

Homework 2

Due: Tuesday March 2 at 11:59pm Champaign (US Central) time

See general homework tips and submit your files via the course website.

For exercises 1, 2, and 3 use the **taevals** data set and for exercise 4 use the **seeds** data set defined in **HW2Data.sas** in the Homework 2 folder on the course website. The **taevals** data is a modification of the **Teaching Assistant Evaluation** data set¹ from the UCI Machine Learning Repository². The **seeds** data is based on the **seeds** data set³ from the UCI Machine Learning Repository⁴.

The raw data are contained in **tae.data** and **seeds_dataset.txt**, which are available from the UCI Machine Learning Repository and in the course website. Variables for the original data sets are described on their respective UCI web pages referenced below. The seeds data set is unchanged from that posted on UCI's website, except with variety indices replaced by names.

The data in the **taevals** data set in **HW2Data.sas** contains the following variables:

- **nativeEnglish**: indicator for whether the TA is a native English speaker (**yes** or **no**)
- **semester**: **Regular** (for Fall or Spring semester) or **Summer**
- **scoregroup**: performance score category (**Low**, **Medium**, or **High**)

Note: To retain the order in which categorical values appear in a data set, you can use **order=data** as an option in the **proc freq** statement. This can allow for testing for ordinal associations, if the data is ordered, like the **taevals** data is.

In the following exercise, limit the number of times you evaluate a procedure. If you can obtain all results for an exercise with one **proc** evaluation, just use one **proc** evaluation.

Exercise 1

A student at UW-Madison is considering taking a Statistics course shortly after these TA evaluations were given. The student wonders if there is any relationship between TA ratings and whether the TA is a native English speaker (e.g. do native or non-native speakers tend to receive higher evaluation scores).

- a) Construct a contingency table for **nativeEnglish** and **scoregroup** and comment on any apparent associations between the TA evaluation score groups and native English speaking. If there appear to be associations, note how native speaking and score groups seem to be associated.
- b) Now test for association. Perform and comment on appropriate tests of association and interpret the results in terms of TA evaluation and native speaking.

¹ <http://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation>

² Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

³ <http://archive.ics.uci.edu/ml/datasets/seeds>

⁴ Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

Exercise 2

The student now wonders if there is any difference in association if the medium score group is ignored and only the highest and lowest evaluated TAs are considered. They look at just the **High** and **Low** score groups to decide.

- a) Construct a contingency table for **nativeEnglish** and **scoregroup** ignoring the **Medium** score group data and comment on any apparent associations between native speaking ability and the highest and lowest score groups. If there appear to be associations, note how speaking and high and low score groups seem to be associated.
- b) Now test for association. Perform and comment on appropriate tests of association and interpret the results in terms of TA English speaking and evaluation score groups.
- c) Finally, test (using risk differences) if native English-speaking TAs have a significantly higher probability than nonnative speaking TAs to be high rated (as opposed to low rated) and state your conclusions.

Exercise 3

The student only plans to take courses during the regular Spring and Fall semesters, so they decide to ignore the summer session data in case there is some difference between regular semesters and Summer.

- a) For the regular semester data, analyze the association between native English speaking and TA evaluation scores, test for statistically significant association, and state your conclusions about associations.
- b) Repeat the analysis in part **a** ignoring the **Medium** score group and state your conclusions for the comparison between the high and low performance evaluation groups.

Exercise 4

For the **seeds** data set, we will consider comparisons of groove length between all three varieties. If the groove lengths differ significantly, measuring the groove length of a seed of unknown variety may provide a good way to guess which variety a seed of unknown type is.

The normality assumption is reasonable for each variety, so there is no need to test normality.

- a) Perform a one-way ANOVA for **groovelength** with **variety** as the categorical predictor; test any assumptions of the model that should be tested (aside from normality, which you do not need to test).
- b) From results for the model in part **a**, comment on the significance of the model, and the amount of variation described by the model. What does this tell us about difference of groove length means for some varieties?
- c) Perform the best test for comparing all pairwise differences of means, and comment on any significantly different groove length means. Interpret what these results tell us about differences of groove lengths between varieties.