-transcriptase–polymerase-chain-reaction for mRNA and calculated the ratios of the expression in the tumors. The distribution of Dicer mRNA levels in the ovarian-cancer specimens were not normally distributed (Kolmogorov–Smirnov test for normality p-value = 0.002). The researchers observe the distribution was

bimodal.

How to calculate in R

The function ks.test{stats}can be used to perform this test. It takes the form ks.test(sample, “cumulative_probability”,alternative = "two.sided"). Note to specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater"). A range of common probability distributions are given below, alongside their name in R.

Beta R-code = pbeta

Lognormal R-code = plnorm

Binomial pbinom R-code =

Negative Binomial R-code = pnbinom

Cauchy R-code = pcauchy

Normal R-code = pnorm

Chisquare R-code = pchisq

Poisson R-code = ppois

Exponential R-code = pexp

Student t R-code = pt

F R-code = pf

Uniform R-code = punif

Gamma R-code = pgamma

Tukey R-code = ptukey

Geometric R-code = pgeom

Weibull R-code = pweib

Hypergeometric R-code = phyper

Wilcoxon R-code = pwilcox

Logistic R-code = plogis

Example: testing against a normal distribution

Enter the following data:

```
> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)
```

The test can be conducted as follows:

```
> ks.test(sample,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  sample

D = 0.1549, p-value = 0.5351

alternative hypothesis: two-sided
```

Since the p-value is greater than 0.05, do not reject the null hypothesis that the data are from the normal distribution.

Example: testing against an exponential distribution

Using the data from the previous example, enter:

```
> ks.test(sample,"pexp")

        One-sample Kolmogorov-Smirnov test

data:  sample

D = 0.59, p-value = 8.04e-09

alternative hypothesis: two-sided
```

Since the p-value is less than 0.05, do reject the null hypothesis that the data come from the exponential distribution.

References

Closas, P., Coma, E., & Méndez, L. (2012). Sequential detection of influenza epidemics by the Kolmogorov-Smirnov test. BMC Medical Informatics and Decision Making, 12(1), 112.

Merritt, W. M., Lin, Y. G., Han, L. Y., Kamat, A. A., Spannuth, W. A., Schmandt, R., ... & Sood, A. K. (2008). Dicer, Drosha, and outcomes in patients with ovarian cancer. New England Journal of Medicine, 359(25), 2641-2650.

Næss, S. K. (2012). Application of the Kolmogorov-Smirnov test to CME

data: Is the universe really weakly random?. Astronomy & Astrophysics, 538.

Back to Table of Contents

TEST 71 ANDERSON-DARLING GOODNESS OF FIT TEST

Question the test addresses

Is there a significant difference between the observed distribution in a sample and a specified population distribution?

When to use the test?

To investigate the null hypothesis that a sample is from a specific distribution. The test compares the fit of an observed cumulative distribution function to a specific cumulative distribution function. It is a modification of the Kolmogorov-Smirnov test giving more weight to the tails of the distribution. Since the test makes use of a specific distribution in calculating critical values it is a more sensitive test than the Kolmogorov-Smirnov test.

Practical Applications

Maximum annual wind speeds in Brazil: Beck and Corrêa (2013) investigate the distribution of maximum annual wind speeds from 104 weather stations over 50 years across Brazil. Individual weather station data was fitted to the Gumbel probability distribution (p-value >0.05 in all cases). The lowest p-values were obtained for the Petrolina and Aracaju weather stations (p-value = 0.14), where basic wind speeds were particularly high. The researchers use wind speeds to build a non-linear regression model, using the p-value of the Anderson-Darling goodness-of-fit test as regression weight. This ensures that extreme value wind distributions for which a higher p-value is obtained are given more importance in the regression model.

Strength and modulus of elasticity of concrete: Kolisko et al (2012) investigates the distribution of the strength and modulus of elasticity of concrete. The sample was obtained in October and November 2010 for a total of 67 prefabricated beams for use in bridges under the management of the Road and Motorway Directorate of the Czech Republic. Cylinders of 150 × 300 mm in size were used to obtain empirical information on strength and modulus of elasticity. The researchers tested the sample using four common probability distributions – normal, lognormal, beta and gamma. Assessment of goodness of fit was made using the Anderson – Darling test. For strength, the researchers report the Beta distribution is the best fit (p-value >0.05). For modulus of elasticity the lognormal distribution is reported as the best fit (p-value >0.05).

Reducing printer paper waste: Hasan et al (2013) study the effect of team-

based feedback on individual printer paper use in an office environment. An email on printer use was sent on a weekly basis to individual participants. The researchers construct a sample based on the difference in printer paper usage before and after the email intervention. In order to check normality of the “difference” sample, the Anderson-Darling test was used (p-value =0.343). The null hypothesis of normality could not be rejected.

How to calculate in R

The function ad.test{ADGofTest} can be used to perform this test. It takes the form ad.test(sample, dist_function). Note dist_function refers to the probability distribution specified under the null hypothesis A range of common probability distributions are given below, alongside their name in R.

Beta R-code = pbeta

Lognormal R-code = plnorm

Binomial pbinom R-code =

Negative Binomial R-code = pnbinom

Cauchy R-code = pcauchy

Normal R-code = pnorm

Chisquare R-code = pchisq

Poisson R-code = ppois

Exponential R-code = pexp

Student t R-code = pt

F R-code = pf

Uniform R-code = punif

Gamma R-code = pgamma

Tukey R-code = ptukey

Geometric R-code = pgeom

Weibull R-code = pweib

Hypergeometric R-code = phyper

Wilcoxon R-code = pwilcox

Logistic R-code = plogis

Example: testing against a normal distribution

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

Let's investigate whether this data are from the lognormal distribution. To do so enter:

> ad.test(sample,plnorm)

Anderson-Darling GoF Test

data: sample and plnorm

AD = Inf, p-value = 2.4e-05

alternative hypothesis: NA

Since the p-value is less than 0.05, reject the null hypothesis that the data are from the lognormal distribution.

References

Beck, A. T., & Corrêa, M. R. (2013). New Design Chart for Basic Wind Speed in Brazil. Latin American Journal of Solids and Structures, 10(4), 707-723.

Hasan, S., Medland, R. C., Foth, M., & Curry, E. (2013). Curbing resourc consumption using team-based feedback: paper printing in a longitudinal case study. In Proceedings of the 8th International Conference on Persuasive Technology. Springer.

Kolisko, J., Hunka, P., & Jung, K. (2012).A Statistical Analysis of the Modulus of Elasticity and Compressive Strength of Concrete C45/55 for Pre-stressed Precast Beams. Journal of Civil Engineering and Architecture. Volume 6, No 11 (Serial No. 60), pp. 1571–1576.

Back to Table of Contents

TEST 72 TWO-SAMPLE KOLMOGOROV-SMIRNOV TEST

Question the test addresses

Do two independent random samples come from the same probability distribution?

When to use the test?

To compare two random samples, in order to determine if they come from the same probability distribution.

Practical Applications

Spectroscopic metallicities: Buchhave et al (2012) analyze spectroscopic metallicities of the host stars of 226 small exoplanet candidates discovered by NASA's Kepler mission. The researchers find smaller planets are observed at a wide range of host-star metallicities, whereas larger planets are detected preferentially around stars with higher metallicity. To investigate the statistical significance of the difference in metallicity, a two-sample Kolmogorov–Smirnov test of the two subsamples of host stars is performed. The probability that the two distributions are not drawn randomly from the same population is calculated to be 99.96%.

Cyclone power dissipation: The power dissipation index (PDI) is an estimate of energy release in individual tropical cyclones. Corral, Ossó and Llebot (2010) calculate PDI in the North Atlantic over the 54-year periods 1900-1953 and 1954-2007, with 436 and 579 storms respectively. A two-sample Kolmogorov-Smirnov test gives a p-value = 0.15, and the null hypothesis cannot be rejected.

Rain Fall: Peters et al (2010) study rain data from all ten diverse locations (Manus, Nauru, Darwin, Niamey, Heselbach, Shouxian, Graciosa Island Point Reye, North Slope of Alaska, Southern Great Plains). A two-sample Kolmogorov-Smirnov test for all pairs of datasets was carried out. The two-sample Kolmogorov-Smirnov test p-value for the samples Manus and Nauru was greater than 0.1. The authors comment that this confirms the similarity of the distributions from these two sites.

How to calculate in R

The function ks.test{stats}can be used to perform this test. It takes the form ks.test(sample1,sample2, alternative ="two.sided"). Note to specify the alternative hypothesis of greater than (or less than) use alternative ="less" (alternative = "greater"). As an alternative the function

ks2Test{fBasics} can also be used. It takes the form ks2Test(sample1,sample2).

Example: testing against a normal distribution

Enter the following data:

sample1<- c(-2.12, 0.08, -1.59, -0.15, 0.9, -0.7, -0.22, -0.66, -2.14, 0.65, 1.38, 0.27, 3.33, 0.09, 1.45, 2.43, -0.55, -0.68, -0.62, -1.91, 1.11, 0.43, 0.42, 0.09, 0.76)

sample2<- c(0.91, 0.89, 0.6, -1.31, 1.07, -0.11, -1.1, -0.83, 0.8, -0.53, 0.3, 1.05, 0.35, 1.73, 0.09, -0.51, -0.95, -0.29, 1.35, 0.51, 0.66, -0.56, -0.04, 1.03, 1.47)

The test can be conducted as follows:

```
> ks.test(sample1,sample2,alternative="two.sided")

        Two-sample Kolmogorov-Smirnov test

data:  sample1 and sample2

D = 0.16, p-value = 0.9062

alternative hypothesis: two-sided
```

Since the p-value is greater than 0.05, do not reject the null hypothesis. We could also use:

```
> ks2Test(sample1,sample2)

Title:

 Kolmogorov-Smirnov Two Sample Test

Test Results:

  STATISTIC:

    D | Two Sided: 0.16

   D^- | Less: 0.08

   D^+ | Greater: 0.16

  P VALUE:

    Alternative       Two-Sided: 0.9062

    Alternative Exact Two-Sided: 0.9062
```

Alternative Less: 0.8521

Alternative Greater: 0.5273

Again the two sided p-value is greater than 0.05, do not reject the null hypothesis.

References

Buchhave, L. A., Latham, D. W., Johansen, A., Bizzarro, M., Torres, G., Rowe J. F., ... & Quinn, S. N. (2012). An abundance of small exoplanets around stars with a wide range of metallicities. Nature, 486(7403), 375-377.

Corral, Á., Ossó, A., & Llebot, J. E. (2010). Scaling of tropical-cyclone dissipation. Nature Physics, 6(9), 693-696.

Peters, O., Deluca, A., Corral, A., Neelin, J. D., & Holloway, C. E. (2010) Universality of rain event size distributions. Journal of Statistical Mechanics: Theory and Experiment, 2010(11), P11030.

Back to Table of Contents

TEST 73 ANDERSON-DARLING MULTIPLE SAMPLE GOODNESS OF FIT TEST

Question the test addresses

Is there a significant difference between the observed distributions in k distinct samples?

When to use the test?

To compare the paired empirical distribution functions of multiple samples. The test does not assume equal variances. The test evaluates the more general null hypothesis that all samples have the same distribution against the alternative that the samples differ in central tendency and/or in variability.

Practical Applications

Transmission of mumps: Fanoy et al (2011) compared mumps viral titers of oral fluid specimens from 60 vaccinated subjects and 110 unvaccinated mumps patients. The sample data was stratified by the time elapsed since onset of disease (≤3 days, >3 and <6 days, ≥6 days). The Anderson-Darling multiple sample goodness of fit test was used to assess the effect of a previous measles, mumps, and rubella vaccination history on the amount of virus detected in the specimens taking into account the time elapsed since onset. The researchers observe the difference between the two groups with samples taken within 3 days after the onset of disease is statistically significant (p-value < 0.01). The difference between the two groups sampled after three and before 6 days was also significant (p-value = 0.01). However, no significant difference appeared among the patients who provided samples 6 or more days after the onset of disease.

Fecal coliform and Escherichia coli in Oregon: Cude (2005) develop a relationship between fecal coliform and Escherichia coli in the context of the Oregon Water Quality Index (OWQI). The OWQI is a primary indicator o general water quality for the Oregon Department of Environmental Quality. Data was collected from long term monitoring stations located in a variety of regions and land uses throughout Oregon. A bacterial sub-index (SIBACT and OWQI values were calculated using paired measurements of Escherichia coli. The Anderson-Darling multiple sample goodness of fit test was used to compare the paired empirical distribution functions (EDF). In all cases (paired SIBACT and paired OWQI EDFs), the null hypothesis wa rejected (p-value < 0.01).

Fish harvest in regulatory area 3a: Meyer (1995) analyzed age, length and sex composition alongside other fishery statistics for the recreational harvest of Pacific halibut in international Pacific halibut commission regulatory Area 3A in 1994. Samples were taken from catches in various areas (Kodiak, Homer, Seward, Valdex, Anchor Point) and at different times of the year. For example in Kodiak, Homer (fish cleaned in port), Seaward and Valdez five samples were obtained between May and September. For Homer (charter with fish cleaned at sea) three samples were taken between July and August. The Anderson-Darling multiple sample goodness of fit test is used to assess differences in the distribution of the length of fish caught within an area. For Kodiak's five samples (p-value <0.01); for Homer (fish cleaned in port) (p-value =0.57) ; for Homer (charter with fish cleaned at sea) (p-value =0.37);for Seaward (p-value =0.26); for Valdez (p-value <0.01).

How to calculate in R

The function adk.test{adk}can be used to perform this test. It takes the form ad.test(sample.1, sample.2,… sample.k).

Example: The distribution of the difference daily differences in stock indices

Let's apply the test to the daily closing first difference of the DAX, SMI,CAC and FTSE stock market indices using data from 1991-1998. This data is contained in the dataframe EuStockMarkets:

```
DAX<-diff(EuStockMarkets[,1],1)

SMI<- diff(EuStockMarkets[,2],1)

CAC<- diff(EuStockMarkets[,3],1)

FTSE<- diff(EuStockMarkets[,4],1)

> adk.test(DAX,SMI,CAC,FTSE)

Anderson-Darling k-sample test.

Number of samples:  4

Sample sizes: 1859 1859 1859 1859

Total number of values: 7436

Number of unique values: 3793

Mean of Anderson-Darling Criterion: 3
```

Standard deviation of Anderson-Darling Criterion: 1.31827

T.AD = (Anderson-Darling Criterion - mean)/sigma

Null Hypothesis: All samples come from a common population.

	t.obs	P-value	extrapolation
not adj. for ties	9.31536	1e-05	1
adj. for ties	9.31007	1e-05	1

Since the p-value is less than 0.05, reject the null hypothesis that the daily difference in stock prices are from a common distribution.

References

Cude, C. G. (2005). Accommodating change of bacterial indicators in long term water quality datasets. Journal of the American Water Resources Association, 41(1), 47-54.

Fanoy, E., Cremer, J., Ferreira, J., Dittrich, S., van Lier, A., Hahné, S., ... & van Binnendijk, R. (2011). Transmission of mumps virus from mumps-vaccinated individuals to close contacts. Vaccine.

Meyer, S. C. (1995). Recreational halibut fishery statistics for southcentral Alaska (Area 3A), 1994. A report to the International Pacific Halibut Commission. Alaska Department of Fish and Game, Special Publication, (96 1).

Back to Table of Contents

TEST 74 BRUNNER-MUNZEL GENERALIZED WILCOXON TEST

Question the test addresses

Are the scores on some ordinally scaled variable larger in one population than in another?

When to use the test?

To test for stochastic equality i.e. $P(X < Y) = P(X > Y)$. The test should be applied when it cannot be assumed that variances are equal and that the distribution is non-symmetric (skewed). It was designed to detect differences between groups without making any assumptions regarding the shape or continuity of the underlying distribution. The test is generally preferable to a transformation of the data, especially when dealing with a small sample size.

Practical Applications

Verb and noun naming deficits in Alzheimer's Disease: Almor et al (2009) address the question is verb performance in AD compatible with graceful degradation in a general feature based framework in terms of error pattern progression? Fourteen patients with Alzheimer's Disease (AD) and fourteen healthy elderly normal controls (EN) participated in this study. The two groups were matched for age, and years of education. Participants from each group performed a verb naming task and a noun naming task first. Error percentages for each group were calculated and the Brunner-Munzel test was used to compare the ranking of the errors made by the two groups (p-value <0.001). The researchers conclude the ranking of errors was higher for the AD patients than for the EN group.

Pre–whole-genome duplication yeast: Wang et al (2011) used the reconstructed gene order of the pre–whole-genome duplication yeast ancestor to compare the co-expression of gene pairs that are conserved between the ancestor and Saccharomyces cerevisiae with the co-expression of gene pairs newly formed in S. cerevisiae. The researchers define co-expression of two genes as the correlation of gene expression values across a large data set of time series experiments. No difference is observed between the co-expression of newly formed divergent gene pairs and convergent gene pairs (Brunner–Munzel p-value = 0.59 comparing new divergent gene pairs with conserved convergent gene; Brunner–Munzel pairs and p-value = 0.59 comparing new divergent gene pairs with newly formed convergent gene pairs). The researchers conclude divergent gene

pairs do not always show higher co-expression compared with other types of adjacent gene pairs in yeast.

Rock ptarmigan: One hundred rock ptarmigan (Lagopus muta), including 30 each of juvenile males and females, and 20 each of adult males and females, were collected in October 2006 in northeast Iceland by Skirnisson et al (2012) to study their parasite fauna. Blastocystis sp was identified as one of many parasite species. The prevalence of Blastocystis sp. was 91%; all adults were infected, and the prevalence in juveniles was 85%. Ranked values for mean intensity indicated no difference among host age groups (Brunner–Munzel test, p-value =0.06).

How to calculate in R

The function brunner.munzel.test{lawstat} can be used to perform this test. It takes the form brunner.munzel.test(x, y, alternative = "two.sided", alpha=0.05)

Note to conduct a one sided test set alternative = “less” or alternative =”greater”.

Example:

Suppose you have collected the ordinal scores from two groups of football fans, at the end of a football game which ended 0-0.

ordinal.score1<-c(2,2,4,1,1,4,1,3,1,5,2,4,1,1)

ordinal.score2<-c(3,3,4,3,1,2,3,3,1,5,4)

The test can be carried out as follows

> brunner.munzel.test(ordinal.score1, ordinal.score2, alternative = "two.sided", alpha=0.05)

Brunner-Munzel Test

data: ordinal.score1 and ordinal.score2

Brunner-Munzel Test Statistic = 1.1588, df = 22.72, p-value = 0.2586

95 percent confidence interval:

0.3953241 0.8709097

sample estimates:

P(X<Y)+.5*P(X=Y)

0.6331169

Since the p-value is greater than 0.05, do not reject the null hypothesis.

References

Almor, A., Aronoff, J. M., MacDonald, M. C., Gonnerman, L. M., Kempler, D. Hintiryan, H., ... & Andersen, E. S. (2009). A common mechanism in verb and noun naming deficits in Alzheimer's patients. Brain and language, 111(1), 8-19.

Skirnisson, K., Thorarinsdottir, S. T., & Nielsen, O. K. (2012). The Parasite Fauna of Rock Ptarmigan (Lagopus muta) in Iceland: Prevalence, Intensity and Distribution Within the Host Population. Comparative Parasitology, 79(1), 44-55.

Wang, G. Z., Chen, W. H., & Lercher, M. J. (2011). Coexpression of Linked Gene Pairs Persists Long after Their Separation. Genome Biology and Evolution, 3, 565.

Back to Table of Contents

TEST 75 DIXON'S Q TEST

Question the test addresses

Do my sample data contain an outlier?

When to use the test?

To investigate if one (and only one) observation from a small sample (typically less than 30 observations) is an outlier. Normal distribution of the sample data is assumed whenever this test is applied. If an outlier has been detected the test should not be reapplied on the set of the remaining observations.

Practical Applications

Moth diet and fitness: In order to assess adult fitness components Cogni et al (2012) added pyrrolizidine alkaloids to an artificial diet at different concentrations fed to the moth Utetheisa ornatrix (Lepidoptera: Arctiidae). A small sample of twenty adults per treatment were used by the researchers. Three replicate spectrophotometer readings were performed for each individual, and the average was used as the dependent variable in further analysis. Dixon's Q-test (p-value <0.05) was used to detect possible outliers among the three replicated readings.

Oyster mushroom production: Oyster mushroom cultivated in banana straw using inocula produced by two different processes - liquid inoculum and solid inoculum was studied by Silveira et al (2008). Different ratios (5, 10, 15, and 20%) were tested. Biological efficiency, yield, productivity, organic matter loss, and moisture of fruiting bodies as well as physical-chemical characteristics of banana straw were analyzed for each ratio and process. Dixon's Q-test was performed to statistically reject outliers (p-value <0.1).

Metabolic syndrome: Sharma et al (2011) compared the use of homeostasis model assessment of insulin resistance with the use of fasting blood glucose to identify metabolic syndrome in African American children. Anthropometric, biochemical and blood pressure measurements were obtained for 108 children. The measurements were first assessed for skewedness and, if significant, Dixon's test for outliers was used to identify unusual values (p-value <0.05). If unusual values were identified, all data for that participant were excluded from further analyses. Using Dixon's test, the researchers exclude 3 children, resulting in a final sample of 105 (45 boys and 60 girls).

How to calculate in R

The function dixon.test{outliers} can be used to perform this test takes the form dixon.test (sample). The parameter sample refers to sample observations to be used in the test.

Example:

Enter the following data, collected, on two variables:

sample<-c(0.189,0.167,0.187,0.183,0.186,0.182,0.181,0.184,0.177)

To perform the test on the smallest value in the sample enter:

> dixon.test(x)

Dixon test for outliers

data: x

Q = 0.5, p-value = 0.1137

alternative hypothesis: lowest value 0.167 is an outlier

The test reports that the p-value on the smallest observation is not significant. As an alternative we can perform the test on the largest observation, in which case you would type:

> dixon.test(x,opposite=TRUE)

Dixon test for outliers

data: x

Q = 0.1667, p-value = 0.8924

alternative hypothesis: highest value 0.189 is an outlier

The test reports that the p-value on the largest observation is not significant.

References

Cogni, R., Trigo, J. R., & Futuyma, D. J. (2012). A free lunch? No cost fo acquiring defensive plant pyrrolizidine alkaloids in a specialist arctiid moth (Utetheisa ornatrix). Molecular ecology, 21(24), 6152-6162.

Sharma, S., Lustig, R. H., & Fleming, S. E. (2011). Peer Reviewed: Identifyin Metabolic Syndrome in African American Children Using Fasting HOMA-I in Place of Glucose. Preventing Chronic Disease, 8(3).

Silveira, M. L. L., Furlan, S. A., & Ninow, J. L. (2008). Development of a alternative technology for the oyster mushroom production using liquid inoculum. Ciência e Tecnologia de Alimentos, 28(4), 858-862.

Back to Table of Contents

TEST 76 CHI-SQUARED TEST FOR OUTLIERS

Question the test addresses

Do my sample data contain an outlier?

When to use the test?

This test requires specification of the population variance. If you do not know the population variance this test statistic should not be used as it is based on the chi-squared distribution of squared differences between data and sample mean, and therefore likely only to reject extreme outliers.

Practical Applications

Vampire calls: Carter et al (2012) investigate whether isolated adult vampire bats produce vocally distinct contact calls when physically isolated. Desmodus vampire bats and Diphylla vampire bats, were placed in physical isolation for up to 24 hours and their calls recorded. The chi-square outlier test was used to assess whether a single Diaemus individual from a different population produced a first note distinct (chi-square outlier test: p-value = 0.025). The researchers observe this discrepancy disappeared when including first and second note, suggesting that the second note in this individual contained much of the signature information. This observation, the researchers suggests, highlights the fact that double-note call structures allow for substantial increases in potential information content.

Oxidoreductase Activity: Gregor et al (2013) study oxygen consumption by enzymatic reactions. Open-system experiments were carried out in two kinds of devices - 0a classical Clark electrode and the novel extracellular flux analyzer. The oxygen consumption rates obtained by both the classical and the novel approach were compared. Outlier exclusion was performed using the Chi-squared test for outliers (p-value <0.05).

Pollination of Podostemaceae: In order to analyze the reproductive system of Mourera fluviatilis (Podostemaceae), Sobral-Leite et al (2011) carryout field experiments involving manual pollination, self-pollination and cross-pollination. A total of 30 and 100 randomly selected individuals were marked, and between one to three flowers per individual for each treatment. Seeds were counted under a stereomicroscope using graph paper and a manual counter. Outliers were tested using the Chi–squared test for outliers and removed from the analysis (p-value <0.05).

How to calculate in R

The function chisq.out.test{outliers} can be used to perform this test takes the form chisq.out.test(data, variance=1). The parameter variance refers to the known population variance.

Example:

Enter the following data, collected, on two variables:

dependent.variable=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,

independent.variable=c(75,78,80,82,84,88,93,97,65,104,109,115,120,127)

To carry out the test enter on the residual of a regression model enter

> regression.model<- lm(dependent.variable ~ independent.variable)

> residual<-rstudent(regression.model)

> regression.model<- lm(dependent.variable ~ independent.variable)

> residual<-rstudent(regression.model)

> chisq.out.test(residual, variance=1)

chi-squared test for outlier

data: residual

X-squared.9 = 1850.439, p-value < 2.2e-16

alternative hypothesis: highest value 46.096060249773 is an outlier

The function identifies 46.09606 as an outlier with a p-value less than 0.01.

References

Carter GG, Logsdon R, Arnold BD, Menchaca A, Medellin RA (2012) Adu Vampire Bats Produce Contact Calls When Isolated: Acoustic Variation by Species, Population, Colony, and Individual. PLoS ONE 7(6): e38791 doi:10.1371/journal.pone.0038791

Gregor Hommes, Christoph A. Gasser, Erik M. Ammann, and Philippe F.-X Corvini.(2013). Determination of Oxidoreductase Activity Using a High- Throughput Microplate Respiratory Measurement. Analytical Chemistry. 85 (1), 283-291.

Sobral-Leite, M., de Siqueira Filho, J. A., Erbar, C., & Machado, I. C. (2011 Anthecology and reproductive system of Mourera fluviatilis (Podostemaceae): Pollination by bees and xenogamy in a predominantly anemophilous and autogamous family?. Aquatic Botany, 95(2), 77-87.

Back to Table of Contents

TEST 77 BONFERRONI OUTLIER TEST

Question the test addresses

Do my sample data contain an outlier?

When to use the test?

This test is frequently used to investigate whether the studentized residuals from a linear or multiple regression model contains an outlier. It uses the standard normal distribution and is based on the largest absolute studentized residual. The null hypothesis is that the largest absolute residual is not an outlier versus the alternative hypothesis that it is an outlier.

Practical Applications

Tree growth and mortality: Wunder et al (2008) study the relationship between growth and mortality among tree species in unmanaged forests of Europe. A total of 10,329 trees of nine tree species (Picea abies, Taxus baccata, Fagus sylvatica, Tilia cordata, Carpinus betulus, Fraxinus excelsior, Quercus robur, Betula spp. and Alnus glutinosa) were analyzed. For each species a logistic regression model was built. The explanatory variables for each model were growth (as measured by relative basal area increment), tree size and site/location. The species-specific model selected was that model which had the highest goodness-of-fit. The researchers checked each species-specific model for outliers using the Bonferroni outlier test. None of the most extreme residuals could be classified as an outlier using the Bonferroni outlier test (p-value >0.05 in all cases).

Vegetation monitoring: Munson et al (2012) used long-term vegetation monitoring results from 39 large plots across four protected sites in the Sonoran Desert region to determine how plant species have responded to past climate variability. To determine if plant species canopy cover was related to the suite of climate variables and time, an analytical method of multiple regression known as hierarchical partitioning was used. Outliers were identified using the Bonferroni Outlier Test. The researchers find a number of significant outliers using this test statistic (p-value < 0.05).

Wild fire prediction: Miranda et al (2012) used linear regression to quantify the influence of drought and temporal trends in the annual number and mean size of wildfires in northern Wisconsin, USA over the period 1985 tc 1997. The regression models included an intercept, linear Annual Palmer

Drought Severity Index (PDSI) variable, PDSI with linear year, and PDSI wit quadratic year. Outliers were evaluated using the Bonferroni outlier test. The researchers report years 1986, 1991, 2004 removed as outliers for Oconto County Mean fire size regression (p-value <0.1).

How to calculate in R

The function outlierTest{car} can be used to perform this test takes the form outlierTest (model). The parameter model refers to the linear regression model. As an alternative outlier{outliers}can be used. It takes the form outlier (residual). The parameter residual is the residual from the regression model.

Example:

Enter the following data, collected, on two variables:

dependent.variable=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,

independent.variable=c(75,78,80,82,84,88,93,97,65,104,109,115,120,127)

To carry out the test enter

> outlierTest (lm(dependent.variable ~ independent.variable))

  rstudent unadjusted p-value Bonferonni p

9 46.09606 6.1286e-14 8.58e-13

The test reports that the p-value on the 9th observation is significant. This observation is an outlier and should be removed from the analysis. As an alternative we can use outlier

> regression.model<- lm(dependent.variable ~ independent.variable)

> residual<-rstandard(regression.model)

> outlier(residual)

[1] 3.45517

The function identifies 3.45517 as an outlier. We can also use the studentized residuals, in which case we would enter

> residual<-rstudent(regression.model)

> outlier(residual)

[1] 46.09606

The function identifies 46.09606 as an outlier.

References

Miranda, B. R., Sturtevant, B. R., Stewart, S. I., & Hammer, R. B. (2012 Spatial and temporal drivers of wildfire occurrence in the context of rural development in northern Wisconsin, USA. International Journal o Wildland Fire, 21(2), 141-154.

Munson, S. M., Webb, R. H., Belnap, J., Andrew Hubbard, J., Swann, D. E., Rutman, S. (2012). Forecasting climate change impacts to plant community composition in the Sonoran Desert region. Global Change Biology.

Wunder, J., Brzeziecki, B., Żybura, H., Reineking, B., Bigler, C., & Bugmann, H. (2008). Growth–mortality relationships as indicators of life-history strategies: a comparison of nine tree species in unmanaged European forests. Oikos, 117(6), 815-828.

Back to Table of Contents

TEST 78 GRUBBS TEST

Question the test addresses

Do my sample data contain an outlier?

When to use the test?

To detect outliers from normal distributed populations. The tested data are the

minimum and maximum sample values. The test is based on the largest absolute deviation from the mean of the sample. If an outlier has been identified and removed, the test should not be repeated without adjusting the critical value. This is because multiple iterations change the probabilities of detection. The test should not be used for sample sizes of six or less.

Practical Applications

Infrared spectroscopy: Seaman and Allen (2010) report on a sample of infrared spectroscopy from mixtures (run in triplicate) of three organic compounds in solution. Outliers needed to be removed before using the results in later chemometric analysis. Grubbs test was used to identify outliers. The researchers report the overall average for one triplicate group was 2.653, with a standard deviation of 2.888 Grubbs (p-value <0.05). After removing the outlier the overall standard deviation was recalculated and dropped substantially, confirming the outlier behavior of the eliminated spectrum.

Health problems in US children: Bethell et al (2011) evaluate national and state prevalence of health problems and special health care needs in children in the United States. Data was collected for 28 health variables (20 chronic conditions, 2 health risks, 6 health summary variables) and quality of care variables, for all children and separately for children with public or private sector health insurance. Finally, a test for the presence of statistical outliers, using Grubbs test, in state distributions of prevalence of health problems and quality of care scores was conducted to assess the degree to which national rates and ranges across states might be impacted by extreme values. Statistical analysis showed no significant outliers in the distribution across states (p-value >0.05 for all samples).

Human Resources: Human resource data from an Indian company was investigated by Sarkar et al (2011). The data consisted of 544 candidate grades from ten functional areas of the company – Purchasing, Finance,

Human Resources, Information Technology, Legal, Vendor management, Pipeline, Engineering, Manufacturing and Retail Sales. The Grubbs test was used to assess outliers in candidate scores from each area. For the functions of Purchasing and Human Resources outliers were identified (p-value < 0.005).

How to calculate in R

The function grubbs.test{outliers}can be used to perform this test takes the form grubbs.test(sample, type = 10, opposite = FALSE, two.sided = TRUE) The parameter sample refers to sample data, type can take on one of three values 10 is a test for one outlier (side is detected automatically and can be reversed by opposite parameter), 11 is a test for two outliers on opposite tails, 20 is test for two outliers in one tail.

Example:

Enter the following data:

sample<-c(0.189,0.167,0.187,0.183,0.186,0.182,0.181,0.184,0.177)

To carry out the test enter

> grubbs.test(sample, type = 10, opposite = FALSE, two.sided = TRUE)

Grubbs test for one outlier

data: sample

G = 2.2485, U = 0.2890, p-value = 0.03868

alternative hypothesis: lowest value 0.167 is an outlier

The test identifies 0.167 as an outlier.

References

Bethell, C. D., Kogan, M. D., Strickland, B. B., Schor, E. L., Robertson, J., & Newacheck, P. W. (2011). A national and state profile of leading health problems and health care quality for US children: key insurance disparities and across-state variations. Academic Pediatrics, 11(3), S22-S33.

Sarkar, A., Mukhopadhyay, A. R., & Ghosh, S. K. (2011). 2011 Issue 2 Performance: Research and Practice in Human Resource Management.

Seaman, J., & Allen, I. (2010). Outlier options. Quality Progress, February 2010.

Back to Table of Contents

TEST 79 GOLDFELD-QUANDT TEST FOR HETEROSCEDASTICITY

Question the test addresses

Are the residuals in a linear regression heteroscedastic?

When to use the test?

To investigate whether the residuals from a linear or multiple regression model are heteroscedastic. It tests whether the estimated variance of the regression residuals are dependent on the values of the independent variables. The null hypothesis is that of homoscedasticity or constant variance.

Practical Applications

South London house prices: The sold price data for 1,251 houses, over a nine year period from April 2000 was studied for Welling, South London by May et al (2011). Their objective was to investigate determinants of residential property values in South London. The researchers collected data on a number of independent variables - house characteristics, health and psychological factors, aesthetic factors, distance to transportation services. A hedonic multiple regression model was adopted to determine the effects of these variables on residential property values. The Goldfeld-Quandt test was used to test heteroscedasticity (p-value > 0.05).

Carbon Dioxide emissions from burning fossil fuels: Karpestam and Andersson (2011) analyzed data from on Carbon Dioxide emissions from burning of fossil fuels for the years 1871 to 2006 for the European Union and the United States. Growth rate in emissions are decomposed into trend and cyclical components using a band pass filter algorithm. The variability of the data is investigated using the Goldfeld-Quandt test. The researchers observe a decline in volatility between the two periods; 1871 to 1959 and 1960 to 2006. The test statistic rejects the hypothesis that the volatility for the United States as well as the European Union is the same for both periods (p-value >0.05). They also find the Goldfeld-Quandt test does not support dividing the modern period of 1960 to 2006 into even shorter sub-periods (p-value >0.05). This result holds for both the European Union and United States.

Wheat production: Carew et al (2009) study the Just-Pope production function and regional-level wheat data from Manitoba, Canada. They examine the relationship between fertilizer inputs, soil quality, biodiversity

indicators, cultivars and climatic conditions on the mean and variance of spring wheat yields. Using data from 2000 to 2006 a mean production function regression model is estimated. The researchers reject the hypothesis of homoskedasticity (Goldfeld-Quandt p-value <0.05).

How to calculate in R

The function gqtest{lmtest} can be used to perform this test. takes the form gqtest (model). The parameter model refers to the linear regression model.

Example: 1

Enter the following data:

dependent.variable=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,

independent.variable=c(75,78,80,82,84,88,93,97,99,104,109,115,120,127)

To carry out the test enter

> gqtest (lm(dependent.variable ~ independent.variable))

Goldfeld-Quandt test

data: lm(dependent.variable ~ independent.variable)

GQ = 1.3841, df1 = 5, df2 = 5, p-value = 0.365

Since the p-value is greater than 0.05, do not reject the null hypothesis.

References

Carew, R., Smith, E. G., & Grant, C. (2009). Factors influencing wheat yield and variability: Evidence from Manitoba, Canada. Journal of Agricultural and Applied Economics, 41(3), 625-639.

Karpestam, P., & Andersson, F. N. (2011). A flexible CO2 targeting regime. Economics Bulletin, 31(1), 297-308.

May, D. E., Corbin, A. R., & Hollins, P. D. (2011). Identifying Determinants of Residential Property Values in South London. Review of Economic Perspectives, 11(1), 3-11.

Back to Table of Contents

TEST 80 BREUSCH-PAGAN TEST FOR HETEROSCEDASTICITY

Question the test addresses

Are the residuals in a linear regression heteroscedastic?

When to use the test?

To investigate whether the residuals from a linear or multiple regression model are heteroscedastic. It tests whether the estimated variance of the regression residuals are dependent 0on the values of the independent variables. The null hypothesis is that of homoscedasticity or constant variance.

Practical Applications

Social influence on guessing: Mavrodiev et al (2013) created an experiment where subjects had to repeatedly guess the correct answer to factual questions, while having only aggregated information about the answers of others. Participants were asked six quantitative questions to which they did not know the answers, and thus could only provide a guess. Each question was repeated for five consecutive rounds. At the end of each round, the subjects were presented with either some or no information about others' guesses, after which they could revise their own estimate. A linear regression model relating the change in guess for each question to past guesses and the groups average guess for that question is tested for heteroscedasticity using the Breusch-Pagan test. For questions 2, 4 and 5 the null hypothesis of homoskedasticity was rejected (p-value <0.05).

25(OH)D concentrations and obesity: De Pergola et al (2013) investigate the relationship between serum 25(OH)D concentrations with measures of obesity such as body mass index (BMI), waist circumference, and subcutaneous and visceral fat. A cohort of 66 healthy overweight and obese patients, 53 women and 13 men were examined. Waist circumference and fasting 25(OH)D, insulin, glucose, lipid (cholesterol, HDL cholesterol, and triglyceride), C-reactive protein (CRP), and complement 3 (C3), and 4 (C4) serum concentrations were measured. Insulin resistance was assessed by the homeostasis model assessment (HOMAIR). A regression model was constructed with 25(OH)D as the dependent variable and BMI (or waist circumferences), fasting insulin (or HOMAIR) triglycerides, and CRP (or C3 or C4) as independent variables. Heteroscedasticity of the regression residuals was assessed using the Breusch-Pagan test (p-value >0.05). The null hypothesis of

homoskedasticity could not be rejected.

Community Pressure and Environmental Compliance: Edirisinghe (2013) use data from rubber processing factories in Sri Lanka to identify the impact of informal regulation on environmental compliance. Three regression models are built using three pollution measures- Chemical Oxygen Demand (COD) Biological Oxygen Demand (BOD) and Total Suspended Solids (TSS) as th dependent variables. The independent variables were Visits, TP, Type and complain. Where, Visits is the number of visits made by officials during the year, TP is the total production of rubber in the factory during the year, Type is the type of natural rubber produced and Complain a variable representing community pressure for abatement. The Breusch-Pagan test rejected the null hypothesis ($p<0.05$) for all three of the models.

How to calculate in R

The function ncvTest{car} can be used to perform this test. takes the form ncvTest (model). The parameter model refers to the linear regression model. Alternatively, bptest{lmtest}will perform the test. It takes the form bptest (model, studentize = FALSE). Note set studentize = TRUE if you wan to use the studentized residuals.

Example:

Enter the following data, collected, on two variables:

dependent.variable=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,

independent.variable=c(75,78,80,82,84,88,93,97,99,104,109,115,120,127)

To carry out the test enter

> ncvTest (lm(dependent.variable ~ independent.variable))

Non-constant Variance Score Test

Variance formula: ~ fitted.values

Chisquare = 0.009994307 Df = 1 p = 0.9203669

Since the p-value is greater than 0.05, do not reject the null hypothesis.

As an alternative we use bptest:

> bptest (lm(dependent.variable ~ independent.variable),studentize = FALSE)

Breusch-Pagan test

data: lm(dependent.variable ~ independent.variable)

BP = 0.01, df = 1, p-value = 0.9204

We obtain the same p-value and therefore do not reject the null hypothesis.

Example: using studentized residuals

On occasion you may want to use the studentized residuals to perform the test. In this case bptest is your best option. It will transform the variables so they have a mean of zero and variance of one and perform the Breusch-Pagan test on the residuals. We can do this we the above example as follows:

> bptest (lm(dependent.variable ~ independent.variable),studentize = TRUE)

studentized Breusch-Pagan test

data: lm(dependent.variable ~ independent.variable)

BP = 0.0199, df = 1, p-value = 0.888

Since the p-value is greater than 0.05, do not reject the null hypothesis.

References

De Pergola, G., Nitti, A., Bartolomeo, N., Gesuita, A., Giagulli, V. A. Triggiani, V., ... & Silvestris, F. (2013). Possible Role of Hyperinsulinemia and Insulin Resistance in Lower Vitamin D Levels in Overweight and Obese Patients. BioMed Research International, 2013.

Edirisinghe, J. C. (2013). Community Pressure and Environmenta Compliance: Case of Rubber Processing in Sri Lanka. Journal o Environmental Professionals Sri Lanka, 1(1), 14-23.

Mavrodiev, P., Tessone, C. J., & Schweitzer, F. (2013). Quantifying the effects of social influence. arXiv preprint arXiv:1302.2472.

Back to Table of Contents

TEST 81 HARRISON-MCCABE TEST FOR HETEROSKEDASTICITY

Question the test addresses

Are the residuals in a linear regression heteroscedastic?

When to use the test?

To investigate whether the residuals from a linear or multiple regression model are heteroscedastic. It tests whether the estimated variance of the regression residuals are dependent on the values of the independent variables. The null hypothesis is that of homoscedasticity or constant variance.

Practical Applications

Rain runoff in Piemonte: Viglione, Claps, and Laio (2007) investigate mean annual runoff in 47 basins in Piemonte and Valle d'Aosta region North-Western Italy using regression models with morphometric and climatic variables as independent variables. A number of regression models are constructed and tested. Two regression models are eventually selected, the first regresses mean annual runoff for a given gauging station as a linear function of mean elevation of the drainage basin above sea level, basin orientation and the Budyko radiational aridity index. The Harrison-McCabe test is used to assess heteroscedasticity (p-value <0.05). The second model regresses mean annual runoff for a for a given gauging station as a linear function of mean elevation of the drainage basin above sea level and annual rainfall areally averaged over the catchment. The Harrison-McCabe test is used to assess heteroscedasticity (p-value <0.05). The researchers conclude both regression models are homoscedastic.

Orangutan genome & humans: Hobolth et al (2011) search the complete orangutan genome for regions where humans are more closely related to orangutans than to chimpanzees due to incomplete lineage sorting (ILS) in the ancestor of human and chimpanzees. To assess the effect of gene density on ILS while controlling for recombination rate, a linear regression model was fitted. The researchers used a stepwise model selection process which retained all interactions between recombination rate, equilibrium GC content, and density of coding site up to the third order. Homoskedasticity was assessed using the Harrison-McCabe test (p-value = 0.245). The null hypothesis of homoskedasticity could not be rejected.

Southern California car purchases: A theoretical model to examine how the

transacted price of a motor car can be affected by the information contained in a buyer's decision to trade in and the traits of the trade-in were studied by Kwon et al (2012). Using a data set of 124,499 new car transactions in Southern California over the period 2002-2008 they develop a linear regression model. The basic model regresses the log consumer price paid for a car as a function of the trade-in incidence and brand loyalty variables. The Harrison-McCabe test was used to assess homoskedasticy (p-value =1).

How to calculate in R

The function hmctest(lmtest} can be used to perform this test. takes the form hmctest (model). The parameter model refers to the linear regression model.

Example:

Enter the following data, collected, on two variables:

dependent.variable=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,

independent.variable=c(75,78,80,82,84,88,93,97,99,104,109,115,120,127)

To carry out the test enter

> hmctest(lm(dependent.variable ~ independent.variable))

Harrison-McCabe test

data: lm(dependent.variable ~ independent.variable)

HMC = 0.439, p-value = 0.365

Since the p-value is greater than 0.05, do not reject the null hypothesis.

References

Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H., & Mailund, T. (2011) Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome research, 21(3), 349-356.

Kwon, O., Dukes, A. J., Siddarth, S., & Silva-Risso, J. M. (2012). Th Informational Role of Product Trade-Ins for Pricing Durable Goods.

Viglione, A., Claps, P., & Laio, F. (2007). Mean annual runoff estimation in North-Western Italy. Water Resources Assessment Under Water Scarcity Scenarios, La Loggia G., G. Aronica and G. Ciraolo (Eds.). CSDU Italy, ISB

978-88.

Back to Table of Contents

TEST 82 HARVEY-COLLIER TEST FOR LINEARITY

Question the test addresses

Is the regression model correctly specified as linear?

When to use the test?

To identify functional misspecification in a regression model. The null hypothesis is the regression model is linear. The test attempts to detect nonlinearities when the data is ordered with respect to a specific variable. If the model is correctly specified the recursive residuals (standardized one step prediction errors) have zero mean. In this sense it is essentially a t test of the recursive residuals.

Practical Applications

Software defects and lines of code: Koru et al (2008) investigated the functional form of the size-defect relationship for large software modules of open-source products. One system known as ACE consists of 174 different C++ classes each corresponding to 11,195 lines of code, and 192 total number of defects. Another system was an IBM relational database management system (IBM-DB) which was developed using C++ at the IBM Software Solutions Toronto Laboratory. It consists of a total of LOC is 185,755 lines of code and the total number of defects is 7,824. The Harvey-Collier test was used to assess the degree of linearity between the logarithms of size and defects. No evidence for nonlinearity was observed (p-value =0.48 for ACE and p-value 0.22 for IBM-DB).

Codon bias as a function of imposed GC bias: Palidwor, Perkins and Xia (2010) generate a continuous-time Markov chain model of codon bias as a function of imposed GC bias for all amino acids. We assess the model by comparing it with codon bias for prokaryote and plant genomes and the genes of the human genome. The Harvey-Collier test is used to assess the null hypothesis of linear usage for all codons as a function of GC3 for prokaryotes. The results indicated a large number of codons exhibit some degree of nonlinear usage in prokaryotes as a function of GC bias. The deviations from linearity were strongest in codons belonging to leucine, isoleucine and arginine (Harvey-Collier test p-value <0.01).

RNA interference: High-content, high-throughput RNA interference (RNAi) is used to functionally characterize genes in living cells. Knapp et al (2011) develop a method that normalizes and statistically scores microscopy based RNAi screens. The approach is tested on two infection screens for hepatitis C (HCV) and dengue virus (DENV). To test whether the effects of

the individual features on the virus signal intensities are linear, the Harvey-Collier test for linearity was computed on the log signal intensities and the raw features. The researchers report all features are significantly nonlinear for all features of the DENV and HCV screen (p-values ≤ 0.0001) except fo the spot border feature of HCV and the Column feature of DENV.

How to calculate in R

The function harvtest{lmtest} can be used to perform this test. It takes the form harvtest (model). The parameter model refers to the linear regression model.

Example:

Enter the following data, collected, on two variables:

dependent.variable=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,

independent.variable=c(75,78,80,82,84,88,93,97,99,104,109,115,120,127)

To carry out the test enter

> harvtest(dependent.variable~independent.variable)

Harvey-Collier test

data: dependent.variable ~ independent.variable

HC = 1.0485, df = 11, p-value = 0.3169

Since the p-value is greater than 0.05, do not reject the null hypothesis of linearity.

References

Knapp, B., Rebhan, I., Kumar, A., Matula, P., Kiani, N. A., Binder, M., ... & Kaderali, L. (2011). Normalizing for individual cell population context in the analysis of high-content cellular screens. BMC bioinformatics, 12(1), 485.

Koru, A. G., Emam, K. E., Zhang, D., Liu, H., & Mathew, D. (2008). Theory o relative defect proneness. Empirical Software Engineering, 13(5), 473-498.

Palidwor, G. A., Perkins, T. J., & Xia, X. (2010). A general model of codon bias due to GC mutational bias. PLoS One, 5(10), e13431.

Back to Table of Contents

TEST 83 RAMSEY RESET TEST

Question the test addresses

Is the regression model correctly specified as linear?

When to use the test?

To identify functional misspecification in a regression model. The test maintains a null hypothesis of a linear specification against the alternative hypothesis of a non-linear specification. The intuition behind the test is that if non-linear combinations of the independent variables have ability to explain the dependent variable, the model is misspecified. More specifically, it tests whether non-linear combinations of the fitted values help explain the dependent variable. If we are unable to reject the null hypothesis, then the results suggest that the true specification is linear and the regression equation passes the Ramsey Reset test.

Practical Applications

High temporal resolution tissue Doppler: Otton et al (2013) build a regression equation for the timing of the period of minimal coronary motion within the RR interval. High temporal resolution tissue Doppler was used to measure coronary motion within diastole. Tissue-Doppler waveforms of the myocardium corresponding to the location of the circumflex artery (100 patients) and mid-right coronary arteries (50 patients) and the duration and timing of coronary motion were measured. The relationship between the RR interval and the time to the E' wave, time to the A' wave, time to the isovolumic relaxation time and time to the center of the period of minimal cardiac motion for the circumflex and right coronary arteries was assessed using the Ramsay RESET test. When hear rates < 50 where excluded, the null hypothesis could not be rejected (p-value >0.3).

East Java shallot production: Saghaian (2013) develop cost functions of small scale shallot production from data collected in a in a Village in East Java. From April to July 2005 a survey was carried out by the researchers. A total of 43 village farmers completed the survey, 7 of which the researchers rejected as outlier observations. The regression cost functions were specified as a linear model, quadratic model, and a cubic model. Ramsey's RESET test was used to assess model fit. For the linear model (power = 2 the null hypothesis was rejected (p-value <0.05). For the quadratic and cubic models (power = 2) the null hypothesis could not be rejected (p-value = 0.69 and 0.23 respectively).

Awareness of management concepts of Bangladeshi managers: Zaman et al (2013) investigate the awareness level of Bangladeshi managers about 96 fashionable management concepts. A total of 130 managers were asked to complete a comprehensive questionnaire. Using awareness of fashionable concepts as the dependent variable, four linear regression equations were specified. The first used gender as the independent variable; the second used gender and level of management as independent variables; the third equation used gender, level of management and functional department; and the fourth used gender, level of management, functional department and industry as the independent variables (Ramsey's RESET test p-value >0.05 for all equations).

How to calculate in R

The function resettest{lmtest} can be used to perform this test. takes the form harvtest (model). The parameter model refers to the linear regression model.

Example:

Enter the following data, collected, on three variables:

dep=c(3083,3140,3218,3239,3295,3374,3475,3569,3597,3725,3794,3959,4043

ind.1=c(75,78,80,82,84,88,93,97,99,104,109,115,120,127)

ind.2=c(5,8,0,2,4,8,3,7,9,10,10,15,12,12)To carry out the test enter

We begin by building our basic linear regression model.

model <- lm(dep~ind.1+ind.2)

Now, we will use the RESET test to assess whether we should include second or third powers of the independent variables - ind.1 and ind.2. We can do this by typing:

> resettest(model, power=2:3, type="regressor")

RESET test

data: model

RESET = 1.6564, df1 = 4, df2 = 7, p-value = 0.2626

Since the p-value is greater than 0.05, do not reject the null hypothesis of linearity.

References

Otton, J. M., Phan, J., Feneley, M., Yu, C. Y., Sammel, N., & McCrohon, J. (2013). Defining the mid-diastolic imaging period for cardiac CT–lessons from tissue Doppler echocardiography. BMC medical imaging, 13(1), 5.

Saghaian, S. H. (2013). Profit Gap Analysis on the Small Scale Production of Shallot: A Case Study in a Small Village in East Java Province of Indonesia. In 2013 Annual Meeting, February 2-5, 2013, Orlando, Florida (No. 142550) Southern Agricultural Economics Association.

Zaman, L., Yasmeen, F., & Al Mamun, M. (2013). An Assessment of Fashionable Management Concepts' Awareness Level amongst Bangladesh Managers in their Move toward Knowledge Economy. International Journal of Applied Research in Business Administration and Economics, 2(1).

Back to Table of Contents

TEST 84 WHITE NEURAL NETWORK TEST

Question the test addresses

Is the sample of timeseries observation linear in the mean?

When to use the test?

The test can be used to investigate the null hypothesis of linearity in the mean. It uses a single hidden layer feed-forward neural network with additional direct connections from inputs to outputs.

Practical Applications

Stalagmite lamina chronologies: Continuous annual lamina chronologies for four stalagmites growing in Oman, China, Scotland and Norway, over the last 1000 years are analyzed by Baker et al (2008). The White neural network test is applied to each of the stalagmites. The null hypothesis is rejected in all cases (p-value <0.05). The researchers conclude all four are statistically nonlinear,

British Pound dynamics: Brooks (1996) investigate the dynamics of the mid-price spot of ten currencies, namely the Austrian schilling/pound, the Canadian dollar/pound, the Danish krone/pound, the French franc/pound, the German mark/ pound , the Hong Kong dollar/pound, the Italian lira/pound , the Japanese yen/pound, the Swiss franc/pound, and the US dollar/pound. The raw daily exchange rates were transformed into log-returns. The sample covers the period from 2 January 1974 until 1 July 1994. The White neural network test is applied to each of currencies and various lags. The null hypothesis is rejected in all cases with the exception of the Canadian dollar Hong Kong dollar and Japanese yen, (p-value >0.05). The researcher concludes the Canadian dollar, and to a lesser extent, the Hong Kong dollar and Japanese yen, show no evidence of non-linearity.

Metal Futures prices: Kyrtsou et al (2004) analyze the nature of the underlying process of metal futures price returns series. The daily first differences of the log of the futures prices of five metals (aluminium, nickel, tin, zinc, and lead) over the period January 1989 to April 1989 were analyzed using the White neural network test. The researchers report test statistics for aluminium, nickel, tin, zinc, and lead as 16.88, 35.99, 334.31, 8.92 and 8.07 respectively. The critical value of the test statistic at the 5% level of significance is 5.99. Since the test statistics values for each metal are greater than 5.99, the null hypothesis of linearity in mean is rejected.

How to calculate in R

The function resettest{lmtest}can be used to perform this test. takes the form harvtest (model). The parameter model refers to the linear regression model.

Example: European Stock Prices

Let's apply the test to the daily difference of the European stock market indices contained in the dataframe EuStockMarkets:

```
> set.seed(1234)

> white.test(diff(EuStockMarkets[,1],1)) # test DAX

        White Neural Network Test

data:  diff(EuStockMarkets[, 1], 1)

X-squared = 3.3931, df = 2, p-value = 0.1833

> white.test(diff(EuStockMarkets[,1],1)) # test SMI

        White Neural Network Test

data:  diff(EuStockMarkets[, 1], 1)

X-squared = 1.8393, df = 2, p-value = 0.3987

> white.test(diff(EuStockMarkets[,1],1)) # test CAC

        White Neural Network Test

data:  diff(EuStockMarkets[, 1], 1)

X-squared = 4.6166, df = 2, p-value = 0.09943

> white.test(diff(EuStockMarkets[,1],1)) # test FTSE

        White Neural Network Test

data:  diff(EuStockMarkets[, 1], 1)

X-squared = 1.1899, df = 2, p-value = 0.5516
```

We cannot reject the null for any of the stock market time series at the 5% level. However, at the 10% level the null hypothesis is rejected for the CAC index (p-value =0.09943).

References

Baker, A., Smith, C., Jex, C., Fairchild, I. J., Genty, D., & Fuller, L. (2008) Annually laminated speleotherms: a review. International Journal of

Speleology, 37(3), 193-206.

Brooks, C. (1996). Testing for non-linearity in daily sterling exchange rates. Applied Financial Economics, 6(4), 307-317.

Kyrtsou, C., Labys, W. C., & Terraza, M. (2004). Noisy chaotic dynamics in commodity markets. Empirical Economics, 29(3), 489-502.

Back to Table of Contents

TEST 85 AUGMENTED DICKEY-FULLER TEST

Question the test addresses

Does the data contain a unit root?

When to use the test?

To investigate whether a time ordered set of observations contains a unit root and is therefore non-stationary.

Practical Applications

Forecasting socioeconomic time series: Frias-Martinez et al (2013) investigate a range of forecasting models using socioeconomic time series data from the local National Statistical Institute of an emerging economy in Latin America. Models are built for six socioeconomic indicator time series that are computed monthly (total assets, measuring both tangible and financial assets of the state, total number of employed citizens, total number of workers employed by private industries and organizations, total number of civil servant employed by public institutions, total number of subcontracted workers and total number of subcontracted civil servants. For each of these series, prior to the time-series models being constructed, stationarity tests are conducted. The Augmented Dickey-Fuller test was used to assess whether the series had a unit root. For example, the null hypothesis could not be rejected for total subcontracted civil servants (Augmented Dickey-Fuller test p-value > 0.05). The researchers apply a log difference to this series and the null hypothesis is rejected on the resultant time-series (Augmented Dickey-Fuller test p-value <0.05).

Gold and Karachi stock prices: Bilal et al (2013) examine the long-run relationship between gold prices and Karachi Stock Exchange (KSE) and Bombay Stock Exchange (BSE). Monthly data on the price of the three variables is collected over the period 2005 to 2011. The Augmented Dickey-Fuller test revealed that the price series contained a unit root (p-value >0.05 for all three variables). The first difference of each of the variables resulted in a p-value <0.05.

Predicting mango cultivation: Mehmood and Ahmad (2013) develop an autoregressive integrated moving average time-series model to forecast the numbers of acres of commercial mangoes cultivation in Pakistan. Data on the size of mango cultivation from 1961 to 2009 were collected. The researchers used the Augmented Dickey-Fuller test to investigate whether this time-series had a unit root. The null hypothesis of a unit root was not rejected (p-value>0.5). To remove the unit root the researchers took the

first difference the time series data and apply the Augmented Dickey-Fuller test to this series (p-value<0.001). The researchers conclude the first difference is stationary.

How to calculate in R

The function adf.test{tseries} can be used to perform this test. It takes the form

adf.test(data,alternative ="stationary",k=21). The timeseries to be tested is contained in data. The parameter alternative refers to the form of the alternative hypothesis you wish to test against. It can be set to "stationary" or "explosive". The default is"stationary". The parameter k refers to the lag length used in the test. If unspecified the test will determine it for you.

As an alternative, the function ur.df{urca} can also be used to perform this test. It takes a slightly more complicated form, ur.df(data, type = "none" or "drift" or "trend"), lags = 21,selectlags = "Fixed" or "AIC" or "BIC"))

The parameter type refers to the form of the alternative hypothesis. You can use lags to specify the number of lags you want the test to use. Alternatively, you can have the test use the Akaike "AIC" or the Bayes "BIC" information criteria. The timeseries to be tested is contained in data. One advantage of this function is that it gives you more specificity on the alternative hypothesis. It also provides you with more detailed information on the test. However, for routine testing adf.test will generally be sufficient.

Example: Simulated data with a unit root

We can simulate a series with a unit root and test as follows:

```
>set.seed(1234)

> data <- cumsum(rnorm(10000)) # contains a unit root
```

To apply the basic form of the test enter

```
> adf.test(data)

        Augmented Dickey-Fuller Test

data:  data

Dickey-Fuller = -2.1219, Lag order = 21, p-value = 0.5267

alternative hypothesis: stationary
```

Since the p-value is greater than 0.05, do not reject the null hypothesis – the data contain a unit root. Of course, we can if we wish specify all the parameters of the test. Let's use a lag length of 10 and an alternative of explosive.

```
> adf.test(data,alternative="explosive",k=10)

        Augmented Dickey-Fuller Test

data:  data

Dickey-Fuller = -2.2628, Lag order = 10, p-value = 0.5329

alternative hypothesis: explosive
```

The test reports the lag order, and alternative hypothesis. Since the p-value is > 0.05, we cannot reject the null hypothesis at the 5% level.

Example: Cointegration of sunspots

The monthly mean relative sunspot numbers from 1749 to 1983 are contained in the object sunspots. We will use the function ur.df, to test for a unit root using the Akaike information criteria. This function supplies slight more information as shown below:

```
> summary(ur.df(sunspots, type = "none",selectlags = "AIC") )

###############################################

# Augmented Dickey-Fuller Test Unit Root Test #

###############################################

Test regression none

Call:

lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:

    Min      1Q  Median      3Q     Max

-72.554  -6.751   0.318   8.811 100.474

Coefficients:

          Estimate Std. Error t value Pr(>|t|)
```

z.lag.1 -0.023397 0.004615 -5.07 4.23e-07 ***

z.diff.lag -0.289217 0.018037 -16.04 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.33 on 2816 degrees of freedom

Multiple R-squared: 0.09876, Adjusted R-squared: 0.09812

F-statistic: 154.3 on 2 and 2816 DF, p-value: < 2.2e-16

Value of test-statistic is: -5.0703

Critical values for test statistics:

1pct 5pct 10pct

tau1 -2.58 -1.95 -1.62

The function reports the regression coefficients of the test and various other statistics. Since the p-value of the overall test is less than <2e-16, we reject the null hypothesis at the 5% and also at the 1% level.

References

Bilal, A. R., Talib, N. B. A., Haq, I. U., Khan, M. N. A. A., & Naveed, M. (2013 How Gold Prices Correspond to Stock Index: A Comparative Analysis o Karachi Stock Exchange and Bombay Stock Exchange. World Applied Sciences Journal, 21(4), 485-491.

Frias-Martinez, V., Soguero-Ruiz, C., Frias-Martinez, E., & Josephidou, M (2013, January). Forecasting socioeconomic trends with cell phone records. In Proceedings of the 3rd ACM Symposium on Computing for Developmen (p. 15). ACM.

Mehmood, S., & Ahmad, Z. (2013). Time Series Model to Forecast Area o Mangoes from Pakistan: An Application of Univariate Arima Model Academy of Contemporary Research, 1.

Back to Table of Contents

TEST 86 PHILLIPS-PERRON TEST

Question the test addresses

Does the data contain a unit root?

When to use the test?

To investigate whether a time ordered set of observations contains a unit root and is therefore non-stationary.

Practical Applications

Net discount ratio: Whether the net discount ratio is a stationary time series is investigated by Haslag et al (1994). They use the Phillips-Perron test for unit roots. Data on the discount rate was analyzed for the time period 1964 to 1993. The test results for the 1964 through 1989 period reject the null hypothesis at lags of 10, 12 and 17 months (p-value <0.05). For the entire sample period, the null hypothesis is also rejected for lags of 10,12 and 17 months (p-value <0.05). The researchers conclude the Phillips-Perron test rejects the notion that a unit root is present in the net discount ratio.

Inflation and economic growth in South Asia: Mallik and Chowdhury (2001) investigate the presence of unit roots in economic and inflation time-series for the economies of Bangladesh (1974-1997); India (1961-1997); Pakistan (1957-1997) and Sri Lanka (1966-1997). Economic growth rates were calculated from the difference of logs of real gross domestic product at 1990 prices. Inflation rates were calculated from the difference of logs of the consumer price index (1990 = 100) for all four countries. Phillips-Perron unit root test is used to investigate the presence of unit roots. The researchers reject the unit root hypothesis for economic growth for all countries (p-value < 0.05). For the inflation series the unit root hypothesis is rejected for India, Pakistan and Sri Lanka. However, for Bangladesh, the researchers find the null hypothesis could not be rejected (p-value > 0.05).

Spanish budget deficit: Using annual data, Bajo-Rubio et al (2004), tested for the order of integration of budget surplus - Gross Domestic Product ratio for the Spanish economy over the time period 1964 to 2001. The Phillips–Perron test for unit roots could not be rejected (p-value > 0.05). The authors conclude the budget surplus– Gross Domestic Product ratio for the Spanish economy is integrated of order 1.

How to calculate in R

The function PP.test{stats} or pp.test{tseries} can be used to perform this test.

Example: Simulated data with a unit root

We can simulate a series with a unit root and test as follows:

```
>set.seed(1234)

> data <- cumsum(rnorm(10000)) # contains a unit root

> PP.test(data)

        Phillips-Perron Unit Root Test

data:  data

Dickey-Fuller = -2.3117, Truncation lag parameter = 12, p-value = 0.4463
```

Since the p-value is greater than 0.05, do not reject the null hypothesis – the data contain a unit root.

Example: 2: Simulated stationary data

We conduct the test on stationary data as follows:

```
>set.seed(1234)

> data <- cumsum(rnorm(10000))

>  diff_data = diff(data,1) # unit root removed

> PP.test(diff_data)

        Phillips-Perron Unit Root Test

data:  diff_data

Dickey-Fuller = -99.7319, Truncation lag parameter = 12, p-value = 0.01

        Phillips-Perron Unit Root Test

data:  diff_data

Dickey-Fuller = -99.1831, Truncation lag parameter = 12, p-value = 0.01
```

Since the p-value is less than 0.01, reject the null hypothesis – the data does not contain a unit root.

References

Bajo-Rubio, O., Díaz-Roldán, C., & Esteve, V. (2004). Searching for threshold

effects in the evolution of budget deficits: An application to the Spanish case. Economics Letters, 82(2), 239-243.

Haslag, J. H., Nieswiadomy, M., & Slottje, D. J. (1994). Are net discount rates stationary?: some further evidence. Journal of Risk and Insurance, 513-518.

Mallik, G., & Chowdhury, A. (2001). Inflation and economic growth: evidence from four south Asian countries. Asia-Pacific Development Journal, 8(1), 123-135.

Back to Table of Contents

TEST 87 PHILLIPS-OULIARIS TEST

Question the test addresses

Is the sample of multivariate observations cointegrated?

When to use the test?

To assess the null hypothesis that a multivariate times series is not cointegrated. Intuitively, the test uses ordinary least squares to estimate the intercept and slope coefficient in a linear regression and then applies a Phillips-Perron test to determine whether the regression residual from the equation is stationary or nonstationary. It is valid when the linear regression residual series are weakly dependent and heterogeneously distributed. The test corrects for serial correlation in the regression error using the Whitney K. Newey and Kenneth D. West's (1987) estimator of the error variance.

Practical Applications

Monetary balances of households and firms: Calza and Zaghini (2010) model US monetary balances of households and firms as a function of the volume of transactions and the nominal interest rate. Two regression specifications are tested. The first a log-log model and the second is a semi-log model. A separate regression model is specified for the monetary balances of households and the monetary balances of firms. Quarterly data on each of the variables was collected from the first quarter of 1959 to the fourth quarter of 2006. The Phillips-Ouliaris test is applied to the residual of the regression models with lag truncation set to 0 and 4. For the four monetary balances of households regressions the authors reject the null hypothesis at the 15% significance level (Phillips-Ouliaris test p-value <0.15). The results are mixed for the regressions on the monetary balances of firms. The semi-log models (Phillips-Ouliaris test p-value < 0.1) reject the null hypothesis of no cointegration. However, this was not the case for the log-log specification (p-value >0.15).

Inflation and unemployment in the US: Westelius (2005) investigate the relationship between the inflation and unemployment in the US over four varying time periods. The Phillips-Ouliaris test is applied in a regression of inflation on unemployment with lag truncation parameter set to 0. The time periods used in the analysis are January 1970 to February 1997; January 1970 to January 2001; January 1970 to April 1979 and January 1980 to January 2001. The null hypothesis of no co-integration can be rejected for all time periods.

Money demand in the US: Ireland (2009) investigate the relationship between the ratio of nominal money balances (m) to nominal income and US short term nominal interest rate (r) in the post-1980 era. Using quarterly data from the Federal Reserve Bank of St. Louis FRED database the authors tests the null hypothesis of no cointegration between the natural logarithm of m and r. The Phillips-Ouliaris test is applied in a regression of the natural logarithm of m on r with lag truncation parameter ranging again between 0 and 8. The researcher reports, for all values of the lag truncation parameter, the null hypothesis is rejected (p-value <0.9).

How to calculate in R

The function po.test{tseries}can be used to perform this test. It takes the form po.test(sample,demean = TRUE), where sample is your multivariate times-series, demean indicates whether to include an intercept in the cointegration regression.

Example: European Stock Prices

Let's apply the test to the daily closing log difference of the European stock market indices contained in the dataframe EuStockMarkets:

```
> po.test(diff(log(EuStockMarkets),1),demean = TRUE)

        Phillips-Ouliaris Cointegration Test

data:  diff(log(EuStockMarkets), 1)

Phillips-Ouliaris demeaned = -1890.53, Truncation lag parameter = 18,

p-value = 0.01
```

The function proceeds by regressing the first series in EuStockMarkets (which is the DAX index) on the remaining series. Since the p-value is less than 0.05, reject the null hypothesis of no cointegration.

References

Calza, A., & Zaghini, A. (2010). Sectoral Money Demand and the Great Disinflation in the United States. Journal of Money, Credit and Banking 42(8), 1663-1678.

Ireland, P. N. (2009). On the welfare cost of inflation and the recent behavior of money demand. The American Economic Review, 1040-1052.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity. and Autocorrelation Consistent Covariance

Matrix." Econometrica, 55(3): 703–08.

Westelius, N. J. (2005). Discretionary monetary policy and inflation persistence. Journal of Monetary Economics, 52(2), 477-496.

Back to Table of Contents

TEST 88 KWIATKOWSKI-PHILLIPS-SCHMIDT-SHIN TEST

Question the test addresses

Is a sample of timeseries observations stationary around a deterministic trend?

When to use the test?

To assess the null hypothesis of stationarity against the alternative hypothesis of a unit root. This test is referred to as efficient unit root test. It can have substantially higher power than the Augmented Dickey-Fuller or Phillips-Perron unit root tests.

Practical Applications

Eucalyptus timber harvest in Galicia: Using annual data (1985-2008) González-Gómez et al (2013) study the long-run relationship between the eucalyptus timber harvest in Galicia, Spain and the three influencing factors – the price in Euros of the eucalyptus timber, pulp exports valued in Euros, the volume of salvage timber damaged by fire, measured in cubic meters. The Kwiatkowski-Phillips-Schmidt-Shin test is applied to all four of these variables. For the price in Euros of the eucalyptus timber and volume of timber damaged by fire the p-values were less than 0.05. This was not the case for timber harvest and exports (p-value > 0.5 for both variables). The researchers apply the first difference to the price in Euros of the eucalyptus timber and volume of salvage timber damaged (Kwiatkowski-Phillips-Schmidt-Shin test p-value > 0.05 for both variables).

Biodiesel, ethanol and commodities: Kristoufek (2013) analyze the relationships between biodiesel, ethanol and related fuels and agricultural commodities. The sample consists of weekly data of Brent crude oil (CO), ethanol (E), corn (C), wheat (W), sugar cane (SC), soybeans (S), sugar beets (SB), consumer biodiesel (BD), German diesel and gasoline (GD and GG) and U.S. diesel and gasoline (UD and UG) from 24.11.2003 to 28.2.2011 Except for the biofuels, the 1-month futures price was used. For biodiesel and ethanol, spot prices were used. Weekly log returns were calculated. The Kwiatkowski-Phillips-Schmidt-Shin test reported a p-value >0.1 for all series. The researchers conclude the log-return series are asymptotically stationary.

Macroeconomic variables in Nigeria: Ozughalu et al (2013) investigate the interrelationships among four macroeconomic variables in Nigeria. The four variables were the unemployment rate, real gross domestic product, real foreign direct investment and real exports. The study is based on annual

time series data from 1984 to 2010. The first differences of these variables were assessed using the Kwiatkowski-Phillips-Schmidt-Shin test. The p-values of all four differenced series were greater than 0.05, and the researchers conclude the first differences to be stationarity.

How to calculate in R

The function kpss.test{tseries} can be used to perform this test. The function takes the form kpss.test(data, null = "Level" or "Trend", lshort = TRUE). Where null refers to the null hypothesis, data is the timeseries to be tested and lshort indicates whether the short or long version of the truncation lag parameter is used.

Example:

We illustrate the use of this test statistics on data which we know to be independent and identically distributed

```
set.seed(1234)

x <- rnorm(7000)
```

To carry out the level stationary test enter:

```
> kpss.test(x, null = "Level")

        KPSS Test for Level Stationarity

data:  x

KPSS Level = 0.0527, Truncation lag parameter = 19, p-value = 0.1
```

The function reports a p-value = 0.1 and the null hypothesis cannot be rejected at the 5% level.

To apply the trend stationary test enter:

```
> kpss.test(x, null = "Trend")

        KPSS Test for Trend Stationarity

data:  x

KPSS Trend = 0.0538, Truncation lag parameter = 19, p-value = 0.1
```

The function reports a p-value = 0.1 and the null hypothesis cannot be rejected at the 5% level.

Example: European Stocks

Let's apply the test to the daily closing first difference of the DAX stock market index using data from 1991-1998. This data is contained in the dataframe EuStockMarkets:

```
> DAX<-EuStockMarkets[,1]

> diff_DAX = diff(DAX,1)

> kpss.test(diff_DAX, null = "Trend")

        KPSS Test for Trend Stationarity

data:  diff_DAX

KPSS Trend = 0.0705, Truncation lag parameter = 9, p-value = 0.1
```

The p-value for the trend test is not significant at the 5% level, Let's apply the level test:

```
> kpss.test(diff_DAX, null = "Level")

        KPSS Test for Level Stationarity

data:  diff_DAX

KPSS Level = 0.7494, Truncation lag parameter = 9, p-value = 0.01
```

The p-value in this case for is significant at the 5% level (p-value 0.01).

References

González-Gómez, M., Alvarez-Díaz, M., & Otero-Giráldez, M. S. (2013 Estimating the long-run impact of forest fires on the eucalyptus timber supply in Galicia, Spain. Journal of Forest Economics.

Kristoufek, L., Janda, K., & Zilberman, D. (2013). Regime-dependent topological properties of biofuels networks. The European Physical Journal B, 86(2), 1-12.

Ozughalu, U. M., & Ogwumike, F. O. (2013). Can Economic Growth, Foreign Direct Investment And Exports Provide The Desired Panacea To The Problem Of Unemployment In Nigeria?. Journal of Economics and Sustainable Development, 4(1), 36-51.

Back to Table of Contents

TEST 89 ELLIOTT, ROTHENBERG & STOCK TEST

Question the test addresses

Does the data contain a unit root?

When to use the test?

To investigate whether a time ordered set of observations contains a unit root and is therefore non-stationary. This test is referred to as an efficient unit root test. It is efficient in the sense that that the local asymptotic power functions are ‘‘close’’ to the asymptotic power envelopes and it can have substantially higher power than the Augmented Dickey-Fuller or Phillips-Perron unit root tests.

Practical Applications

Japanese tourist arrivals: Chang et al (2011) investigated the properties of the time series of monthly Japanese tourist arrivals to New Zealand and Taiwan over the period January 1997 to December 2007. The Elliott, Rothenberg & Stock test is used to assess the presence of a unit root. The truncation lag length is selected using a modified Akaike information criterion. The null hypothesis of a unit root is not rejected for the levels of Japanese tourist arrivals to New Zealand and Taiwan in the models with a constant (Elliott, Rothenberg & Stock test p-value > 0.05 for both Taiwan and New Zealand timeseries) and with a constant and trend (Elliott, Rothenberg & Stock test p-value > 0.05 for both Taiwan and New Zealand timeseries) as the deterministic terms. The researchers apply the Elliott, Rothenberg & Stock test to the logarithm of monthly Japanese tourist arrivals to each country. The tests do not reject the null hypothesis of a unit root for the models with a constant and with a constant and trend for Japanese tourism to New Zealand (p-value > 0.05 in all cases). However, for the series in log differences for Japanese tourists to New Zealand and Japanese tourists to Taiwan, the null hypothesis of a unit root is rejected (p-value < 0.01 in all cases). The researchers conclude the unit root tests suggest the use of log differences in monthly Japanese tourist arrivals to estimate timeseries and volatility models.

Births to unmarried women in the United States: Ermisch (2009) explores the proportion of women who are unmarried and proportion of births to unmarried women in the United States. The focus is on four age groups (20–24, 25–29, 30–34, and 35–39) and two race groups (black and white) over the period 1965–2002. For all women in any age group, the researchers cannot reject the unit root hypothesis for stationarity around

a nonzero mean, but no linear time trend, with a maximum of 2 lags (Elliott, Rothenberg & Stock test p-value >0.05 for the proportion of women who are unmarried women and the proportion of births to unmarried women for all groups and races).

United States –China real exchange rate: The real exchange rate between the United States and China is tested for a unit root by Gregory and Shelley (2011). The researchers analyze monthly data from the International Monetary Fund, over the period January 1986 through May 2010. Both Akaike information criteria and the Bayesian Information Criterion indicate an optimal augmenting lag length of 12 for the Elliott, Rothenberg & Stock test. The researchers observe the test fails to reject the presence of a unit root in the real exchange rate from between the United States and China (p-value >0.01).

How to calculate in R

The function ur.ers{urca} can be used to perform this test. It takes form, ur.ers

ur.ers(data, type = "DF-GLS" or "P-test", model = "constant"or "trend"),lag.max = 4).The parameter type refers to whether to conduct a DF-GLS test or P-test. You can use the maximum numbers of lags used for testing with lag.max. The parameter model refers to the deterministic model used for de-trending. The timeseries to be tested is contained in data. One advantage of this function is that provides you great detail on the regression model used in the test.

Example:

We illustrate the use of this test, with model= "trend" with data which we know to be independent and identically distributed:

```
set.seed(1234)

x <- rnorm(7000)
```

To carry out the level stationary test enter:

```
> summary(ur.ers(x, type="DF-GLS", model= "trend", lag.max=4))

###############################################

# Elliot, Rothenberg and Stock Unit Root Test #

###############################################
```

Test of type DF-GLS

detrending of series with intercept and trend

Call:

lm(formula = dfgls.form, data = data.dfgls)

Residuals:

Min	1Q	Median	3Q	Max
-3.4360	-0.5560	0.1505	0.8471	3.7827

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
yd.lag	-0.34688	0.01658	-20.92	<2e-16	***
yd.diff.lag1	-0.51432	0.01744	-29.48	<2e-16	***
yd.diff.lag2	-0.37361	0.01711	-21.83	<2e-16	***
yd.diff.lag3	-0.24001	0.01545	-15.54	<2e-16	***
yd.diff.lag4	-0.12426	0.01186	-10.47	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.053 on 6990 degrees of freedom

Multiple R-squared: 0.4298, Adjusted R-squared: 0.4294

F-statistic: 1054 on 5 and 6990 DF, p-value: < 2.2e-16

Value of test-statistic is: -20.9164

Critical values of DF-GLS are:

	1pct	5pct	10pct
critical values	-3.48	-2.89	-2.57

The function reports an overall p-value < 2.2e-16 [F-statistic = 1054] and the null hypothesis of a unit root is rejected at the 5% level.

References

Chang, C. L., McAleer, M., & Lim, C. (2011). Modelling the volatility in shor and long haul Japanese tourist arrivals to New Zealand and Taiwan. KIEF Discussion Paper, 783.

Ermisch, J. (2009). The rising share of nonmarital births: is it only compositional effects?. Demography, 46(1), 193-202.

Gregory, R. P., & Shelley, G. (2011). Purchasing power parity and the Chinese yuan. Economics Bulletin, 31(2), 1247-1255.

Back to Table of Contents

TEST 90 SCHMIDT - PHILLIPS TEST

Question the test addresses

Does the data contain a unit root?

When to use the test?

To investigate whether a time ordered set of observations contains a unit root and is therefore non-stationary. This is another variant of tests for the null hypothesis of a unit root when a deterministic linear trend is present. It estimates the deterministic term in a first step under the unit root hypothesis. Then the timeseries is adjusted for the deterministic terms and a unit root test is applied to the adjusted series.

Practical Applications

The yen-dollar exchange rate and the business cycle: The effects of fluctuations in the yen/dollar exchange rate on the business cycle of the smaller East Asian economies are examined by Olson (2011). The analysis used monthly data over the period 1990 to 2005 on the following variables: the yen/dollar exchange rate, the GDP of the Asean4 economies (Indonesia, Malaysia, Philippines and Thailand), and the GNP of the Newly Industria economies (Hong Kong, Korea, Singapore, and Taiwan). The Schmidt - Phillips test is used to assess unit roots in the log of each of the series (p-value > 0.05 for all series). The null hypothesis of unit roots could not be rejected. The first differences of the log of each variable are also analyzed using the Schmidt - Phillips test (tau and rho p-value < 0.01 for all series). The researcher concludes the null hypothesis is rejected and the variables are found to be stationary when the series is differenced.

Bid-ask orders of Australian stocks: Härdle et al(2012) investigate the dynamics of ask and bid orders of four stocks in a limit order book traded on the Australian Stock Exchange using a vector autoregressive model. The four companies analyzed were Broken Hill Proprietary Limited (BHP) National Australia Bank Limited (NAB), MIM and Woolworths (WOW). Dat was collected covering the period from July 8 to August 16, 2002 (30 trading days). The researchers observed more buy orders than sell orders implying that the bid side of the limit order book was changing more frequently than the ask side. BHP and NAB are significantly more actively traded thar MIM and WOW shares. The Schmidt-Phillips test was used to test for uni roots separately in the bid and ask orders for each of the four stocks. For all processes the null hypothesis of a unit root can be rejected at the 5% significance level (p-value >0.05).

Dow Jones return behavior: Chikhi (2013) investigate the memory of the Dow Jones through a range of semiparametric timeseries models with non-constant errors. The objective is to construct models which can be applied to explore the persistence of informational shocks; and to the search for long memory properties in Dow Jones returns. The sample consists of the logarithmic series of daily Dow Jones covering from May 26, 1896 to August 17, 2006, a total of 30,292 observations. This series is characterized by a unit root (Schmidt and Phillips tau and rho p-value >0.05). The first difference of the series were taken and used in the subsequent analysis.

How to calculate in R

The function ur.sp{urca} can be used to perform this test. It takes form, ur.sp(data, type = "tau" or "rho", pol.deg = 1, signif = 0.05).The parameter type refers to whether to conduct a tau or rho test, researchers frequently report both forms of the test. You can specify the degree of polynomial in the test regression ranging from one to four. The timeseries to be tested is contained in data. If you specify a value for signif, the function will return the value of the test statistic as well as the p-value.

Example:

The monthly mean relative sunspot numbers from 1749 to 1983 are contained in the object sunspots. We will use the function ur.sp, to test for a unit root of type tau, with a first degree polynomial and 5% level of significance:

> summary(ur.sp(sunspots, type="tau", pol.deg=1, signif=0.05))

##

Schmidt-Phillips Unit Root Test

##

Call:

lm(formula = sp.data)

Residuals:

Min	1Q	Median	3Q	Max
-65.795	-8.855	-1.505	8.030	102.129

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)

(Intercept) 3.1674276 0.6991430 4.53 6.13e-06 ***

y.lagged 0.9197909 0.0073972 124.34 < 2e-16 ***

trend.exp1 0.0006636 0.0003949 1.68 0.093 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.85 on 2816 degrees of freedom

Multiple R-squared: 0.8497, Adjusted R-squared: 0.8496

F-statistic: 7961 on 2 and 2816 DF, p-value: < 2.2e-16

Value of test-statistic is: -8.3987

Critical value for a significance level of 0.05 is: -3.02

The function reports an overall p-value < 2.2e-16 [F-statistic = 7961] and the null hypothesis of a unit root is rejected at the 5% level.

References

Chikhi, M., Péguin-Feissolle, A., & Terraza, M. (2013). SEMIFARMA-HYGARC Modeling of Dow Jones Return Persistence. Computational Economics, 41(2), 249-265.

Härdle, W. K., Hautsch, N., & Mihoci, A. (2012). Modelling and forecasting liquidity supply using semiparametric factor dynamics. Journal of Empirical Finance.

Olson, O. (2011). Exchange Rates Under The East Asia Dollar Standard: Th Future Of East Asian Economies. International Business & Economics Research Journal (IBER), 6(3).

Back to Table of Contents

TEST 91 ZIVOT AND ANDREWS TEST

Question the test addresses

Does the data with an expected structural break contain a unit root?

When to use the test?

To test for a unit root in a timeseries, allowing for a structural break in the series. The structural break may appear in intercept, trend or both. The Augmented Dickey Fuller, Phillips-Perron and Schmidt - Phillips type tests are not appropriate if the time series contains structural changes. In order to test for the unknown structural break, the Zivot and Andrews test uses a data dependent algorithm that regards each data point as a potential structural break and runs a regression for every possible structural break sequentially. This involves running three regressions models. The first allows for a one-time change in the intercept of the series; the second permits a one-time change in the slope of the trend function; and the third combines a one-time structural break in the intercept and trend.

Practical Applications

Labor productivity function for Argentina: Ramirez (2012) estimates a dynamic labor productivity function for Argentina that incorporates the impact of public and private investment spending, the labor force, and export growth. Data for the period 1960-2010 is collected on the following variables, the labor force (thousands occupied); the ratio of private investment to GDP; public investment spending on economic and social infrastructure as a proportion of GDP; the ratio of foreign direct investment to GDP; real government consumption expenditures as a proportion of GDP and exports of goods and services. The natural logarithm of each variable is subjected to the Zivot-Andrews with a structural break in both the intercept and the trend (p-value > 0.05 for all variables). The researcher concludes the null hypothesis with a structural break in both the intercept and the trend cannot be rejected at the 5 percent level of significance.

Relationship between immigration and real GDP in the US: Islam, Khan and Rashid (2012) study the long-run equilibrium relationship between immigration and real GDP in the United States. Annual data, from 1952 to 2000, on real Gross Domestic Product (GDP) and immigration is transformed to natural logarithms for the analysis. The authors expect both series to contain structural breaks and so use the Zivot and Andrews test to assess the presence of a unit root in each series. For the Immigration variable the researchers assumed a break in trend. For Real

GDP, a break in Intercept was assumed. Test statistics were obtained by using 1-lag for both tests. The results of the test fail to reject the null hypothesis of unit root for both series at the 5% significance level. The Zivot and Andrews test identified 1964 as a break point for the real GDP and 1992 for the immigration series. The researchers suggest the break in real GDP to have been caused by the escalation of the Vietnam War and federal Medicare. For the immigration series, they suggest the break was caused by the amnesty (Immigration Reform and Control Act of 1986) which granted legal status to a large number of undocumented immigrants.

Trade and tourism with India: Gautam and Suresh (2012) examine the relationship between tourism arrivals and bilateral trade of India with Germany, Netherland, Switzerland, France, Italy, USA, UK and Canada. Their analysis uses monthly bilateral trade data and tourist arrivals data over the period 1994January to 2008 December. The researchers test for a unit root, but expect a structural break in the data and so deploy Zivot and Andrews test. The test reject the unit root null hypothesis in thirteen out of the sixteen study variables (p-value <0.1). The three variables where the null hypothesis could not be rejected were Trade with France, Netherlands and Germany.

How to calculate in R

The function ur.za{urca} can be used to perform this test. It takes form, ur.za(data, model = "intercept" or "trend" or "both", lag=NULL).The parameter model refers to whether to conduct the test on trend, intercept or both. You can specify the highest number of lagged endogenous differenced variables to be included in the test regression with lag.

Example: real money supply

The object USeconomic{tseries} contains seasonally adjusted log of real U.S. money M1 and log of GNP in 1982 Dollars; discount rate on 91-Day treasury bills rs and yield on long-term treasury bonds rl. To apply the test where the structural break may appear in both intercept and trend to the real money M1with a lag of 3 enter:

```
>data(USeconomic)

>m1<- diff(USeconomic[,1],1)

>summary(ur.za(m1, model="both", lag=3))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.287e-03	2.618e-03	-0.492	0.623809	
y.l1	4.139e-01	1.149e-01	3.602	0.000457	***
trend	6.807e-05	5.803e-05	1.173	0.243047	
y.dl1	-1.771e-03	1.122e-01	-0.016	0.987430	
y.dl2	-1.825e-02	1.009e-01	-0.181	0.856807	
y.dl3	6.232e-02	8.858e-02	0.704	0.483041	
du	-1.227e-02	4.136e-03	-2.966	0.003627	**
dt	2.461e-04	1.091e-04	2.255	0.025888	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01013 on 123 degrees of freedom

(4 observations deleted due to missingness)

Multiple R-squared: 0.3895, Adjusted R-squared: 0.3547

F-statistic: 11.21 on 7 and 123 DF, p-value: 6.402e-11

Teststatistic: -5.1001

Critical values: 0.01= -5.57 0.05= -5.08 0.1= -4.82

Potential break point at position: 76

The test reports a potential break at position 76 (Q1 1973). The overall p-value is <0.05 and we reject the null hypothesis.

References

Gautam, V., & Suresh, K. G. (2012). An Empirical Investigation About Relationship Between International Trade And Tourist Arrival: Evidence From India. Business Excellence and Management, 2(3), 53-62.

Islam, F., Khan, S., & Rashid, S. (2012). Immigration and Economic Growth Further Evidence from US Data. Review of Applied Economics, 8(1).

Ramirez, M. D. (2012). Are Foreign and Public Investment Spending Productive in the Argentine Case? A Single Break Unit Root and Cointegration Analysis, 1960-2010. Modern Economy, 3(6), 726-737.

Back to Table of Contents

TEST 92 GRAMBSCH-THERNEAU TEST OF PROPORTIONALITY

Question the test addresses

Is the assumption of proportional hazards for a Cox regression model fit valid?

When to use the test?

If you are building a Cox proportional hazard model a key assumption is proportional hazards. This can be assessed using this test. Essentially it tests for a non-zero slope in a generalized linear regression of the scaled Schoenfeld residuals on functions of time. A non-zero slope is an indication of a violation of the proportional hazard assumption.

Practical Applications

Still births in Scotland: Smith et al (2004) study whether the risk of antepartum stillbirth varies in relation to circulating markers of placental function measured during the first trimester of pregnancy. A total of 7934 women who had singleton births at or after 24 weeks' gestation, who had blood taken during the first 10 weeks after conception, and who were entered into national registries of births and perinatal deaths in Scotland from 1998 to 2000 were analyzed in the study. The association between pregnancy-associated plasma protein level and stillbirth was assessed via various statistical methods. Hazard ratios were estimated using a Cox proportional hazards model. To assess the proportional hazards assumption the Grambsch-Therneau test was used (p-value = 0.25). The researchers observe there was no evidence of non-proportionality.

Breast Cancer Recurrence: Brewster et al (2008) investigate the residual risk of breast cancer recurrence 5 years after adjuvant therapy. The researchers evaluated the residual risk of recurrence and prognostic factors of 2838 patients with stage I–III breast cancer who were treated with adjuvant or neo-adjuvant therapy (AST) between January 1, 1985, and November 1 2001, and remained disease free for 5 years. Recurrence-free survival modeled with a multivariable Cox proportional hazards models. The independent factors considered in the model included age at diagnosis (≤35, 36–59, or ≥60 years), year of start of AST (before 1992 or 1992 or later), hormone receptor status, chemotherapy (anthracycline, anthracycline and taxane, other, or none), endocrine therapy (tamoxifen, aromatase inhibitor, tamoxifen and aromatase inhibitors, other, or none), stage (I, II, or III), surgery type (breast conserving or mastectomy), radiation

(yes or no), and grade (1, 2, or 3) .The proportionality assumption was tested using the Grambsch-Therneau test (p-value = 0.2). The researchers observe that the assumption of proportionality was not violated for their fitted model.

Modeling blood pressure risk: Glynn (2002) develop models that quantify the risk associated with both systolic and diastolic blood pressure and to infer the benefits of antihypertensive therapy. A total sample of 22,071 males and 39,876 women were used to develop gender-specific predictive models via Cox regression. Independent variables included age, body mass index, current hypertension treatment, diabetes, parental history of MI before 60 years, smoking status (never, former, current), exercise (none, <2 times/week, ≥2 times/week), and alcohol intake (<1 drink/week, 1–6 drinks/week, ≥1 drink/day). The proportional hazards assumption was tested using the Grambsch-Therneau test (p-value for all models >0.05). The researchers observe that the assumption of proportionality was tenable for all models.

How to calculate in R

The function cox.zph{survival} can be used to perform this test. It takes the form cox.zph (fit), where fit is the Cox regression model fit.

Example:

Suppose you have collected the data given below:

```
sample<- list(time=c(3,3,3,4,3,1,1,2,2,3,3,4),
              status=c(0,0,1,1,1,1,0,1,1,0,0,1),
              factor.1=c(2,2,1,0,2,1,1,1,0,0,0,0),
              factor.2=c(1,0,0,0,0,0,0,1,1,1,1,1))
```

A Cox proportional hazards model can be fitted to this data by entering:

```
fit<-coxph(Surv(time, status) ~ factor.1 + factor.2, sample)
```

To apply the Grambsch-Therneau test of proportionality enter.

```
> cox.zph(fit)
            rho  chisq     p
factor.1  0.2773 0.2125 0.645
factor.2 -0.0441 0.0115 0.915
```

GLOBAL NA 0.2884 0.866

The test returns a p-value on each of the factors (p-value factor.1 = 0.645, p-value factor.2 = 0.915), and also a globally (p-value=0.866). For this example, we cannot reject the null hypothesis of proportionality.

References

Brewster, A. M., Hortobagyi, G. N., Broglio, K. R., Kau, S. W., Santa-Maria, C A., Arun, B., ... & Esteva, F. J. (2008). Residual risk of breast cancer recurrence 5 years after adjuvant therapy. Journal of the National Cancer Institute, 100(16), 1179-1183.

Glynn, R. J., Gilbert, J. L., Sesso, H. D., Jackson, E. A., & Buring, J. E. (2002 Development of predictive models for long-term cardiovascular risk associated with systolic and diastolic blood pressure. Hypertension, 39(1), 105-110.

Smith, G. C., Crossley, J. A., Aitken, D. A., Pell, J. P., Cameron, A. D., Connor J. M., & Dobbie, R. (2004). First-trimester placentation and the risk of antepartum stillbirth. JAMA: the journal of the American Medica Association, 292(18), 2249-2254.

Back to Table of Contents

TEST 93 MANTEL-HAENSZEL LOG-RANK TEST

Question the test addresses

Are there statistically significant differences between two or more survival curves?

When to use the test?

When the tail of the survival curve is of primary interest because the long-rank test emphasizes the tail of the survival curve in that it gives equal weight to each failure time.

Practical Applications

Comparing cardiovascular Interventions: Hannan et al (2013) investigate whether patients with coronary artery disease (CAD) without ST-elevation myocardial infarction (STEMI) have significantly different 3-year mortality rates with staged percutaneous coronary intervention (PCI) than when they undergo complete revascularization (CR). A total of 15,955 patients in New York between 2007 and 2009 were analyzed in the study. Patients with acute coronary syndrome (ACS) (unstable angina, or recent myocardial infarction within 7 days, without STEMI) and patients without ACS are analyzed separately. Patients without STEMI undergoing PCI were separated into 2 group (staged CR and unstaged CR: those with acute coronary syndrome but no STEMI, and those without acute coronary syndrome). Mortality of staged and unstaged patients for a 3-year follow-up period was assessed using the Mantel-Haenszel log-rank test. The researchers report the three-year mortality for propensity-matched multivessel CAD patients without ACS (Mantel-Haenszel log-rank test p value =0.68); and three-year mortality for propensity-matched multivessel CAD patients with ACS (Mantel-Haenszel log-rank test p-value =0.22).

Protocadherin-10 protein levels and bladder cancer survival rates: Ma et al (2013) assess the difference of overall survival between patients with normal and down-regulated levels of protocadherin-10 (PCDH10) protein immunoreactivity. Tumour samples from patients with bladder transitional cell carcinoma were collected during surgery at the Department of Urology, Second Hospital of Tianjin Medical University, Tianjin, China between January 2003 and June 2006. A total of 38 samples were taken from patients with normal levels of PCDH10 protein immunoreactivity, and 67 from patients with down-regulated levels of PCDH10 protein immunoreactivity. The Mantel-Haenszel log-rank test was used to assess the difference in survival between the two groups (p-value = 0.0055). The

researchers conclude down-regulated levels of PCDH10 were significantly associated with decreased overall survival rates.

Vitamin D and chronic obstructive lung disease mortality: Holmgaard et al (2013) investigate whether vitamin D deficiency or insufficiency was associated with mortality rate in patients suffering from advanced Chronic Obstructive Lung Disease (COPD) in a 10-Year Prospective Cohort Study. 25-OHD serum levels (vitamin D) were measured in 462 patients suffering from moderate to very severe COPD. Participants were stratified into 3 groups according to serum levels of 25-OHD, >30 ng/ml, 30–20 ng/ml and <20 ng/ml. The Mantel-Haenszel log-rank test was used to assess overal survival of the three groups (p-value = 0.26). Three-year survival according to levels of serum 25-OHD distributed on tertiles was also assessed using the Mantel-Haenszel log-rank test (p-value =0.26). The researchers conclude vitamin D does not appear to be associated with mortality rate, suggesting no or only a minor role of vitamin D in disease progression in patients with moderate to very severe COPD.

How to calculate in R

The function survdiff{survival} can be used to perform this test. It takes the form survdiff (formula,rho=0), where formula refers to the curves to be tested.

Example:

Suppose you have collected the data on time, two factors and the status as given below:

```
time <- c(13, 18, 28, 26, 21, 22, 24, 25, 10, 13, 15, 16, 17, 19, 25, 32)#months

status <- c(1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1)

treatment.group <- c(1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2)

sex <- c(1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 2, 1, 1, 1, 2, 2)# 1 = male
```

To apply the test enter.

```
> survdiff(Surv(time, status) ~ treatment.group, rho=0)

Call:

survdiff(formula = Surv(time, status) ~ treatment.group, rho = 0)

                    N Observed Expected (O-E)^2/E (O-E)^2/V

treatment.group=1 8         6     6.25    0.0102    0.0281
```

treatment.group=2 8 5 4.75 0.0135 0.0281

Chisq= 0 on 1 degrees of freedom, p= 0.867

The test returns a p-value of 0.867, the null hypothesis that the survival times are similar between the two groups cannot be rejected.

References

Hannan, E. L., Samadashvili, Z., Walford, G., Jacobs, A. K., Stamato, N. J Venditti, F. J., ... & King, S. B. (2013). Staged Versus One-time Complete Revascularization With Percutaneous Coronary Intervention for Multivessel Coronary Artery Disease Patients Without ST-Elevation Myocardia Infarction. Circulation: Cardiovascular Interventions, 6(1), 12-20.

Holmgaard, D. B., Mygind, L. H., Titlestad, I. L., Madsen, H., Fruekilde, P. E N., Pedersen, S. S., & Pedersen, C. (2013). Serum Vitamin D in Patients with Chronic Obstructive Lung Disease Does Not Correlate with Mortality-Results from a 10-Year Prospective Cohort Study. PloS one, 8(1), e53670.

Ma, J. G., He, Z. K., Ma, J. H., Li, W. P., & Sun, G. (2013). Downregulation o protocadherin-10 expression correlates with malignant behaviour and poor prognosis in human bladder cancer. Journal of International Medical Research, 41(1), 38-47.

Back to Table of Contents

TEST 94 PETO AND PETO TEST

Question the test addresses

Are there statistically significant differences between two or more survival curves?

When to use the test?

When the tail of the survival curve is of primary interest because the Peto test emphasizes the beginning of the survival curve in that earlier failures receive higher weights.

Practical Applications

Nuclear radio components in Seyfert galaxies: Thean et al (2002) study the properties of compact nuclear radio components in Seyfert galaxies from a 12 μm Active Galactic Nuclei sample. The sample was obtained from radio observations made with the VLA in A–configuration at 8.4 GHz. These 0.25–arcsec–resolution observations allow elongated radio structures tens of parsecs in size to be resolved and enable radio components smaller than 3.5 arcsec to be isolated from kiloparsec–scale, low–brightness–temperature emission. The researchers make a number of observations. First, there is no significant difference between the 8.4 GHz A–configuration flux densities of type 1 and type 2 Seyferts (Peto and Peto test p-value = 0.919); Second, the luminosity distributions of type 1 and type 2 Seyferts are drawn from the same parent distribution (Peto and Peto test p-value = 0.7122); third, the nuclear radio structures in type 1 and type 2 Seyferts are drawn from the same parent distribution (Peto and Peto test p-value = 0.5969).

Encephalopathic crises: Harting et al (2009) analyzed magnetic resonance images (MRIs) in 38 patients with glutaric aciduria type I diagnosed before or after the manifestation of neurological symptoms. As part of their analysis they test differences in the time course of these MRI abnormalities among patients with and without encephalopathic crises (AEC). They report deep grey matter structures- putamen versus without AEC (Peto and Peto test p-value 0. 0.005) and deep grey matter structures- caudate versus without AEC (Peto and Peto test p-value 0. 0.037). The researchers observe the test showed that striatal (putamen, caudate) MRI abnormalities differed between patients with and without encephalopathic crises.

Propagation of Agave macroacantha: The establishment and survival of bulbils and seedlings of Agave macroacantha in the Tehuacán Valley,

Mexico, between 1991 and 1994 was studied by Arizaga and Ezcurra (2000). A total of 102 bulbils were collected and divided into three categories: small (<4.0 cm height, 48 in total), intermediate (4.0–5.9 cm, 30 in total), and large bulbils (≥6 cm, 24 in total). The bulbils were planted under three nurse shrubs (Acacia coulteri) of similar size. No significant differences were found in bulbil survivorship between the three size classes (Peto and Peto test p-value >0.05). However, the researchers report non-nursed plants died faster when planted during the rainy season (Peto and Peto test p-value <0.05).

How to calculate in R

The function survdiff{survival} can be used to perform this test. It takes the form survdiff (formula,rho=0), where formula refers to the curves to be tested.

Example:

Suppose you have collected the data on time, two factors and the status as given below:

```
time <- c(13, 18, 28, 26, 21, 22, 24, 25, 10, 13, 15, 16, 17, 19, 25, 32)#months

status <- c(1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1)

treatment.group <- c(1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2)

sex <- c(1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 2, 1, 1, 1, 2, 2)# 1 = male
```

To apply the test enter.

```
> survdiff(Surv(time, status) ~ treatment.group, rho=1)

Call:

survdiff(formula = Surv(time, status) ~ treatment.group, rho = 1)

                    N Observed Expected (O-E)^2/E (O-E)^2/V

treatment.group=1 8     3.76     4.14    0.0349     0.127

treatment.group=2 8     3.06     2.68    0.0541     0.127

Chisq= 0.1  on 1 degrees of freedom, p= 0.721
```

The test returns a p-value of 0.721, the null hypothesis that the survival times are similar between the two groups cannot be rejected.

References

Arizaga, S., & Ezcurra, E. (2002). Propagation mechanisms in Agave macroacantha (Agavaceae), a tropical arid-land succulent rosette. American Journal of Botany, 89(4), 632-641.

Harting, I., Neumaier-Probst, E., Seitz, A., Maier, E. M., Assmann, B., Baric, I ... & Kölker, S. (2009). Dynamic changes of striatal and extrastriatal abnormalities in glutaric aciduria type I. Brain, 132(7), 1764-1782.

Thean, A., Pedlar, A., Kukula, M. J., Baum, S. A., & O'Dea, C. P. (2002). High resolution radio observations of Seyfert galaxies in the extended 12-µm sample–II. The properties of compact radio components. Monthly Notices of the Royal Astronomical Society, 325(2), 737-760.

Back to Table of Contents

TEST 95 KUIPER'S TEST OF UNIFORMITY

Question the test addresses

Is the sample equally distributed with respect to angle?

When to use the test?

To assess the null hypothesis that a sample is uniformly distributed on the circle. The test was originally designed for problems defined on a circle, for example, to test whether the distribution in longitude of something agrees with some theory. The test is as sensitive in the tails as at the median and invariant under cyclical data transformations.

Practical Applications

Throbbing pain and arterial pulsations: Mirza et al (2012) recorded the subjective report of the throbbing rhythm and the arterial pulse in subjects with throbbing dental pain. A total of 29 records were analyzed in the study. The phase synchronization between the heart rate and throbbing pain rate waveforms was assessed using Kuiper's test. The researchers report uniformity of the relative phase distribution using (p-value >0.1). The researchers conclude synchrony between arterial pulse and throbbing rhythm shows no relationship.

Odor and fly orientation: Bhandawat et al (2010) used experimental methods for studying tethered flight of the Drosophila melanogaster fly. A fly was rigidly oriented into a stream of air. Odors were injected into the air stream using a computer-controlled valve while the wing movements of the fly were monitored with an optical sensor. A total of 22 trials from 17 flies was used in the analysis. The researchers observe the orientation distributions were significantly different during the odor period and the pre-odor period (Kuiper's test p-value <0.05).

U.S. football games tickets: Lu and Giles (2010) study the psychological barriers in prices for pro-football tickets in the eBay auction market. Their sample consisted of 1,159 successful auctions for tickets for professional U.S. football games in the eBay “event tickets” category between 25 November and 2 December 2004. The researchers test for psychological barriers using cyclical permutations of the data. The null hypothesis is that there are no psychological barriers in prices for pro-football tickets in the eBay auction market. The researchers report that psychological barriers are absent in eBay auctions for pro-football tickets (bootstrapped Kuiper's one-sample test of uniformity >0.4).

How to calculate in R

The functions Kuiper{CircStats}and Kuiper.test{circular}can be used to perform this test. They take the form kuiper(data_radan)and Kuiper.test(data_radan). Note data_radan is a vector of angular measurements in radians.

Example: Green sea turtles

Luschi et al (2001) investigate the navigational abilities of green sea turtles. The dataframe turtles, from the circular package contains observations on the directions from which 10 green sea turtles approached their nesting island (Ascension Island, South Atlantic Ocean) after having been displaced to open-sea sites. We convert the data to radians and then apply the test using both functions:

> turtles_radan<- 0.0174532925*turtles[,2] # convert degrees to radans

> kuiper(turtles_radan)

Kuiper's Test of Uniformity

Test Statistic: 2.3281

P-value < 0.01

> kuiper.test(turtles_radan)

Kuiper's Test of Uniformity

Test Statistic: 3.3428

P-value < 0.01

Although both functions report slightly different test statistics, they both reject the null hypothesis (p-value <0.01). The data are not uniformly distributed on the circle.

References

Bhandawat, V., Maimon, G., Dickinson, M. H., & Wilson, R. I. (2010) Olfactory modulation of flight in Drosophila is sensitive, selective and rapid. The Journal of experimental biology, 213(21), 3625-3635.

Lu, O. F., & Giles, D. E. (2010). Benford's Law and psychological barriers ir certain eBay auctions. Applied Economics Letters, 17(10), 1005-1008.

Luschi, P., Åkesson, S., Broderick, A. C., Glen, F., Godley, B. J., Papi, F., & Hays, G. C. (2001). Testing the navigational abilities of ocean migrants: displacement experiments on green sea turtles (Chelonia mydas). Behavioral Ecology and Sociobiology, 50(6), 528-534.

Mirza, A. F., Mo, J., Holt, J. L., Kairalla, J. A., Heft, M. W., Ding, M., & Ahn, A H. (2012). Is There a Relationship between Throbbing Pain and Arteria Pulsations?. The Journal of Neuroscience, 32(22), 7572-7576.

Back to Table of Contents

TEST 96 RAO'S SPACING TEST OF UNIFORMITY

Question the test addresses

Is the sample equally distributed with respect to angle?

When to use the test?

To assess the null hypothesis of uniformity (or bimodal opposing directions). The statistics is based on the mean angle of the data and Rayleigh's measure of circular spread. The test was originally designed for problems defined on a circle, for example, to test whether the distribution in longitude of something agrees with some theory.

Practical Applications

Artificial fish aggregating devices and Tuna catch: Hallier and Gaertner (2008) compare the migratory patterns between drifting artificial fish aggregating devices (FAD) - and free school-caught Yellow fin and skipjack tuna. The fish were tagged and monitored during 11 cruises (4 in the Atlantic Ocean and 7 in the Indian Ocean). Rao's spacing test of uniformity was used to test whether the angular migrations differed significantly from randomness. Four circular distributions were considered by fishing mode: Yellow fin /FAD (n =167, p-value <0.01), Yellow fin/ free caught (n=13, p-value<0.01), skipjack /FAD (n=519, p-value <0.01), skipjack / free caught (n=52,p-value<0.01), where n is the number of fish tagged and monitored for each fishing mode. The authors conclude the null hypothesis of uniformity was rejected for all of the 4 circular distributions considered.

Angular analysis of tree roots: Di Iorio et al (2005) assess the influence of slope on the architecture of woody root systems. Five mature, single-stemmed Quercus pubescens trees growing on a steep slope and five on a shallow slope were excavated to a root diameter of 1 cm. The center of volume (COV) of the first- and second-order laterals and the center of branching (COB) of the first-order at increasing radial distance from the root-stump center was assessed using Rao's spacing test of uniformity. The researchers observe in steep-slope trees, the COVs for first-order roots showed a clustering tendency (Rao's spacing test, $P < 0.01$). In shallow-slope trees, the centers of root COV were randomly distributed (Rao's spacing test, $P > 0.05$).

Solar orientation of sandhoppers: Experiments on solar orientation of adult sandhoppers (Talitrus saltator) were undertaken by Ugolini et al (2002). The research involved a reduction and/or phase shift of the hours of light or dark. The sandhoppers were released into an apparatus which

prevented the sandhoppers from viewing the surrounding landscape but allowed them to see the sun and sky. Groups of approximately five individuals were released into the bowl containing approximately 1 cm of seawater. Each individual was tested only once, and a single direction per individual was recorded. Rao's test was applied to assess whether the distribution differed from uniformity (p-value ≤0.05).

How to calculate in R

The functions rao.spacing{CircStats}and rao.spacing.test{circular}can be used to perform this test. They take the form rao.spacing (data_radan, rad=TRUE)and rao.spacing.test (data_radan). Note data_radan is a vector of angular measurements in radians, if the data are measured in degrees set rad=FALSE.

Example: Green sea turtles

Luschi et al (2001) investigate the navigational abilities of green sea turtles. The dataframe turtles, from the circular package contains observations on the directions from which 10 green sea turtles approached their nesting island (Ascension Island, South Atlantic Ocean) after having been displaced to open-sea sites. Since the data are recorded in degrees we can carry out the test directly using rao.spacing:

```
> rao.spacing(turtles[,2],rad=FALSE)

    Rao's Spacing Test of Uniformity

Test Statistic = 227

P-value < 0.001
```

Since the p-value is less than 0.05, we reject the null hypothesis at the 5% level. Next we convert the data to radians and then apply the test using both functions:

```
> turtles_radan<- 0.0174532925*turtles[,2] # convert degrees to radans

> rao.spacing.test(turtles_radan)

    Rao's Spacing Test of Uniformity

Test Statistic = 227

P-value < 0.001

> rao.spacing(turtles_radan,rad=TRUE)
```

Rao's Spacing Test of Uniformity

Test Statistic = 227

P-value < 0.001

Both functions report reject the null hypothesis (p-value <0.01). The data are not uniformly distributed on the circle.

References

Di Iorio, A., Lasserre, B., Scippa, G. S., & Chiatante, D. (2005). Root systen architecture of Quercus pubescens trees growing on different sloping conditions. Annals of Botany, 95(2), 351-361.

Hallier, J. P., & Gaertner, D. (2008). Drifting fish aggregation devices could act as an ecological trap for tropical tuna species. Marine Ecology Progress Series, 353, 255-264.

Luschi, P., Åkesson, S., Broderick, A. C., Glen, F., Godley, B. J., Papi, F., & Hays, G. C. (2001). Testing the navigational abilities of ocean migrants: displacement experiments on green sea turtles (Chelonia mydas). Behavioral Ecology and Sociobiology, 50(6), 528-534.

Ugolini, A., Tiribilli, B., & Boddi, V. (2002). The sun compass of the sandhopper Talitrus saltator: the speed of the chronometric mechanism depends on the hours of light. Journal of experimental biology, 205(20), 3225-3230.

Back to Table of Contents

TEST 97 RAYLEIGH TEST OF UNIFORMITY

Question the test addresses

Is the sample equally distributed with respect to angle?

When to use the test?

To assess whether the distribution of sample angles is uniformly distributed. The test was originally designed for problems defined on a circle, for example, to test whether the distribution in longitude of something agrees with some theory.

Practical Applications

Magnetoencephalography and electroencephalography phase angles: Rana et al (2013) analyze magnetoencephalography and electroencephalography phase angles from a pre-stimulus period or from resting-state data. The phase angle difference between two regions of interest is computed and the corresponding unit vector is found. The vectors were averaged across trials and the magnitude of the average vector was assessed using the Rayleigh test of uniformity applied (p-value < 0.05).

Navigational Efficiency of Nocturnal Ants: To better understand the evolution of nocturnal life, Narendra, Reid and Raderschall (2013) investigate the navigational efficiency of the nocturnal ants (Myrmecia pyriformis) at different light levels. Ants were allowed individually to travel in a narrow corridor from the nest to their main foraging tree. The initial mean heading direction of ants before sunset (51.016 degrees) and after sunset (54.033 degrees) was close to the true nest direction (60 degrees). The researchers observe the orientation of ants before sunset was distributed uniformly around a circle (p-value 0.30); this was not the case, at the 10% level of significance, after sunset (p –value =0.07).

Reproductive peaks in swamp forests and savannas: Silva et al (2011) investigate whether the reproductive peaks in riparian forests are different from those of the savannas in Brazil. The first day of January was coded to correspond to 15 degrees, first day of February corresponded to 45 degrees; the first day of March corresponded to 75 degrees, and so on. Four combinations of vegetation type – phonological were assessed using the Rayleigh test of uniformity – Cerrado/ Flowering (mean angle 320.4, p-value = 0.032), Cerrado/ Fruiting (mean angle 3.5, p-value = 0.062), Swamp forest/ Flowering (mean angle 281.2, p-value = 0.001), Cerrado/ Fruiting (mean angle 312.7, p-value = 0.001).

How to calculate in R

The functions r.test{CircStats}and rayleigh.test{circular}can be used to perform this test. They take the form r.test(data_radan,degree=FALSE)and rayleigh.test (data_radan). Note data_radan is a vector of angular measurements in radians, if the data are measured in degrees rather than radians set degrees =TRUE.

Example: Desert ants

Wehner and Müller (1985) examine interocular transfer in the desert ant (Cataglyphis fortis). In one experiment measurements are recorded on the directions of 11 ants after one eye on each ant was 'trained' to learn the ant's home direction, then covered and the other eye uncovered. The data is stored as a list (first column) in the dataset fisherB10 from the circular package. Since the data are recorded in degrees we can carry out the test directly using r.test:

```
> ants<- as.numeric (fisherB10[[1]])

> r.test(ants,degree=TRUE)

$r.bar

[1] 0.9735658

$p.value

[1] -3.558271e-07
```

To use rayleigh.test we first convert the data into radians and then apply the test.

```
> ants_radians<-0.0174532925*circular(ants)

> rayleigh.test(ants_radians)

    Rayleigh Test of Uniformity

    General Unimodal Alternative

Test Statistic:  0.9736

P-value:  0
```

Since the p-value is less than 0.01, we can reject the null hypothesis at the 1% level.

References

Narendra, A., Reid, S. F., & Raderschall, C. A. (2013). Navigational Efficiency of Nocturnal Myrmecia Ants Suffers at Low Light Levels. PLOS ONE, 8(3 e58801.

Rana, K. D., Vaina, L. M., & Hämäläinen, M. S. (2013). A fast statistica significance test for baseline correction and comparative analysis in phase locking. Frontiers in Neuroinformatics, 7.

Silva, I. A., da Silva, D. M., de Carvalho, G. H., & Batalha, M. A. (2011 Reproductive phenology of Brazilian savannas and riparian forests: environmental and phylogenetic issues. Annals of forest science, 68(7), 1207-1215.

Wehner, R., & Müller, M. (1985). Does interocular transfer occur in visual navigation by ants?.

Back to Table of Contents

TEST 98 WATSON'S GOODNESS OF FIT TEST

Question the test addresses

Is the sample uniformly distributed or from the Von Mises distribution?

When to use the test?

To test a given distribution to determine the probability that it derives from a Von Mises or uniform distribution. The test uses a mean square deviation and is especially powerful for small sample sizes, unimodal and multimodal data.

Practical Applications

Narwhal movement in Kolutoo Bay: An estimated 12,650 narwhals (8,750 in 2007 and 3,900 in 2008) grouped in 4,568 clusters were observed travelling into Kolutoo Bay by Marcoux (2011). Watson's test for uniformity was used to evaluate the evenness of the movements around the tidal and the circadian cycle as well as a Watson's test for the von Mises distribution to evaluate the normality of the observed sample. The researchers find in both years, the movements of clusters into and out of the bay were not distributed uniformly around the tidal cycle (2007: Watson's Test for uniform distribution p-value < 0.01, 2008: Watson's test for uniform distribution p-value < 0.01). The researchers also report the sample was neither unimodal and linearly normally distributed (2007: Watson's test for the von Mises distribution p-value<0.01; 2008: Watson's test for the von Mises distribution p-value<0.01). However, the herds were distributed uniformly around the tidal cycle (Watson's test uniform distribution p-value>0.1 and followed the von Mises distribution (Watson's test for the von Mises Distribution p-value > 0.1).

Precise axon growth: Precise axon growth is required for making proper connections in development and after injury. Li and Hoffman-Kim (2008) study axon in vitro outgrowth assays using circular statistical methods to evaluate directional neurite response. The direction of neurite outgrowth from dorsal root ganglia derived neurons on different substrate types was measured. A variety of types of substrates were used and an assessment on the directionality of neurite outgrowth made. For the adsorbed uniform protein coating on glass the researchers report phase contrast images of neurons showed neurite outgrowth in all directions (Watson test for uniform distribution p-value >0.05). The null hypothesis of uniformity of neurite angle distributions could not be rejected.

Gaze behavior and eye–hand coordination: A total of 10 students (4

women and 6 men) with normal vision participated in a gaze behavior and eye–hand coordination study by Sailer et al (2005). Participants learned a visual motor task which involved hitting a target with a rigid tool held freely between two hands. Learning occurred in stages that could be distinguished by changes in performance (target–hit rate) as well as by gaze behavior and eye–hand coordination. In a first exploratory stage, the hit rate was consistently low. In a second skill acquisition and refinement stage, the hit rate improved rapidly. The directional distribution of saccades in the second half of the skill acquisition stage and in the skill refinement stage did not differ significantly from a uniform distribution of saccades in all directions (Watson's test p-value > 0.12 for both stages), whereas the direction of sub-movements did (Watson's test p-value < 0.0001 for both stages).

How to calculate in R

The functions watson{CircStats}and watson.test{circular}can be used to perform this test. They take the form watson(data_radan, dist='uniform' or dist='vm') and watson.test (data_radan, dist= 'uniform' or dist= 'vonmises'). Note data_radan is a vector of angular measurements in radians, if the data are measured in degrees rather than radians set degrees =TRUE.

Example: Desert ants

Wehner and Müller (1985) examine interocular transfer in the desert ant (Cataglyphis fortis). In one experiment measurements are recorded on the directions of 11 ants after one eye on each ant was 'trained' to learn the ant's home direction, then covered and the other eye uncovered. The data is stored as a list (first column) in the dataset fisherB10 from the circular package. Since the data are recorded in degrees we first convert to radians and then apply Watson's test for the von Mises distribution using the function watson:

```
> ants<- as.numeric (fisherB10[[1]])

> ants_radians<-0.0174532925*circular(ants)

> watson(ants,dist='vm')

    Watson's Test for the von Mises Distribution

Test Statistic: 0.025
```

P-value > 0.10

Since the p-value is greater than 0.05, we cannot reject the null hypothesis that the data are from the Von Mises distribution.

References

Li, G. N., & Hoffman-Kim, D. (2008). Evaluation of neurite outgrowth anisotropy using a novel application of circular analysis. Journal of neuroscience methods, 174(2), 202-214.

Marcoux, M. (2011). Narwhal communication and grouping behaviour: A case study in social cetacean research and monitoring (Doctoral dissertation, McGill University).

Sailer, U., Flanagan, J. R., & Johansson, R. S. (2005). Eye–hand coordination during learning of a novel visuomotor task. The Journal of neuroscience, 25(39), 8833-8842.

Back to Table of Contents

TEST 99 WATSON'S TWO-SAMPLE TEST OF HOMOGENEITY

Question the test addresses

Is the sample uniformly distributed or from the Von Mises distribution?

When to use the test?

To test a given distribution to determine the probability that it derives from a Von Mises or uniform distribution. The test uses a mean square deviation and is especially powerful for small sample sizes, unimodal and multimodal data. Note other circular distributions are the wrapped normal and the wrapped Cauchy distribution. These have similar properties to the Von Mises Distribution, but the Von Mises distribution can be parameterized to match any of the other distributions. The Von Mises Distribution is a popular choice because the concentration parameter has a close association to the mean vector length, and it has other convenient statistical properties similar to the linear normal distribution.

Practical Applications

Eastern Screech-Owl nest sites: Belthoff and Ritchison (1990) compare usec nest sites to randomly chosen unused nest sites to determine which features of nest tree/cavity and surrounding vegetation influenced nest site selection for the Eastern Screech-Owl (Otus asio). Over the period 1985-1987 Eastern Screech-Owl nest sites were located in the central Kentucky wildlife management area in Madison County, Kentucky. The area consists of small deciduous woodlots and thickets interspersed with cultivated fields. Nests sites were obtained by following radio- tagged adult Owls to nest cavities and by systematically inspecting tree cavities within the study area. Mean entrance orientation (direction) for screech-owl nest cavities and random cavities was 204.5 degrees and 48.5 degrees respectively. There was no significant difference in mean entrance orientation between used and unused sites (Watson's two-sample test of homogeneity p-value> 0.10).

Magnetic field and butterfly orientation: Srygley et al (2006) investigated whether migrating Aphrissa statira butterflies, captured over Lake Gatun, Panama, orient with a magnetic compass. Butterflies were collected during the migratory seasons of 2001, 2002 and 2003 (specifically 24 June-7 July 2001, 13 May-23 July 2002 and 21 May-6 June 2003). The researchers randomly selected butterflies by coin-flip to undergo an experimental or control treatment immediately prior to release over the lake. Butterflies in

the experimental group were swiped through a strong magnetic field. The distributions of orientations between the two groups were significantly different (Watson's test p-value < 0.001; control group contained 57 butterflies; experimental group contained 59 butterflies). The researchers conducted another experiment where they reversed the Magnetic Field. Again they found the distributions of orientations between the two groups was significantly different (Watson's test p-value < 0.001; control group contained 61 butterflies; experimental group contained 64 butterflies).

Idiopathic clubfoot in Sweden: Danielsson (1992) performed a prospective multicenter study in order to assess the cumulative incidence of Idiopathic clubfoot in Sweden over the years 1995 and 1996. The medical records of 280 children with clubfoot born during 1995– 1996 were collected and analyzed in the study. The distribution of clubfoot births by month was compared to other newborn births using Watson's two-sample test of homogeneity (p-value >0.5). The researchers conclude there was no significant difference in distribution of birth month between clubfoot children and all other live births in Sweden.

How to calculate in R

The functions watson.two{CircStats}and watson.two.test{circular}can be used to perform this test. They take the form watson.two (sample.1_radan, sample.2_radan, plot=FALSE) and watson.two.test (sample.1_radan, sample.2_radan). Note sample.1_radan and sample.2_radan represent the vector of angular measurements in radians. If the plot =TRUE, the empirica cumulative density functions of both samples are plotted.

Example: Orientation of barn swallows

Giunchi, D and Baldaccini (2004) investigate the role of visual and magnetic cues during the first migratory journey of the Juvenile barn swallow. Orientation experiments were performed in both local and shifted magnetic fields. The data is contained in the swallows list in the circular package. Let’s investigate the difference in distribution between the control group and the experimental group (shifted). The data can be put in suitable form by entering:

```
sample <- split(swallows$heading, swallows$treatment)

treatment<-circular(as.numeric (sample[[2]]) *0.0174532925)

control<-circular(as.numeric (sample[[1]]) *0.0174532925)
```

The test can be conducted using both functions by entering:

```
> watson.two(control,treatment, plot=FALSE)

      Watson's Two-Sample Test of Homogeneity

Test Statistic: 0.4044

P-value < 0.001
```

Or alternatively by typing:

```
> watson.two.test(control,treatment)

     Watson's Two-Sample Test of Homogeneity

Test Statistic: 0.4044

P-value < 0.001
```

The null hypothesis is rejected at the 5% level (p-value <0.001).

References

Belthoff, J. R., & Ritchison, G. (1990). Nest-site selection by Eastern Screech-Owls in central Kentucky. Condor, 982-990.

Danielsson, L. G. (1992). Incidence of congenital clubfoot in Sweden. Acta Orthopaedica, 63(4), 424-426.

Giunchi, D., & Baldaccini, N. E. (2004). Orientation of juvenile barn swallows (Hirundo rustica) tested in Emlen funnels during autumn migration. Behavioral Ecology and Sociobiology, 56(2), 124-131.

Srygley, R. B., Dudley, R., Oliveira, E. G., & Riveros, A. J. (2006). Experimental evidence for a magnetic sense in Neotropical migrating butterflies (Lepidoptera: Pieridae). Animal behaviour, 71(1), 183-191.

Back to Table of Contents

TEST 100 RAO'S TEST FOR HOMOGENEITY

Question the test addresses

Is the mean direction and dispersion of two or more circular samples different?

When to use the test?

To compare the mean direction and dispersion between two or more directional samples.

Practical Applications

Nocturnal passerine migration: Gagnon et al (2010) documented the pattern of nocturnal passerine migration on each side of the St. Lawrence estuary using Doppler radar. Doppler radar data on bird flight paths were collected from 29 July to 31 October 2003. The main flight track direction (i.e., the resulting direction between bird heading and wind drift) was determined at each Doppler elevation angle for two regions (Cote-Nord north and Gaspesie south) at three times per night: 1.5 hours past sunset, as well as 1/3 and 2/3 of the night length. To assess if mean flight directions and their variance differed between regions within a given period Rao's test of homogeneity was used. The researchers report the following results. For 1.5 hours (equality of means p-value =0.003, equality of dispersions p-value <0.001); For 1/3 of night length (equality of means p-value < 0.001, equality of dispersions p-value <0.001); For 2/3 of night length (equality of means p-value =0.001, equality of dispersions p-value <0.001). The researchers also used flight directions within each region. For Cote-Nord north (equality of means p-value =0.013, equality of dispersions p-value =0.293); For Gaspesie south (equality of means p-value =0.733, equality of dispersions p-value =0.688).

Whale behavior on observing a seismic ship: The behavior of a bowhead whale (Balaena mysticetus) as a function of distance from a seismic ship was investigated by Quakenbush et al (2010). During September of 2006, a satellite-tagged bowhead whale was in the vicinity of a seismic ship for 17 days. The whale was located 160 times during the seismic survey. Researchers collected data on the whale's velocity, turn angle relative to the seismic ship, and the dispersion in turn angles. To determine if the distribution of turning angles changed in dispersion between distance categories Rao's test for Homogeneity was used (p-value = 0.52). The researchers observe there to be no statistical relationship between whale behavior and distance from the seismic ship. This result, they conjecture, is

due to the ship shutting down seismic operations when the whale came closest.

Monkey learning: Zach et al (2012) compared responses of single cells in the primary motor cortex and premotor cortex of primates to interfering and noninterfering tasks. Two female monkeys (Macaca fascicularis) were trained on an 8-direction center-out reaching task, using a 2-joint manipulation at their elbow level. Measurements were taken during rotation (n = 127 cells from Monkey 1; n = 67 from Monkey 2), arbitrary association (n = 104 from Monkey 1, n = 36 from Monkey 2), rotation and arbitrary association (n = 76 from Monkey 1, n = 241 from Monkey 2) and rotation and opposite rotation sessions (n = 40 from Monkey 1, n = 100 from Monkey 2). The researchers calculated the signal-to-noise ratio (SNR) for different movement directions before and after learning the arbitrary association task alone, where movements were made to the same direction, but without any perturbation. No SNR trend toward any direction was observed (Rao's test for homogeneous distribution p-value >0.3).

How to calculate in R

The function rao.homogeneity{CircStats}and rao.test{circular}can be used to perform this test. They take the form rao.homogeneity (sample) and rao.test (sample.1, sample.2,…, sample.2). Note sample values should be represented as angular measurements in radians.

Example:

Let’s investigate the test using four samples from the Von Mises distribution. Sample x has a larger dispersion than the other samples:

```
set.seed(1234)

w <- list(rvonmises(300, circular(0), kappa=10))

x <- list(rvonmises(300, circular(0), kappa=20))

y <- list(rvonmises(300, circular(0), kappa=10))

z <- list(rvonmises(300, circular(0), kappa=10))

sample<-c(w,x,y,z)
```

The test can be conducted using both functions by typing:

> rao.homogeneity(sample)

Rao's Tests for Homogeneity

Test for E`quality of Polar Vectors:

Test Statistic = 4.91964

Degrees of Freedom = 3

P-value of test = 0.17778

Test for Equality of Dispersions:

Test Statistic = 178.886

Degrees of Freedom = 3

P-value of test = 0

> rao.test(w,x,y,z)

Rao's Tests for Homogeneity

Test for Equality of Polar Vectors:

Test Statistic = 2.3665

Degrees of Freedom = 3

P-value of test = 0.4999

Test for Equality of Dispersions:

Test Statistic = 89.2427

Degrees of Freedom = 3

P-value of test = 0

The functions report different p-values for the equality of polar vectors; however, they report similar p-values for the test of equality of Dispersions. The null hypothesis of homogeneity across the four samples is therefore rejected at the 5% level.

References

Gagnon, F. G. F., Ibarzabal, J. I. J., Bélisle, M. B. M., & Vaillancourt, P. V. P. (2010). Autumnal patterns of nocturnal passerine migration in the St. Lawrence estuary region, Quebec, Canada: a weather radar study. Canadian Journal of Zoology, 89(1), 31-46.

Quakenbush, L. T., Small, R. J., Citta, J. J., & George, J. C. (2010). Satellit tracking of western Arctic bowhead whales. Satellite Tracking of Western Arctic Bowhead Whales, 69.

Zach, N., Inbar, D., Grinvald, Y., & Vaadia, E. (2012). Single Neurons in M1 and Premotor Cortex Directly Reflect Behavioral Interference. PloS one 7(3), e32986.

Back to Table of Contents

TEST 101 PEARSON CHI SQUARE TEST

Question the test addresses

Is the sample from a normal distribution?

When to use the test?

To test of the null hypothesis that the sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.

Practical Applications

Fuzzy logic skin incision: Zbinden et al (1995) contrast the use of fuzzy logic to control arterial pressure in 10 patients during intra-abdominal surgery by automatic adjustment of the concentration of isoflurane in gas. Experiments contrasting human adjustment of gas concentration with fuzzy logic adjustment were carried out to two types of incision in live patients – skin incision and non-skin incision. Measurement values were calculated as the difference of the measured minus the desired pressure value divided by the desired pressure value. The distribution of skin incision for fuzzy logic and Human's was non normal (Pearson chi square test of normality p-value < 0.05 in both cases). Similar results were observed for non-skin incision (Pearson chi square test of normality p-value < 0.05 in both fuzzy logic and Human experiments).

Genetic algorithm image enhancement: Munteanu and Rosa (2001) develop a method for image enhancement of gray-scale images using genetic algorithm based model. Several greyscale images were used in the analysis (plane, cape, lena, goldhill, mandrill, boat). For each of these images the model residuals were tested against normality using the Pearson chi-squared test. The researchers find outliers to the normal distribution of the residuals occur in the case of the goldhill and lena images, which were the only images not to pass the Pearson chi-square test for normality (p-value <0.05).

Explosion versus earthquake identification: Identifying explosions versus earthquakes is investigated using the ratio of Pg/Lg waves between frequencies of 0.5 and 10 Hz using 294 by Taylor (1996). Nevada Test Site explosions and 114 western U.S. earthquakes recorded at four broadband seismic stations located at distances of about 200 to 400 km are used in the analysis. Event magnitudes ranged from about 2.5 to 6.5 and propagation paths for the earthquakes range from approximately 175 to 1300 km. The Pearson chi-square test was used to test for normality of the

log(Pg /Lg) ratios. In general, it was observed that the 1-2, 2-4, and 4-6 Hz frequency bands were normally distributed (p-value >0.05).

How to calculate in R

The function pearson.test{nortest} or pchiTest{fBasics}can be used to perform this test. It takes the form pearson.test(sample) or pchiTest (sample).

Example: testing against a normal distribution

Enter the following data:

> sample <-c(-1.441,-0.642,0.243,0.154,-0.325,-0.316,0.337,-0.028,1.359,-1.67,-0.42,1.02,-1.15,0.69,-1.18,2.22,1,-1.83,0.01,-0.77,-0.75,-1.55,-1.44,0.58,0.16)

The test can be conducted as follows:

> pchiTest (sample)

Title:

Pearson Chi-Square Normality Test

Test Results:

PARAMETER:

Number of Classes: 8

STATISTIC:

P: 2.84

P VALUE:

Adhusted: 0.7246

Not adjusted: 0.8994

The adjusted p-value at 0.7246 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution. Alternatively using pearson.test

> pearson.test(sample)

Pearson chi-square normality test

data: sample

P = 2.84, p-value = 0.7246

The p-value at 0.7246 is greater than 0.05, therefore do not reject the null hypothesis that the data are from the normal distribution.

References

Munteanu, C., & Rosa, A. (2001). Evolutionary image enhancement with user behavior modeling. ACM SIGAPP Applied Computing Review, 9(1), 8 14.

Taylor, S. R. (1996). Analysis of high-frequency Pg/Lg ratios from NTS explosions and western US earthquakes. Bulletin of the Seismological Society of America, 86(4), 1042-1053.

Zbinden, A. M., Feigenwinter, P., Petersen-Felix, S., & Hacisalihzade, S (1995). Arterial pressure control with isoflurane using fuzzy logic. British journal of anaesthesia, 74(1), 66-72.

Back to Table of Contents