ABSTRACT

Data collected in many biology laboratory classes are on ratio or interval scales where the size interval between adjacent units on the scale is constant, which is a critical requirement for analysis with parametric statistics such as t-tests or analysis of variance. In other cases, such as ratings of disease or behavior, data are collected on ordinal scales in which observations are placed in a sequence but the intervals between adjacent observations are not necessarily equal. These data can only be interpreted in terms of their order, not in terms of the differences between adjacent points. They are unsuitable for parametric statistical analyses and require a rank-based approach using nonparametric statistics. We describe an application of one such approach, the Kruskal-Wallis test, to biological data using online freeware suitable for classroom settings.

Key Words: Kruskal-Wallis test; ordinal scales; parametric statistics; nonparametric statistics; inquiry teaching.

Introduction

Inquiry teaching in high school and university biology classes requires students to consider their experimental design, the data they will collect, and the statistical analysis they will use to draw conclusions from the data (Kamin, 2010; Giese, 2012; Bennett, 2015). Failure to consider all steps before data collection risks a design that does not address the central hypothesis, collecting data inadequate for the planned analysis, or using an incorrect analysis that may give misleading results (Ruxton & Colegrave, 2011).

Here, we focus on an issue we believe is often neglected in planning – choosing an appropriate statistical analysis for data collected on an ordinal scale, as opposed to a ratio or interval scale. We begin by describing briefly the different types of scale that may be used with biological data before explaining why rank-based tests, as opposed to the better-known t-test and analysis of variance (ANOVA), are needed for analyzing ordinal data. Example data are then analyzed with the rank-based Kruskal-Wallis test, using online freeware suitable for classroom settings.

Types of Scale in Biology

Biological data fall into four broad types: ratio scale, interval scale, ordinal scale, and nominal (Zar, 2010). In ratio scales the difference between any two sequential points on the scale is the same irrespective of which two sequential points we pick. For example, if one measures height above sea level, the difference between a height of 70 m and a height of 71 m is the same as the difference between 83 m and 84 m. Furthermore, ratios are meaningful. A point 60 m above sea level is three times as high as a point 20 m above sea level. On an interval scale the difference between any two sequential points on the scale is the same irrespective of which two sequential points we pick, but ratios no longer have a clear meaning. On the Celsius scale a temperature of 26° is not “twice as hot” as a temperature of 13°. Ordinal data, by contrast, place observations in an ordered sequence but the distance between adjacent observations is not measurable. For example, if one records the order in which five runners finished a race, this gives no information about the relative times taken by each runner. It would be wrong to assume that the time between the first and the second runner was the same as that between the second and the third. Lastly, nominal data arise when observations are placed in unordered categories such as those for sex (male, female, intersex) or eye color (brown, blue, green, etc.).

The type of scale in which data are measured restricts the range of statistical analyses that may be appropriate. Most students have little difficulty in appreciating that nominal data such as sex cannot...
be analyzed using t-tests or ANOVAs, not least because the data cannot be forced into these tests in statistical software. However, numerical coding of ordinal data can obscure their basic distinction from ratio or interval scale data. While it is possible to force the numbers into t-tests or ANOVAs, such approaches are incorrect because ordinal data violate key assumptions of these tests and the tests can generate misleading results (Shah & Madden, 2004).

**Why Parametric Tests Are Inappropriate for Ordinal Data**

In parametric statistical tests the numerical data come from populations with particular probability distributions, the parameters of which may be the subject of the test or constrained by its assumptions. Both t-tests and ANOVAs are parametric in this sense, since they assume numerical data sampled from normally distributed populations, the hypotheses concern the distribution means, and ANOVA (at least) assumes that the distribution variances are equal. The term nonparametric is somewhat elastic, but we use it here to refer to statistical tests that do not make assumptions about population distributions. That is, we use nonparametric in the sense of “without distribution assumptions.” The only statistical assumptions required are those concerning sampling (e.g., simple random sampling within groups, independence between groups) and perhaps size (e.g., a minimum group size may be needed so that the sampling distribution of the test statistic is sufficiently close to a known probability distribution). Accordingly, the following comments focus on difficulties that can arise when ordinal data are analyzed by parametric methods—such methods assume population probability distributions that are meaningful only for (genuinely) numerical data.

Use of parametric tests for analyzing ordinal data leads to problems of (1) interpretation and (2) potentially misleading results. Regarding interpretation, differences between ordinal values cannot be interpreted in a quantitative sense, nor can we compare differences between ordinal values. Returning to the earlier example of the order in which runners finish a race, we do not know whether the difference between first and second place is “the same” as the difference between third and fourth place, nor is it meaningful to suggest that places 2 and 4 “average” to a place of 3. Consequently, analyses in terms of means and differences between means are very likely to be misleading and may encourage poor interpretation.

The potential (and often actual) occurrence of misleading results arises mainly from breaches of distribution assumptions underlying parametric tests such as ANOVA in its various forms. Even when ranks (ordinal transformations of numerical data) are derived from normally distributed populations with equal variance, the ranks are neither normally distributed nor of equal variance, and further problems arise in two-way (or more) factorial designs (Shah & Madden, 2004). When the original data are ordinal (i.e., there are no underlying data of interval or ratio scale), it is hardly even meaningful to ask about assumptions that presuppose numerical variables.

**Appropriate Analyses for Ordinal Data**

A growing array of nonparametric statistical tests has emerged during the twentieth century (though some were introduced much earlier). Here, we describe the Kruskal-Wallis test (Kruskal & Wallis, 1952, 1953). This is a nonparametric analogue of a one-way ANOVA and thus can be used to test for differences between two or more groups. The Kruskal-Wallis test does not entirely lack distribution assumptions, because it effectively tests whether the groups of observations plausibly come from the same distribution. However, no particular population distribution is assumed. If the shapes and spreads of the distributions are approximately the same, the Kruskal-Wallis test functions as a test for difference between the group locations. Although the Mann-Whitney test can be used where there are only two groups as a nonparametric analogue of a t-test, we present the Kruskal-Wallis test as having the broadest application because it can deal with designs involving two groups as well as those with more than two groups.

Given numerical data in k groups, the logic of the Kruskal-Wallis test considers the number of cases in each group that fall above or below the common median (i.e., the middle value if the observations in all groups are ranked in order from smallest to largest). If there is no difference between the groups, we expect approximately half of the observations in each group to fall above the common median and approximately half to fall below the common median. The Kruskal-Wallis test statistic, $H$, rises as the data deviate from these expectations. If the deviation is sufficiently great, we conclude that the different groups do not all come from the same population. The Kruskal-Wallis test proceeds as follows.

(a) Calculation of ranks. Ranks are calculated by sorting all the observations as one group from smallest (rank = 1) to largest. For ties (equal observations), the rank is the mean of the ranks that would have applied if the observations had differed slightly.

(b) Test statistic (no ties). If no two observations are equal, the test statistic is

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

where $k$ = the number of groups, $n_i$ = the number of observations in the $i$th group, $n$ = the number of observations in all groups combined, and $R_i$ = the sum of the ranks in the $i$th group.

(c) Test statistic (with ties). If there are tied observations, the test statistic $H$ (1) is divided by

$$1 - \frac{\sum_{i=1}^{n} t_i^3 - t_i}{n^3 - n}$$

where $g$ = the number of groups of ties in the combined data, and $t_i$ = the number of ties in the $i$th group of tied observations.

If the underlying data are true numerical observations (measurement or ratio scale), there are usually few ties and the adjustment for ties makes little difference. However, in ordinal data the number of distinct values is usually limited, and ties are common. Thus, for ordinal data, the adjustment for ties may be quite important.

If all $n_i$ are $\geq 5$, the sampling distribution of Kruskal-Wallis’s (adjusted) $H$ statistic is approximately $\chi^2$ on $k - 1$ degrees of freedom.

A significant result in a Kruskal-Wallis test indicates that not all the treatment groups come from the same population, but it will not indicate which groups are significantly different from others. If $H$ is significant, post hoc testing can be used to determine which groups are significantly different from others in a series of tests between pairs of groups, analogous to similar approaches in ANOVA. Such
multiple pairwise tests increase the likelihood of spuriously finding a significant result (for a detailed exposition, see Field et al., 2012, pp. 428–431). Thus, depending on the software package used, these tests may or may not include corrections for multiple comparisons such as the Bonferroni correction or the sequential Bonferroni correction (often favored because the Bonferroni correction is deemed too severe; Holm, 1979).

**An Example from Plant Pathology**

Consider a fictitious example from plant pathology, a field in which ordinal scales are often used for assessing the severity of disease. The example deals with the imaginary "tomato blight" disease. The tomato blight fungus infects plants via the leaves and spreads through the vascular tissue, weakening the plant and ultimately killing it. Fungicides administered as foliar sprays treat the disease, but the dose is critical – too low a dose has little protective effect, while too high a dose is toxic to the tomato plant. An experiment is designed in which 50 tomato plants are infected with tomato blight and then assigned at random to one of five dose levels of fungicide. The response of the plants is assessed on a five-point rating scale, so the data are ordinal and unsuited to a parametric test (Table 1).

These data can be analyzed using a Kruskal-Wallis test, an option in most commercial statistical analysis software packages as well as the freeware R (https://www.r-project.org). If commercial software is unavailable and R is unsuitable because of the need to learn coding, freeware with a graphical user interface can be used. For example, PAST (Hammer et al., 2001) is available as a free download for Windows and Macintosh from http://folk.uio.no/ohammer/past/, while VassarStats (http://vassarstats.net) offers an interactive online environment. In the following sections we analyze the data in this example using both PAST and VassarStats.

**Analyzing the Plant Pathology Data in PAST**

Data entry in PAST is via a spreadsheet interface, with the Kruskal-Wallis test available via the "several-sample tests" option under the "univariate" menu. Selecting the Kruskal-Wallis tab in the output gives the $H$ statistic with and without correction for tied ranks (but PAST does not display the degrees of freedom, calculated simply as one less than the number of groups; Figure 1). In this example the groups all have identical sample sizes, but this need not be the case and PAST will accept unequal group sizes.

The treatments lead to significant differences in survivorship. PAST also offers post hoc pairwise comparison of treatments using Mann-Whitney tests (Figure 2), with options for adjusting these with a Bonferroni correction or sequential Bonferroni correction. Using the sequential Bonferroni correction, treatments C and D (doses of fungicide at 2 mg/mL and 3 mg/mL) lead to the best health ratings. Treatment B (1 mg/mL) is no better than treatment A (the control), while at treatment E (4 mg/mL) toxicity may be setting in, with the rating significantly less than treatment D although still better than the control. For a more detailed version of these instructions, see Appendix S1 (available as Supplemental Material with the online version of this article).

**Analyzing the Plant Pathology Data in VassarStats**

VassarStats (http://vassarstats.net) offers the Kruskal-Wallis test via the "ordinal data" option on the home page. There are options

| Treatments | A | B | C | D | E |
|------------|---|---|---|---|---|
| 0          | 1 | 2 | 4 | 2 |
| 1          | 3 | 3 | 1 |   |
| 1          | 1 | 3 | 3 | 3 |
| 0          | 1 | 2 | 4 |   |
| 0          | 2 | 4 | 4 | 2 |
| 0          | 1 | 3 | 4 | 2 |
| 0          | 1 | 3 | 4 | 2 |
| 1          | 0 | 2 | 3 | 1 |
| 2          | 3 | 3 | 4 | 0 |
| 0          | 2 | 4 | 3 | 2 |

**Table 1.** Fictitious data showing plant health ratings for tomato seedlings two weeks after infection with the imaginary “tomato blight” disease and administration of one of five treatments by foliar spray: (A) distilled water, (B) 1 mg/mL fungicide, (C) 2 mg/mL fungicide, (D) 3 mg/mL fungicide, or (E) 4 mg/mL fungicide. Ratings: 0 = dead, 1 = severely unhealthy, 2 = moderately unhealthy, 3 = minor symptoms, and 4 = healthy.

**Figure 1.** Output from a Kruskal-Wallis analysis of data in Table 1, using PAST.
for entering data directly online or copying and pasting from Excel, with a maximum of five groups available. In this example the groups all have identical sample sizes, but this need not be the case and VassarStats will accept unequal group sizes. The output includes a display of the ranks calculated from the raw data and the $H$ statistic with degrees of freedom, but not $H$ corrected for ties (Figure 3). There is no provision for post hoc pairwise testing, although this could be done by selecting the option for Mann-Whitney tests and analyzing each paired combination individually. Bonferroni or sequential Bonferroni corrections would need to be applied manually. VassarStats does, though, offer a handy “Mean Ranks for Sample” table in the output that is useful for a quick visual inspection of trends in the data, albeit without pairwise significance testing (Figure 3). This shows an increase in ranks from treatment A to treatment D, with a fall in treatment E. This is consistent with improved protection at increasing doses of fungicide up to a dose of 3 mg/mL, above which toxicity appears to set in. For a more detailed version of these instructions, see Appendix S2 (available as Supplemental Material with the online version of this article).

**Suggestions for Classroom Practice**

The distinction between different scales and the implications for statistical analysis are important points in inquiry-based biology teaching. Although a $t$-test or ANOVA will “work” with ordinal data, such an analysis is incorrect because there is no information on the distance between measurements, only their order. Fortunately, easy-to-use freeware is available for nonparametric analyses of ordinal data to draw robust conclusions. We include a table to support choice of statistical analysis (Table 2) that teachers may find useful in guiding students in how to think about their data and choose a relevant test.

We suggest that if an ordinal scale is implicit in a laboratory exercise – for example, the nightcrawler decomposition exercise described by Bennett (2015) – then an explanation of the special nature of ordinal scales should be part of the lesson. If statistical analysis is planned, then PAST or VassarStats offers quick freeware solutions via the Kruskal-Wallis test without the need to learn coding. VassarStats is simpler to use (there is no software download and no need to learn conventions of data entry or search menus). It also shows ranks and mean ranks for treatment groups in its output, which can help students appreciate the underlying logic of the test and identify trends across groups by inspection of the mean ranks. However, more advanced students will appreciate the extra features in PAST, including the correction for ties and the post hoc pairwise testing (complete with corrections for multiple tests). Of course, if students have access to site-licensed statistical software or already have proficiency in R, these may be preferred solutions.

**Figure 2.** Significance values for post hoc pairwise comparisons of the treatments in Table 1, using the Mann-Whitney pairwise option in PAST.
Table 2. Choosing a statistical analysis. Choice of analysis depends on (at least) two things: the type of data (column headings) and the research question (row headings). It is the intersection of data type and research question that determines the appropriate analysis. Note that this table does not deal with data where there are repeated measures (i.e., more than one measurement taken on each subject). So, where “groups” are mentioned, these are independent groups with only one measurement on each subject. Similarly, there are many more kinds of research question that can arise in practice, and complex combinations of data types can arise. The table is limited to some basic research questions that can be resolved using fairly simple data.

| Data Type | Nominal | Ordinal | Interval | Ratio |
|-----------|---------|---------|----------|-------|
| The research question concerns the . . . | $\chi^2$ goodness-of-fit | $\chi^2$ test of association | Wilcoxon | Kruskal-Wallis |
| . . . distribution of one categorical variable. | . . . relationship between variables of same type. | . . . difference between two groups. | . . . differences between multiple groups. | Analysis of variance |
References

Bennett, S.L. (2015). Exploring the methods of science using nightcrawler decomposition. American Biology Teacher, 77, 600–605.

Field, A., Miles, J. & Field, Z. (2012). Discovering Statistics Using R. London: Sage.

Giese, A.R. (2012). Heads up! A calculation- & jargon-free approach to statistics. American Biology Teacher, 74, 339–340.

Hammer, Ø, Harper, D.A.T. & Ryan, D. (2001). PAST: paleontological statistics software package for education and data analysis. Palaeontologia Electronica, 4(1). http://palaeoelectronica.org/2001_1/past/issue1_01.htm.

Holm, B. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65–70.

Kamin, L.F. (2010). Using a five-step procedure for inferential statistical analyses. American Biology Teacher, 72, 186–188.

Kruskal, W.H. & Wallis, W.A. (1952). The use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47, 583–621.

Kruskal, W.H. & Wallis, W.A. (1953). Errata in: The use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 48, 907–911.

Ruxton, G.D. & Colegrave, N. (2011). Experimental Design for the Life Sciences, 3rd ed. Oxford, UK: Oxford University Press.

Shah, D.A. & Madden, L.V. (2004). Nonparametric analysis of ordinal data in designed factorial experiments. Phytopathology, 94, 33–43.

Zar, J.A. (2010). Biostatistical Analysis, 5th ed. Upper Saddle River, NJ: Prentice Hall.

MICHAEL CALVER (m.calver@murdoch.edu.au) is a Professor of Environmental and Conservation Sciences and DOUGLAS FLETCHER is an Associate Lecturer in Information Technology, Mathematics and Statistics, both at Murdoch University, Murdoch, Western Australia, Australia 6150.

Affiliate Members

Biology Teachers Association of New Jersey (BTANJ)
Colorado Biology Teachers Association (CBTA)
Cleveland Regional Association of Biologists (CRABS)
Connecticut Association of Biology Teachers (CTABT)
Delaware Association of Biology Teachers (DABT)
Empire State Association of Two-Year College Biologists (ESATYCB)
Hong Kong Association of Biology Teachers (HKABT)
Illinois Association of Biology Teachers (IABT)
Illinois Association of Community College Biologists (IACCB)
Indiana Association of Biology Teachers (IABT)

Kansas Association of Biology Teachers (KABT)
Louisiana Association of Biology Teachers (LABT)
Massachusetts Association of Biology Teachers (MABT)
Michigan Association of Biology Teachers (MABT)
Mississippi Association of Biology Educators (MSABE)
Missouri Association of Biology Teachers (MOBioTA)
New York Biology Teachers Association (NYBTA)
South Carolina Association of Biology Teachers (SCABT)
Tennessee Association of Biology Teachers (TNABT)
Texas Association of Biology Teachers (TABT)
Virginia Association of Biology Teachers (VABT)

The National Association of Biology Teachers supports these affiliate organizations in their efforts to further biology & life science education.