

Instructions

For each of the following problems, an exercise from the textbook (OpenIntro Statistics) has been worked, with any requisite R (or RStudio) commands provided. Use each as a reference when answering the similar exercise specified. The inputs that will require new values are highlighted in yellow. All input and output will be in the RStudio console. It is recommended that sections 6.3 - 6.4 on pp. 286 - 302 be read, or at least reviewed, prior to completing this lab assignment.

1. *Exercise 6.44 on page 323*

Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined from 2001 to 2002. In this region, woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, and deciduous forests makes up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.

Woods	Cultivated grass plot	Deciduous forests	Other	Total
4	16	61	345	426

(a) *Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.*

$$H_0: p_{\text{woods}} = 4.8\%; p_{\text{grass}} = 14.7\%; p_{\text{forest}} = 39.6\%; p_{\text{other}} = 40.9\%$$

H_A : at least two p_i differ from the hypothesized value

p_i represents the proportion of foraging sites for barking deer that are habitat type i , where i = woods, cultivated grass plot, deciduous forest, or other.

Note: since we are testing a claim about the distribution of a [categorical] variable, the sum of the percentages must be 100%; further we could have said at least one p_i differs in the alternative hypothesis, however, since the sum of all proportions must be 1, by definition, if at least one p_i differs, then at least one other would as well. In words, the hypotheses can be stated as

H_0 : the distribution of habitat usage for foraging by barking deer follows the composition of the region, i.e., there is a no preference.

H_A : the distribution of habitat usage for foraging by barking deer differs from the composition of the region, i.e., there is a preference.

(b) *What type of test can we use to answer this research question?*

We shall use the *Chi-square Goodness of Fit* test, since we are testing a claim about the distribution of a categorical variable from a single population.

(c) *Check if the assumptions and conditions required for this test are satisfied.*

Referring to page 293, we find that “there are two conditions that must be checked to perform a chi-square test:

Independence. Each case that contributes a count to the table must be independent of all other cases in the table.

Sample size / distribution. Each particular scenario (i.e., cell count) must have at least 5 expected cases.”

In this study, we shall assume that each of the 426 foraging sites (cases) of barking deer in Hainan Island, China that were examined (observed) from 2001 to 2002 are independent. Further, if we apply the hypothesized distribution defined in H_0 to the stated sample size, that is evaluate np_i for all i , we would expect to observe the following counts:

Woods	Cultivated grass plot	Deciduous forests	Other	Total
20.448	62.622	168.696	174.234	426

Here the smallest expected count (20.448) is more than four times the minimum threshold of 5.

(d) *Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.*

As stated on page 293, the test statistic for this particular test has the form

$$\begin{aligned}\chi^2 &= \sum_{i=1}^4 \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(4 - 20.448)^2}{20.448} + \frac{(16 - 62.622)^2}{62.622} + \frac{(61 - 168.696)^2}{168.696} + \frac{(345 - 174.234)^2}{174.234} \approx 284.0609\end{aligned}$$

Under the assumption the null hypothesis is true and the conditions in (c) are met, the test statistic follows a chi-square distribution with $k - 1$ degrees of freedom, where k denotes the number of categories (or number of levels) of the categorical variable. For this study, k is 4 and the degrees of freedom is 3. The resulting p-value is then

$$P(\chi^2 > 287.4741) \approx 2.8 \times 10^{-61}$$

Again, the calculations can be done altogether by typing (or copying and pasting) the following R commands into the console in RStudio, where `observed` is a vector containing the four observed counts, `p0` is a vector contain the corresponding hypothesized proportions (it is important that the order of the categories is preserved between both), and `n` is the sample size:

```

# inputs
observed <- c(4, 16, 61, 345)
p0 <- c(0.048, 0.147, 0.396, 0.409)
n <- 426

# calculations
expected <- n*p0
x2 <- sum((observed-expected)^2/expected)
df <- length(observed) - 1
p.value <- pchisq(x2, df, lower.tail=F)

# print results
expected
x2
p.value

```

which yields the resulting [edited] console output:

```

> expected
[1] 20.448 62.622 168.696 174.234

> x2
[1] 284.0609

> p.value
[1] 2.799724e-61

```

Again, we could have also obtained the value of the test statistic and corresponding p-value by using the more convenient `chisq.test` command. Here we only need to provide the vectors for the observed counts and hypothesized proportions; if the latter is omitted, a uniform distribution is assumed, i.e., all proportions are the same. Also, to match the results of the “by hand” calculation performed earlier, we will need to “turn off” the continuity correction in the `chisq.test` procedure by including the `correct=F` input. Copy and paste the command below (into the Rstudio console) to obtain the corresponding [formatted] output.

COMMAND

```
chisq.test(x=c(4,16,61,345), p=c(.048,.147,.396,.409), correct=F)
```

CONSOLE OUTPUT

```

      Chi-squared test for given probabilities
data:  c(4, 16, 61, 345)
X-squared = 284.06, df = 3, p-value < 2.2e-16

```

Notice that the test statistic and p-values are identical.

Decision:

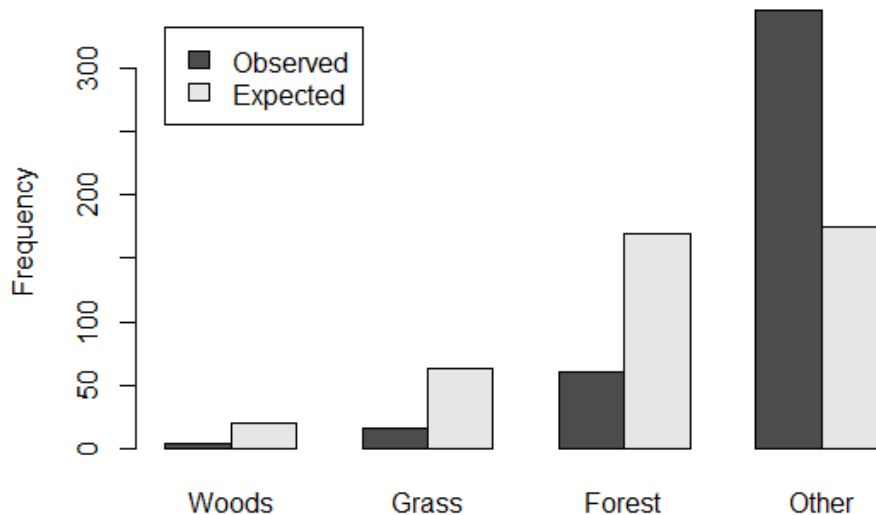
Since the p-value ($2.799724e-61$) $\ll \alpha$ (0.05), we reject H_0 .

Conclusion:

At the 5% significance level, there is sufficient (statistically significant) evidence to conclude the distribution of habitat usage for foraging by barking deer differs from the composition of the region. In other words, barking deer prefer to forage in certain habitats over others; foraging location does not appear to be random.

Finally, constructing a side-by-side bar plot comparing the observed and expected counts for each category is a useful tool for identifying if any [large] deviations between the two exist, or if they are likely due to random chance. The R commands and resulting plot are provided below - due to spacing, the plot is shown first; we reference the observed and expected objects defined earlier. Notice the largest deviation(s) and thus greatest contribution(s) to the chi-square test statistic occur in the *Other* (much higher than expected) and then *Deciduous forests* (much lower than expected).

PLOT



COMMAND

```
barplot(
  height      = rbind(observed,expected),
  names.arg   = c("woods","Grass","Forest","Other"),
  beside      = T,
  legend.text = T,
  args.legend = list(x=4, legend=c("Observed","Expected")),
  ylab        = "Frequency"
)
```

Note: the above R code is a single command, `barplot`, that spans multiple lines.

On Your Own

Exercise 6.50 (a) on page 326.

2. Exercise 6.47 on page 324

The table below summarizes a data set we first encountered in Exercise 6.29 that examines the responses of a random sample of college graduates and non-graduates on the topic of oil drilling. Complete a chi-square test for these data to check whether there is a statistically significant difference in responses from college graduates and non-graduates.

	<i>College Grad</i>	
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131
Total	438	389

Here we will be performing chi-square test for independence (in a two-way table) to determine if the observed data provides convincing (statistically significant) evidence that the distribution of the levels of one categorical variable (RESPONSE: support, oppose, don't know) differs between two or more levels of the other (COLLEGE GRAD: yes, no). For a more detailed discussion of the topic, we refer the reader to section 6.4 on pages 297 - 302.

- *State hypotheses.*

H_0 : There is no association between the opinion of registered voters in California on drilling for oil and natural gas off the Coast of California and whether or not they graduated college; i.e., *response* is independent of *college grad*.

H_A : There is an association between the opinion of registered voters in California on drilling for oil and natural gas off the Coast of California and whether or not they graduated college; i.e., *response* is dependent on *college grad*.

DATA ENTRY AND CALCULATIONS

Before continuing with the remaining steps in the analysis, we will first look at how to enter a two-way table in RStudio as a `data.frame` object, and then (re)use the `chisq.test` command to perform the computations. Rather than simply using the default output printed to the console, we will instead store the results in another object and then access the desired output from the (`htest`) object by referencing the specific value. Please compare the cell values and row/column labels in the two-way table above to the inputs in the `data.frame` command to understand the input format.

COMMANDS

```
# define two-way table
obs <- data.frame(
  Yes = c(154, 180, 104),
  No = c(132, 126, 131),
  row.names = c("Support", "Oppose", "Do not know")
)
```

```
# perform chi-square test
htest <- chisq.test(obs, correct=F)

# print selected results
htest$observed
htest$expected
htest$statistic
htest$p.value

# print default output
htest
```

CONSOLE OUTPUT

```
> htest$observed
      Yes  No
Support  154 132
Oppose   180 126
Do not know 104 131

> htest$expected
      Yes      No
Support 151.4728 134.5272
Oppose  162.0653 143.9347
Do not know 124.4619 110.5381

> htest$statistic
X-squared
11.46082

> htest$p.value
[1] 0.003245752

> htest

Pearson's Chi-squared test

data:  obs
X-squared = 11.461, df = 2, p-value = 0.003246
```

- *Check relevant conditions.*

Independence. Each case that contributes a count to the table must be independent of all other cases in the table.

Sample size / distribution. Each particular scenario (i.e., cell count) must have at least 5 expected cases.”

In this study, we can assume each case (registered California voter) is independent as the 827 in the study were randomly selected. Further, if we apply the hypothesized distribution defined in H_0 to the stated sample size, that is perform the computation for expected counts in a two-way table (see page 299), which is an application of the multiplication rule for independent events in probability theory - or more conveniently, just reference the R output above, we would expect to observe the following counts:

	<i>College Grad</i>	
	Yes	No
Support	151.4728	134.5272
Oppose	162.0653	143.9347
Do not know	124.4619	110.5381
Total	438	389

Here all expected counts are well above the minimum threshold of 5; the smallest is approximately 111.

- *Interpret results in context of the data.*

Decision:

Since the p-value (0.003246) $< \alpha$ (0.05), we reject H_0 .

Conclusion:

At the 5% significance level, we have sufficient (statistically significant) evidence to conclude there is an association between the opinion of registered voters in California on drilling for oil and natural gas off the Coast of California and whether or not they graduated college; in other words, the responses from college graduates and non-graduates differ.

On Your Own

Exercise 6.50 (b) on page 326.