**INTRODUCTION TO NON-PARAMETRIC STATISTICS**

by Simon Moss

|  |
| --- |
| **Introduction** |

 Non-parametric statistics comprise a collection of tests, such as the Mann-Whitney U test, the Kruskall-Wallis one-way ANOVA, the sign test, and many other procedures. Yet, researchers do not agree on

* the precise definition of nonparametric statistics
* which tests belong to this collection
* when these tests are most applicable

**Limitations when the outcome measure is a ranking**

 Nevertheless, as most researchers agree, whenever data are rankings, nonparametric statistics are often suitable. To illustrate, consider a researcher who plans to investigate whether physical fitness increases the likelihood that research candidates will submit their thesis on time. To answer this question, the researcher organizes a 5-km run and records the placing or ranking of each candidate. Four years later, the researcher determines which of these candidates submitted their thesis on time. The following table presents an extract of these data.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | … |
| Completed |  Yes | No | No | Yes | No | Yes | Yes | No | No |  |

Usually, to analyse whether some numerical measure, such as running speed, differs between two groups, researchers would conduct an independent t-test. But, in this instance, this independent t-test is unsuitable because

* t-tests assume the data within each group are independent of one another
* for example, whether one score is high should not affect whether another score is high
* but, in this instance, the scores are not independent of one another
* to illustrate, if one person is ranked 1, the other individuals must have ranked higher than 1, violating the assumption of independence.

Consequently, if researchers conducted an independent t-test in this circumstance, the p value would not be accurate. Instead, whenever the data are rankings, researchers often apply nonparametric statistics. To illustrate, in this instance

* researchers would conduct a nonparametric statistic called the Mann-Whitney U test
* this test, for example, might generate a p value of .021
* because this p value is less than .05, the researcher would conclude the rankings differ between candidates who submitted on time and candidates who had not submitted on time

Nevertheless, this finding does not indicate whether rankings are higher or lower in the candidates who complete on time. To resolve this problem, the researcher might calculate the median rank in each group. For example, the researcher might show

* the median is 14 for candidates who had not submitted—that is, half these individuals ranked lower than 14
* but, the median is 6 for candidates who had submitted—that is, half these individuals ranked higher than 6
* so, the candidates who had submitted tended to complete the race faster

**Limitations when the outcome measure is categorical**

 As the previous section demonstrates, whenever the outcome is a ranking, non-parametric statistics are appropriate. Whenever the outcome is categorical, sometimes called nominal, another suite of tests is appropriate. Some, but not all, researchers utilize the term nonparametric to describe these tests as well. To illustrate, consider a researcher who wants to assess whether research candidates, after completing a workshop on R, prefer SPSS or R to analyse data. As the following table indicates,

* before the workshop, most of the candidates preferred SPSS
* after the workshop, most of the candidates preferred R
* note, in this example, the participants before the workshop and after the workshop are the same people—sometimes called a repeated measures design

|  |  |  |
| --- | --- | --- |
| Time | Before workshops on R | After workshops on R |
| Prefers SPSS | 47 participants | 19 participants |
| Prefers R | 13 participants | 28 participants |

 To determine whether this change in attitudes towards SPSS and R changes over time, researchers would conduct a nonparametric test called a McNemar test. If the p value was less than .05, the researcher would conclude the degree to which candidates prefer R instead of SPSS increased after the workshop.

**Rating scales and ordinal scales**

Some researchers like to use non-parametric tests when the outcome measure is a rating scale. To illustrate, suppose 6 people indicated their level of happiness on a 5-point scale, in which 5 implies very happy. The following table displays their responses

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Adam | Betty | Carl | Donna | Ernie | Fred |
| 1 | 1 | 2 | 3 | 4 | 5 |

 We can derive only limited information from these data. In particular,

* we can order the people from most happy to least happy; Fred is happier than Ernie for example
* we cannot, however, readily comment on the magnitude of difference between people
* for instance, we cannot be certain the extent to which Fred is happier than Ernie equals the extent to which Ernie is happier than Donna
* that is, these ratings were mainly designed to order individuals rather than specify the quantitative difference between individuals precisely
* consequently, rating scales are sometimes called ordinal; they clarify the order or ranking of individuals.

Classical tests will assume the difference between 5 and 4 is equivalent to the difference between 4 and 3; therefore, when the outcome is ordinal, classical tests, such as t-tests, might not be suitable. Instead, tests that examine rankings, such as the Mann-Whitney U test, may be more appropriate.

But actually, most researchers do not adopt this perspective. Instead, as most researchers assume, when participants complete rating scales, they assume the intervals between consecutive numbers are roughly equal. The difference between 5 and 4 mirrors the difference between 4 and 3, and so forth. So, if the outcome is a rating scale, most researchers would apply classical tests rather than nonparametric tests. These ratings are different to rankings because

* ratings are independent of one another; if one person scores a 1, the likelihood that another person scores a 1 does not change
* rankings are not independent; if one person is ranked 1, the likelihood that another person is ranked 1 is zero
* classical tests assume the scores are independent of one another

**Violations of assumptions about frequency distributions**

 To illustrate, many classical tests assume the outcome measure conforms to a pattern called a normal distribution, depicted in the following graph. In this graph, the Y axis indicates the number of people who specified each level of happiness. Other tests assume a different pattern.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |

But…

* if this assumption is violated, the test will generate an inaccurate p value
* however, in general, nonparametric tests do not assume the outcome measure conforms to a specific pattern, such as a normal distribution
* consequently, nonparametric tests generate accurate p values, regardless of whether these assumptions are fulfilled.
* thus, when assumptions about these distributions are violated, many researchers will use non-parametric tests instead.

Yet, not all researchers abstain from classical tests, such as t-tests, when these assumptions are violated. In particular

* even when assumptions are violated to a moderate extent, the p values of classical statistical tests are usually quite accurate.
* in these circumstances, researchers could utilize a classical test, such as a t-test, but utilize a more conservative level of alpha, such as .01.
* in other words, the researcher might decide the findings are significant only if p is less than .01
* here is the rationale: if the p value they generate is less than .01 when the assumption is violated, the p value would probably still be less than .05 had the assumption not been violated.

**The definition of non-parametric statistics**

As the previous discussion shows, the definition of non-parametric tests remains contentious. For example

* some people utilize this term to indicate the test does not assume a specific distribution, such as a normal distribution
* but some tests, such as the 2 test of independence, assume a normal distribution, but this assumption is invariably fulfilled if the sample size is sufficiently high
* therefore, whether these tests should be regarded as non-parametric is contentious
* some people utilize the term nonparametric to refer to techniques that determine which the underlying model is not specified in advance but might change during the analysis—such as nonparametric Bayesian models

Despite this complexity, this document will primarily revolve around techniques that are suitable whenever the outcome measure is ordinal or categorical.

|  |
| --- |
| **How to decide which non-parametric test to apply** |

 In general, nonparametric tests are not hard to conduct or to interpret. The most challenging phase is to decide which test to utilize. The following tables can be utilized to guide this decision.

|  |  |
| --- | --- |
|  | Tests that compare groups on some outcome measure |
|  | Outcome measure is ordinalsuch as a ranking | Outcome measure is categorical |
|  | 2 groups | 3 groups or more | 2 groups | 3 groups or more |
| Each group comprises different and unrelated individuals or units | * The Mann-Whitney U test assesses whether scores are higher in one group compared to another group
* Moses extreme reactions test assess whether scores are more extreme or variable in one group compared to another group
 | * The Kruskall-Wallis H test assesses whether scores differ across three groups
* The Jonckheere-Terpsra test is more powerful than is the Kruskall-Wallis H test if the researcher can order the groups from highest to lowest scores in advance
 | * 2 test of independence
 | * 2 test of independence
 |
| Each group comprises the same, or matched, individuals or units | * The sign test assesses whether one group is greater on the outcome than is the other group more often than vice versa
* The Wilcoxon signed rank test is applicable if the outcome measure is numerical rather than ordinal—or the differences between the two groups are ranked
 | * The Friedman test assess whether the magnitude of scores tends to differ across the groups
 | * The McNemar test assesses whether two groups differ on a dichotomous outcome measure—a measure with one of two possible outcomes
 | * Cochrane’s Q test assesses whether three or more groups differ on a dichotomous outcome measure
 |

|  |
| --- |
| Tests that assess the degree to which two variables are related—like a correlation |
| Variables are ordinalsuch as a ranking | Variables are categorical |
| 2 variables | 3+ variables | 2 variables | 3+ variables |
| * Kendall’s rank correlation coefficient assesses the extent to which two ordinal variables are related to each other, on a scale from 0 to 1
* Spearman rank order correlation also assesses the extent to which two ordinal variables are related to each other, on a scale from 0 to 1
 | * Kendall’s W assesses the extent to which three or more ordinal variables are related to each other, on a scale from 0 to 1
 | * Cohen’s kappa assesses the extent to which two categorical variables are related to each other, on a scale from 0 to 1
* Statisticians have developed many alternatives to Cohen’s kappa as well
 |  |

When you apply these tables, you need to consider a few caveats. In particular

* tests that analyze ordinal variables can also analyze numerical variables. The reason is that any numerical variables, such as heights, can be converted to ranks. For example, 152 cm, 153 cm, and 189 cm can be converted to ranks of 1, 2, and 3
* these tables do not include all nonparametric tests, such as permutation tests.
* these tables also do not include binomial tests, run tests, or 2 goodness of fit tests—tests that assess whether one group diverges from some expected pattern

**SPSS expert system**

 Rather than utilize these tables, some software can determine which test to conduct. For example, if you use SPSS, choose the “Analyse” menu and then “Nonparametric tests”. Then select

* “One sample” if you want to assess whether one group diverges from some expected pattern
* “Independent samples” if you want to compare groups that comprise different participants, or
* “Related samples” if you want to compare groups with the same or matched participants.

You will receive a screen with three options at the top left: Objective, Fields, and Settings. Choose “Fields” to generate the following screen. Then, specify your variables in the relevant boxes.



 Now choose settings and then an option entitled “Automatically choose the tests based on the data”. After you press “Run”, SPSS will choose the right test and then conduct this test as well. If this option does not work

* Check that, for each column in your data file, you have selected the right level of measurement in “Variable View”.
* For example, your ranking variables should be labelled as “Ordinal”
* Your categorical variables should be labelled as “Nominal”
* Otherwise, you might need to choose and complete these tests yourself.

|  |
| --- |
| **Illustration with a Mann Whitney U test in SPSS** |

 Once you have chosen a suitable non-parametric test, the procedure is not hard to conduct. Rather than illustrate all possible non-parametric tests, this section will show how you can undertake one of these tests: the Mann Whitney U test, designed to compare two groups on a ranking or ordinal variable. This illustration should instil in you the confidence to complete other nonparametric tests as wlels.

**Enter the data**

First, you obviously need to enter the data. The following data file corresponds to the previous example, designed to assess whether physical fitness increases the likelihood that research candidates will submit their thesis on time. Specifically

* the first column specifies the placing or ranking of each candidate during the 5-km run.
* the second column indicates whether the individuals submitted their thesis on time, denoted as 1, or did not submit their thesis on time, denoted as 0.
* if you use SPSS, you could enter these data. If not, follow the discussion anyway.



**Conduct the test**

 To conduct the test, choose the “Analyse” menu, “Nonparametric Test”, and then “Legacy dialogues”, generating a series of options, such as chi-square, binomial, runs, and so forth.



In this instance, because you want to compare two independent groups, choose “2 Independent Samples”. This procedure will generate the following screen.



Then, to conduct the test, you would simply

* specify the “Test variable”—in this instance “Rank in Race”
* specify the “Grouping variable”—in this instance, “Submitted\_on\_time”. You would also press “Define groups” to indicate that you want to compare groups 1 and 0
* press OK to generate the following output



 As this output shows, the p value, designated as the “Asymp Sig” or asymptotic significance is less than .05. Consequently

* you would conclude the two groups differ on the rankings
* as the top table shows, the mean rank is lower in Group 1—the candidates who submitted on time
* hence, the candidates who submitted on time ran faster.

**Other tests**

 To conduct other tests, the procedure is similar, besides a few exceptions. The following table outlines some of these exceptions.

|  |  |
| --- | --- |
| Test | Details |
| The 2 test of independence | To conduct this test in SPSS* choose “Analyze”, “Descriptive Statistics”, and “Cross-tabs”
* choose the relevant variables
* select “Statistics” and specify “Chi-square”
* Press continue and OK
 |
| Runs test | * Determines whether a sequence, such as head, tails, tails, tails, heads, and heads, is random or not
* Very long or very short sequences of repetitions generate significant p values
 |

|  |
| --- |
| **Illustrations of nonparametric tests in R** |

 Unsurprisingly, you can also use R to conduct a range of nonparametric tests. The following table presents some examples of the R codes that can be used to conduct nonparametric tests. To conduct the range of nonparametric tests available, you may need to install and load two key packages: npsm and Rfit, using these commans

* install.packages(“npsm”)
* library(npsm)
* install.packages(“Rfit”)
* library(Rfit)

|  |  |  |
| --- | --- | --- |
| Test | Code | Notes |
| Mann-Whitney U test | * wilcox.test(rank~group)
 | * “rank” is the ordinal variable
* “group” is the grouping variable
 |
| Wilcoxon signed rank test | * wilcox.test(time1, time2, paired=TRUE)
 | * time 1 and time 2 are the numeric outcome measures at two times
 |
| Kruskall-Wallis test | * kruskal.test(rank~group)
 | * “rank” is the ordinal variable
* “group” is the grouping variable
 |

For more comprehensive information about how to use R to conduct non-parametric statistics, Google “Nonparametric Statistical Methods Using R – Pindex”