CHAPTER 12

Categorical Data

CONTENTS

12.1 Introduction 634
12.2 Hypothesis Tests for a Multinomial Population 634
12.3 Goodness of Fit Using the χ^2 Test 637
12.4 Contingency Tables 641
12.5 Loglinear Model 649
12.6 Chapter Summary 655
12.7 Chapter Exercises 655

■ Example 12.1: Developmental Research

A study by Aylward *et al.* (1984), reported in Green (1988), examines the relationship between neurological status and gestational age. The researchers were interested in determining whether knowing an infant's gestational age can provide additional information regarding the infant's neurological status. For this study, 505 newborn infants were cross-classified on two variables: overall neurological status, as measured by the Prechtl examination, and gestational age. The data are shown in Table 12.1.

Note that the response variable, Prechtl status, is a categorical variable; hence a linear model of the type we have been using is not appropriate. Additionally, in this example, the independent variable, the age of the infant, is recorded by intervals and can therefore also be considered a categorical variable. We will return to this example in Section 12.5.

Statistical Methods, DOI: 10.1016/B978-0-12-374970-3.00012-3

Table 12.1 Number of Infants

	GESTATIONAL AGE (IN WEEKS)				
Prechtl Status	**31 or Less**	**32–33**	**34–36**	**37 or More**	**All Infants**
Normal	46	111	169	103	409
Dubious	11	15	19	11	56
Abnormal	8	5	4	3	20
All infants	65	131	192	117	505

12.1 INTRODUCTION

Up to this point we have been primarily concerned with analyses in which the response variable is ratio or interval and usually continuous in nature. The only exceptions occurred in Sections 4.3 and 5.5, where we presented methods for inferences on the binomial parameter p for an outcome variable that is binary (has only two possible values).

Nominal variables are certainly not restricted to having only two categories. Variables such as flower petal color, geographic region, and plant or animal species, for example, are described by many categories. When we deal with variables of this nature we are usually interested in the frequencies or counts of the number of observations occurring in each of the categories; hence, these types of data are often referred to as categorical data.

This chapter covers the following topics:

- Hypothesis tests for a multinomial population.
- The use of the χ^2 distribution as a goodness-of-fit test.
- The analysis of contingency tables.
- An introduction to the loglinear model to analyze categorical data.

12.2 HYPOTHESIS TESTS FOR A MULTINOMIAL POPULATION

When the response variable has only two categories, we have used the binomial distribution to describe the sampling distribution of the number of "successes" in n trials. If the number of trials is sufficiently large, the normal approximation to the binomial is used to make inferences about the single parameter p, the proportion of successes in the population.

When we have more than two categories, the underlying distribution is called the **multinomial distribution**. For a multinomial population with k categories, the distribution has k parameters, p_i, which are the probabilities of an observation occurring in category i. Since an observation must fall in one category, $\sum p_i = 1$. The

actual function that describes the multinomial distribution is of little practical use for making inferences. Instead we will use large sample approximations, which use the χ^2 distribution presented in Section 2.6.

When making inferences about a multinomial population, we are usually interested in determining whether the probabilities p_i have some prespecified values or behave according to some specified pattern. The hypotheses of interest are

$$H_0: p_i = p_{i0} \quad i = 1, 2, \ldots, k,$$
$$H_1: p_i \neq p_{i0} \quad \text{for at least two } i,$$

where p_{i0} are the specified values for the parameters.

The values of the p_{i0} may arise either from experience or from theoretical considerations. For example, a teacher may suspect that the performance of a particular class is below normal. Past experience suggests that the percentages of letter grades A, B, C, D, and F are 10, 20, 40, 20, and 10%, respectively. The hypothesis test is used to determine whether the grade distribution for the class in question comes from a population with that set of proportions. In genetics, the "classic phenotypic ratio" states that inherited characteristics, say, A, B, C, or D, should occur with a 9:3:3:1 ratio if there are no crossovers. In other words, on the average, 9/16 of the offspring should have characteristic A, 3/16 should have B, 3/16 should have C, and 1/16 should have D. Based on sample data on actual frequencies, we use this hypothesis test to determine whether crossovers have occurred.

The test statistic used to test whether the parameters of a multinomial distribution match a set of specified probabilities is based on a comparison between the actually observed frequencies and those that would be expected if the null hypothesis were true. Assume we have n observations classified according to k categories with observed frequencies $n_1, n_2, \ldots, n_k$. The null hypothesis is

$$H_0: p_i = p_{i0}, \quad i = 1, 2, \ldots, k.$$

The alternate hypothesis is that at least two of the probabilities are different. The expected frequencies, denoted by E_i, are computed by

$$E_i = n\, p_{i0}, \quad i = 1, 2, \ldots, k.$$

Then the quantities $(n_i - E_i)$ represent the magnitudes of the differences and are indicators of the disagreement between the observed values and the expected values if the null hypothesis were true. The formula for the test statistic is

$$X^2 = \sum \frac{(n_i - E_i)^2}{E_i},$$

where the summation is over all k categories. We see that the squares of these differences are used to eliminate the sign of the differences, and the squares are