Statistical Hypothesis Testing [Nonparametric]

Satishkumar L. Varma
Professor, Department of Computer Engineering
PCE, New Panvel
www.sites.google.com/view/vsat2k
www.vsat2k.wordpress.com
www.vsat2k.moodlecloud.com

Overview: Statistical Nonparametric Hypothesis Testing

- Introduction to Nonparametric Statistics
- Uses of Nonparametric Tests
- Advantages of Nonparametric Test
- Disadvantages of Nonparametric Test
- Nonparametric Tests
  - Sign test
  - U test

Learning Objectives

- By the end of this topic, I will be able to . . .
  - Explain what a nonparametric hypothesis test is, and why we use it
  - Describe what is meant by the efficiency of a nonparametric test
  - Explain commonly used Nonparametric Test Procedures
  - Know uses, advantages and disadvantages of nonparametric tests
  - Perform hypothesis tests using nonparametric procedures
  - Know when to apply parametric & nonparametric tests

Statistical Analysis: Classification

- **Descriptive Statistics**
  - Graphical
    - Organizing and presenting the data
    - Example: histogram, box plot, probability plot
  - Numerical:
    - Summarizing the sample set
    - Example: mean, median, mode, range, quartile, variance, stad.dev
- **Inferential Statistics**
  - Estimation:
    - Estimate parameters of the pdf along with its confidence region
  - Hypotheses Testing:
    - Making judgements about f(x) and its parameters

Motivation for Hypothesis Testing

- **Need of (Motivation for) Hypothesis Testing**
  - Will an investment in MF yield > desired value?
  - Is incidence of diabetes > among male than female?
  - Are women > than male to change mobile service provider?
  - Has the efficiency of a pump < form its original value due to aging?
- **Statistical tests are either**
  - Parametric or
  - Nonparametric (distribution-free hypothesis tests)
- **When the conditions for the parametric test are met**
  - It is preferable to perform parametric test as opposed to nonparametric test

Need of Nonparametric Tests

- Motivation of performing nonparametric test
  - Should not perform a parametric hypothesis test (t test for μ)
    - if the conditions are not met
  - Data analyst don’t take a chance and use a parametric test
    - when the conditions may not be satisfied?
  - So, there are advantages and disadvantages to each method

Nonparametric Tests

- Distinguish Between Parametric & Nonparametric Test

	Parametric tests		Nonparametric tests
1	Depends of probability distribution	1	Distribution-free methods
2	Require more conditions to be satisfied	2	Require fewer conditions than their parametric counterparts
3	Involve population parameters (Mean)	3	Do not involve population parameters
4	Use Mean, P	4	Use median
5	Have stringent assumptions (Normality)	5	Data measured on any scale (Ratio/Interval, Ordinal/ Nominal)
6	More efficient (at the cost of more stringent required conditions for the parametric tests)	6	Less efficient
7	Require either a normal population or a large sample ($n \geq 30$)	7	Require neither a normal population nor a large sample ($n < 30$)

Nonparametric Tests

- Advantages
  - Easy to understand
  - No lengthy calculations
  - No requirement of assumption of distribution
  - Applicable to all kind of data
- Disadvantages
  - Less efficient in comparison with parametric tests
    - for a given level of significance, it require a larger sample size to reject H_0
  - Result may or may not provide actual answer because it is distribution free

Parametric & Nonparametric Test Procedures

- **Parametric tests**
  - t test
  - Z test
  - x^2 Test
  - F test (Analysis of variance)
  - Linear correlation
- **Nonparametric tests**
  - Sign Test
  - Mann-Whitney U Test
  - Mann-Whitney-Wilcoxon Test
  - Wilcoxon Signed-Rank Test
  - Wilcoxon Rank Sum Test
  - Kruskal-Wallis H-Test
  - Rank Correlation Test
  - Runs Test

Nonparametric Tests Requirements

- Situation in which nonparametric tests are used
  - Parametric tests are not satisfied
  - Test hypothesis does not have any distribution
  - Population don not follow a particular distribution (say normal distribution)
  - Neither a normal population nor a large sample available
  - Data is categorical (qualitative)
  - Distributions are not available
  - Requirement of a quick data analysis
  - Unscaled data
  - Tends to be easier than their parametric counterparts

Nonparametric Test Procedures

- Sign Test Statistics
- Mann-Whitney U Test
- Wilcoxon Signed-Rank Test
- Wilcoxon ranak sum test
- Mann-Whitney-Wilcoxon Test
- Kruskal-Wallis H Test
- Rank Correlation Test
- Runs test

Sign Test

- Objectives
  - 1. Perform the sign test for a single population median
  - 2. Perform the sign test for matched-pair data from two dependent samples
  - 3. Perform the sign test for binomial data
- Sign test is a nonparametric hypothesis test
- Here the original data are transformed into plus or minus signs
- Test is based on signs and not magnitudes
- The sign test may be conducted for
  - 1. a single population median
  - 2. matched-pair data from two dependent samples, or
  - 3. binomial data
- Note: Sign test is a hypothesis test for the population median, not the population mean

Steps for Performing Sign Test

- Determine whether the conditions required for parametric test are not met
- Write the null and alternative hypotheses for nonparametric sign test
- Change the data values to plus or minus signs
- Perform Sign Test (may be conducted for)
  - **The population median M**
    - Step 1: State the hypotheses
    - Step 2: Find the critical value and state the rejection rule
    - Step 3: Find the value of the test statistic
    - Step 4: State the conclusion and the interpretation
  - **The matched-pair data from two dependent samples (median of differences)**
    - Step 1: State the hypotheses
    - Step 2: Find the critical value and state the rejection rule
    - Step 3: Find the value of the test statistic
    - Step 4: State the conclusion and the interpretation
  - **The p-value method**
    - Step 1: State the hypotheses
    - Step 2: Find the p-value using technology
    - Step 3: State the conclusion and the interpretation
  - **Binomial Data**
    - Step 1: State the hypotheses
    - Step 2: Find the critical value and state the rejection rule
    - Step 3: Find the value of the test statistic
    - Step 4: State the conclusion and the interpretation

Sign Test for a Single Population Median

- Example
  - A random sample size n = 7 from the population of daily sell of banana on the road is given.
  - We are interested to find whether the conditions required for the parametric hypothesis test are met
    - Is population normal?
    - Is sample size at least 30?
- If conditions are not met then we are left with nonparametric hypothesis tests
  - Perform Sign Test for the population median M
  - Let we are interested in testing whether population median M of sells per day < 10
    - Step 1: State the hypotheses
    - Step 2: Change data value to plus or minus sign
    - Step 3: Find the critical value and state the rejection rule
    - Step 4: Find the value of the test statistic
    - Step 5: State the conclusion and the interpretation

Day	1	2	3	4	5	6	7
Sell	7	5	4	25	9	1	3

Step 1: State the hypotheses

- Only requirement for performing the sign test for the population median M is
  - The sample data have been randomly selected and
  - It is not necessary to have a population that is normally distributed
- Interested in testing whether the population median M number of sell per day < 10
  - Write the null and alternative hypotheses for this test
    - $H_0 : M = 10$
    - $H_a : M < 10$
- Hypotheses for the sign test for the population median M is in one of the forms

Null hypothesis	Alternative hypothesis	Type of test
$H_0 : M = M_0$	$H_\alpha : M > M_0$	Right-tailed test
$H_0 : M = M_0$	$H_\alpha : M < M_0$	Left-tailed test
$H_0 : M = M_0$	$H_\alpha : M \neq M_0$	Two-tailed test

- M0 is the value of the population median M for which a claim is being made

Step 2: Change data value to plus or minus sign

- Changing the data values to plus or minus signs
  - if (sell per day < 10) assign a minus sign
  - if(sell per day > 10) assign a plus sign
  - if(sell per day = 10) ignore data values
  - So, we have 6 minus signs, 1 plus sign and sample size = 6 + 1 = 7

Day	1	2	3	4	5	6	7
Sell	7	5	4	25	9	1	3
Sign	-	-	-	+	-	-	-

Step 3: Find the critical value and state the rejection rule

- Small-Sample Case (sample size n <= 25)
  - Use table to find the critical value S_{crit}
  - Choose the column with the appropriate level of significance (a) and
  - The applicable one-tailed or two-tailed test
  - Then select row with the appropriate sample size n = # pluses & minuses
  - The number in that row and column is our critical value S_{crit}
  - The rejection rule is to reject H_0 if $S_{data} \leq S_{crit}$

Step 3: Find the critical value and state the rejection rule

- Large-Sample Case (sample size n > 25)
  - Use the standard normal table
  - Z_{crit} value for this sign test is always found in the left tail of the st.nor.dist.
    - So that Z_{crit} is always less than 0
  - For a left-tailed test or a right-tailed test
    - the critical value Z_{crit} is the value of Z with area α to the left of it
  - For a two-tailed test
    - the critical value Z_{crit} is the value of Z with area $\alpha/2$ to the left of it
  - Table contains values of Z_{crit} for some common values of α
  - The rejection rule is to reject $H0$ if $Z_{data} \leq Z_{crit}$

Step 4: Find the value of the test statistic

- Find the value of the test statistic S_{data}
  - For small-Sample Case (sample size n <= 25) find S_{data} as

Type of test	Test statistic S_{data}
Right-tailed test	S_{data} = number of minus signs
Left-tailed test	S_{data} = number of plus signs
Two-tailed test	S_{data} = number of minus signs or plus signs, whichever is smaller

  - For large-Sample Case (sample size n > 25) find test statistic Z_{data} as

$$Z_{data} = \frac{\left(S_{data} + 0.5\right) - \frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

Step 5: State the conclusion and the interpretation

- State the conclusion and the interpretation
  - Compare the test statistic with the critical value, using the rejection rule
  - If(H0 = rejected)
    - State=> Evidence exists that [whatever Ha says]
  - If(H0 ≠ rejected)
    - State=> There is insufficient evidence that [whatever Ha says]

Small-Sample size n <= 25 sign test for population median

- Example 2: Use the sign test to determine
  - whether the population median M number of sell per day < 10
  - using level of significance $\alpha = 0.05$
- The hypotheses are H0 : M = 10 and Ha : M < 10
- Find the critical value and state the rejection rule
  - The total number of plus signs and minus signs is $n = 6+1=7$
  - which is not greater than 25
  - so we use the small-sample case
  - We have a one-tailed test, with $\alpha = 0.05$ and $n = 7$
  - which gives us $S_{crit}=1$ (see Figure on next slide)
  - The rejection rule is to reject H_0 if $S_{data} \leq 1$

Day	1	2	3	4	5	6	7
Sell	7	5	4	25	9	1	3
Sign	-	-	-	+	-	-	-

Small-Sample size n <= 25 sign test for population median

- **Find the value of the test statistic**
  - We have a left-tailed test, and so
  - from above S_{data} Table, our test statistic is
  - S_{data} = number of plus signs = 1
- **State the conclusion and the interpretation**
  - The value of our test statistic is S_{data} =1 which is = 1
  - so we reject H_0
  - Evidence exists that the population median number of sell < 10 per day

	α			
n	0.005 (one tail) 0.01 (two tails)	0.01 (one tail) 0.02 (two tails)	0.025 (one tail) 0.05 (two tails)	0.05 (one tail) 0.10 (two tails)
1	*	*	*	*
2	*	*	*	*
3	*	*	*	*
4	*	*	*	*
5	*	*	*	0
6	*	*	0	0
7	*	0	0	0
8	0	0	0	1

Sign Test Using Excel

- Excel does not have a built-in sign test function
- However, use the BINOMDIST function to calculate the *p*-value for a sign test
  - **Step 1 On the main menu bar, click fx**
    - Where it says Search for a function, type BINOMDIST and click Go
    - Where it says Select a function, select BINOMDIST and click OK
  - **Step 2 In the dialog box, enter the following values:**
    - For Number_s, enter the value of S_{data} (from Table 3)
    - For Trials, enter n = the sum of the number of pluses and minuses
    - For Probability_s, enter 0.5
    - For Cumulative, type True
  - **Step 3 Click OK**
    - The result is the *p*-value for a one-tailed test
    - Double this value for a two-tailed test
    - Reject *H*0 if the *p*-value < *alpha*

Efficiency of Nonparametric Hypothesis Test

- **Efficiency of Parametric tests are > corresponding nonparametric tests**
- Nonparametric test Efficiency, $\eta = \frac{sample\ size\ required\ in\ parametric\ test}{samle\ size\ required\ innonparametric\ test}$
  - Provide conditions for parametric and nonparametric tests have been met
  - in order to achieve the same result (such as correctly rejecting the H_0)

Efficiency of nonparametric tests vs. parametric tests

- **Efficiency of Parametric tests are > corresponding nonparametric tests**

Parametric Test	Nonparametric Test	Situation	η
t test or Z test	Sign test	Matched pairs (dependent samples)	0.63
t test or Z test	Wilcoxon signed rank test	Matched pairs (dependent samples)	0.95
t test or Z test	Wilcoxon rank sum test	Two independent samples	0.95
Analysis of variance (F test)	Kruskal-Wallis test	Several independent samples	0.95
Linear correlation	Rank correlation test	Correlation	0.91
No parametric test	Runs test	Randomness	-

Examples

- goo.gl/f5rC77

Summary

- For solved examples:
  - goo.gl/f5rC77

Bibliography

[1] Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, 3rd edition Publisher: Morgan Kaufmann; 3 edition
[2] Business Intelligence, 2/E; Efraim Turban, Ramesh Sharda, Dursun Delen, David King; pearson Education
[3] Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with Xlminer; 2nd edition, Galit Shmueli, Nitin R. Patel and Peter C. Bruce; John Wiley
[4] Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Author: Berry, Gordon S. Linoff, Format: Paperback, 648 pages, Edition: 3; Publisher: John Wiley & Sons Inc.
[5] Robert Groth, Data Mining: Building Competitive Advantage, Prentice Hall, 2000.
[6] P. N. Tan, M. Steinbach, Vipin Kumar, "Introduction to Data Mining", Pearson Education
[7] Alex Berson and Smith, "Data Mining and Data Warehousing and OLAP", McGraw Hill Publication
[8] E. G. Mallach, "Decision Support and Data Warehouse Systems", Tata McGraw Hill.
[9] Michael Berry and Gordon Linoff "Mastering Data Mining- Art & science of CRM", Wiley Student Edition

Thank You.