QSCI 381 HW 5

Simon-Hans Edasi

2/10/2023

1 Part 1. Chi-squared Goodness-of-Fit Test

A private landowner has contracted with a logging company to harvest trees from their property. A condition of the contract is that following the harvest, the remaining trees on the property must be composed of the following percentages:

12% big leaf maple, 40% Douglas-fir, 8% red alder, 28% western hemlock, 12% western redcedar

Following the harvest, the private landowner claims that the logging company has violated the terms of the contract. The private landowner conducts a survey on the trees remaining after harvest; these data are in the treesoil.csv dataset under the column Tree.

(1a) What is H_o and H_a in a test of the claim? Is this a left-tailed, right-tailed, or two-tailed test? (6 points)

Two tail test due to the inequality of the alternative hypothesis.
$H_o \rightarrow blm = 0.12, df = 0.40, ra = 0.08, wh = 0.28, wrc = 0.12$
$H_a \rightarrow blm \neq 0.12, df \neq 0.40, ra \neq 0.08, wh \neq 0.28, wrc \neq 0.12$

(1b) Determine how many trees of each species were observed in the survey. To do this, create a frequency table of the tree species data using the table() command. Paste in your frequency table below. (6 points)

```
> tree_table = table(tree.data$Tree)
> tree_table

  big leaf maple     Douglas fir        red alder western hemlock
              41             220               21             140
western redcedar
              42
```

(1c) Use your observed frequencies for each tree species from (1b) to test Ho from (1a) using a Chi-squared goodness-of-fit test:

```
chisq.test(dataset, p = c(expected frequencies))
```

Paste your code and output below. (6 points)

```
> chisq.test(tree_table, p = c(.12, .4, .08, .28, .12))

	Chi-squared test for given probabilities

data: tree_table
X-squared = 21.39, df = 4, p-value = 0.000265
```

(1d) Based on your output from (1c), what is your statistical conclusion and interpretation at alpha = 0.05? (10 points)

The p-value is lower than the chosen alpha, so we reject the null hypothesis that the trees occur at the correct rates.

2 Part 2. Chi-squared Test of Independence

In the survey, the private landowner also collected information on the type of soil in which each tree was growing; these data are in the treesoil.csv dataset under the column Soil.

(2a) What is H_o and H_a in a Chi-squared test of independence for the relationship between tree species and soil type? (4 points)

$H_o \to$ Tree and soil are independent of each other
$H_a \to$ Tree and soil are not independent of each other

2b) Use R/RStudio to create a contingency table of tree species by soil type using the following command in R/RStudio:

```
treesoiltab<- table(tree.data)
```

Paste your table below. (6 points)

```
> treesoiltab
                Soil
Tree             clay loam sand silt
  big leaf maple   13    7   12    9
  Douglas fir      51   62   61   46
  red alder         6    7    3    5
  western hemlock  29   49   32   30
  western redcedar 16   10   11    5
```

(2c) Create a mosaic plot in R/RStudio from your contingency table using mosaicplot(). Add an appropriate title, and appropriate labels for the x-axis and y-axis. Use different colors in your plot using col = c("color1", "color2"...); you will need a different color for each soil type from the data in the contingency table. Recall that contingency tables are described as r x c; for example, a 2 x 3 contingency table has 2 rows and 3 columns. Paste your mosaic plot below. (15 points)

(2d) Based on the mosaic plot from (2c), do you think there is an association between tree species and soil type? Justify your answer. (6 points)

(2e) Conduct a Chi-squared test of independence to test your hypotheses in (2a). Paste your code and output below. (6 points)

(2f) Based on your output from (2e), what is your statistical conclusion and interpretation at alpha = 0.05? (10 points)