# Chapter 12

# Modeling categorical relationships

So far we have discussed the general concept of statistical modeling and null hypothesis significance testing, and we have implemented statistical modeling and null hypothesis significance testing in some analyses.

In this chapter we will focus on the modeling of *categorical* relationships, by which we mean relationships between variables that are measured discretely. Discrete data are usually expressed in terms of counts; that is, for each value of the variable (or combination of values of multiple variables), how many observations take that value?

## 12.1 Chi-square test for goodness of fit

Let's say that I have purchased a bag of 99 candies, which are labeled as having 1/3 chocolates, 1/3 licorices, and 1/3 gum balls. Because the bag promises 99 candies of three types and 1/3 of each type, I expect there to be 33 chocolates, 33 licorices, and 33 gum balls.

However, when I count the candies in the bag, I get the following numbers: 30 chocolates, 33 licorices, and 37 gum balls. Because I like chocolate much more than licorice or gum balls, I feel slightly ripped off, and I'd like to know if this was just a random accident.

To answer that question, I need to know: Does the frequency that I observed differ significantly from the frequency I expected?

Table 12.1: Observed frequencies, Expected frequencies under the null hypothesis, and squared differences in the candy data

| Candy Type | Observed frequency | null Expected | difference | squared difference |
|---|---|---|---|---|
| chocolate | 29 | 33 | -4 | 16.000 |
| licorice | 33 | 33 | 0 | 0.000 |
| gum ball | 37 | 33 | 4 | 16.000 |

## 12.2 Calculating Pearson's chi-square test for goodness of fit

The Pearson chi-square test for goodness of fit provides us with a way to test whether the frequencies of discrete data that we observe differ from the frequencies of discrete data that we expect under the null hypothesis:

$$\chi^2 = \sum_i \frac{(observed_i - expected_i)^2}{expected_i}$$

To compute the chi-square statistic, we first need to come up with our expected frequencies under the null hypothesis.

In the case of our candy example, the null hypothesis is that the number of each of the three types of candies should all the same: 33 chocolates, 33 licorices, and 333 gum balls -- meaning an even split across the three categories (as shown in Table 12.1).

Then,

(1) we calculate the difference between each Observed frequency and its (null) Expected frequency,

(2) we square each of those differences,

(3) we divide each of the squared differences by the (null) Expected frequency, and

(4) we add up them up to obtain the chi-square statistic.

The chi-square statistic we calculate for our candy test comes out to 0.486.

When we consult a chi-square probability table, we find that this chi-square value suggests a *p*-value greater than .050; therefore, we cannot reject the null hypothesis that there was an even frequency of each of the three types of candy in the bag.