11-1 Review and Preview

By introducing basic concepts of estimating population parameters (with confidence intervals) and methods of hypothesis tests, Chapters 7 and 8 moved us into methods of inferential statistics. Chapters 9 and 10 then involved us with different configurations of data. In this chapter we use statistical methods for analyzing categorical (or qualitative, or attribute) data that can be separated into different cells.

In Section 11-2 we consider hypothesis tests of a claim that observed frequency counts agree with some claimed distribution, so there is a "good fit" of the sample data with the claimed distribution. In Section 11-3 we analyze contingency tables (or two-way frequency tables), which consist of frequency counts arranged in a table with at least two rows and two columns. The objective is to determine whether there appears to be some dependence between the row variable and the column variable.

The methods of this chapter use the same χ^2 (chi-square) distribution that was first introduced in Section 7-4. See Section 7-4 for a quick review of properties of the χ^2 distribution.

11-2 Goodness-of-Fit

Key Concept By "goodness-of-fit" we mean that sample data consisting of observed frequency counts arranged in a single row or column (called a *one-way frequency table*) agree with some particular distribution being considered. We will use a hypothesis test for the claim that the observed frequency counts agree with some claimed distribution.

DEFINITION A **goodness-of-fit test** is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.

Objective

Conduct a goodness-of-fit test. That is, conduct a hypothesis test to determine whether a single row (or column) of frequency counts agrees with some specific distribution (such as uniform or normal).

Notation

O represents the *observed frequency* of an outcome, found from the sample data.

E represents the *expected frequency* of an outcome, found by assuming that the distribution is as claimed.

k represents the *number of different categories* or cells.

n represents the total *number of trials* (or observed sample values).

Requirements

1. The data have been randomly selected.
2. The sample data consist of frequency counts for each of the different categories.
3. For each category, the *expected* frequency is at least 5. (The expected frequency for a category is the frequency that would occur if the data actually have the distribution that is being claimed. There is no requirement that the *observed* frequency for each category must be at least 5.)