11-2 Goodness of Fit Test

In This section we consider sample data consisting of observed frequency counts arranged in a single row or column (called a one-way frequency table). We will use a hypothesis test for the claim that the observed frequency counts agree with some claimed distribution, so that there is a *good fit* of the observed data with the claimed distribution.

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

H_0: The random variable follows a particular distribution.

H_1: The random variable does not follow the distribution specified in H_0.

Ex 1) Consider the observed frequencies and relative frequencies of browser preference form a survey of 200 Internet users.

Browser	Observed frequency	Relative frequency
Microsoft Internet Explorer	140	0.785
Firefox	40	0.15
Safari/other	20	0.065

The following model shows how the market shares are distributed in the null hypothesis:

H_0: $P_{Ms\ IE}$= 0.785 , $P_{firefox}$=0.15, $P_{Safari/other}$= 0.065

$H_{1:}$ The random variable does not follow the distribution specified in H_0.

How a Goodness- of -Fit Test Works

The goodness -of -fit test is based on a comparison of the observed frequencies (actual data from the field) with the expected frequencies when H_0 is true. That is, we compare what we actually see with what would expect to see if H_0 were true. If the difference between the observed and expected frequencies is large, we reject H_0.

As usual, it comes down to how large a difference is large. The hypothesis we conduct to answer this question relies on χ^2 distribution.

Performing the χ^2 Goodness of Fit Test

The following conditions must be met:

- The data have been randomly selected.
- The sample data consist of frequency counts for each of the different categories.
- None of the expected frequencies is less than 1.
- For each category, the expected frequency is at least 5.

Finding Expected Frequencies

The expected frequency for a category is the frequency that would occur if the data actually have the distribution that is being claimed. For the i^{th} category, the expected frequency is $E_i = n \cdot p_i$, where n represent the number of trials and p_i represents the population proportion for the i^{th} category.

If we assume that all expected frequencies are equal, then each expected frequency is $E = n/k$, where n is the total number of observations and k is the number of categories.

The χ^2 goodness of fit test may be performed using (a) the critical value, and

(b) the p-value method.

(a) χ^2 goodness of fit test. (Critical value method)

Step 1: State the hypotheses and check the conditions.

The null hypothesis is states that the qualitative random variable follows a particular distribution. The alternative hypothesis states that the random variable does not follow that distribution.

Step 2: Find the χ^2 critical value, χ^2_{critical}, from table A-4 by using k -1 degrees of freedom, where k is the number of categories.

Note, Goodness-of-fit hypothesis are always right tailed.

And state the rejection rule.

Reject if $\chi^2_{\text{data}} > \chi^2_{\text{critical}}$.

Step 3: Find the test statistic χ^2_{data}.

$$\chi^2_{\text{data}} = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where O_i = observed frequency, and E_i = expected frequency.

Step 4: State the conclusion and the interpretation.

Ex 2) Perform the hypothesis test shown in example 1, use 0.05 as a significance level.

H_0: $P_{Ms\ IE}$= 0.785 , $P_{firefox}$=0.15, $P_{Safari/other}$= 0.065

$H_{1:}$ The random variable does not follow the distribution specified in H_0.

Browser Category	Observed freq O_i	Relative freq p_i
MS IE	140	0.785
Firefox	40	0.15
Safari/other	20	0.065

Interpretation: There is evidence that the random variable browser does not follow the distribution in H_0. In other words, there is evidence that the market shares for internet browsers have changed.

Note carefully what this conclusion says and what it doesn't say. The χ^2goodness of fit test shows that there is evidence that the random variable does not follow the distribution specified in H_0. In particular, the conclusion does not state, for example, that Firefox's proportion is significantly greater.

(b) **χ^2 goodness of fit test: (p-value Method)**

Step 1: State the hypotheses and check the conditions.

Step 2: Find the test statistic χ^2_{data}.

$$\chi^2_{data}=\sum\frac{(O_i-E_i)^2}{E_i}$$

Where O_i = observed frequency, and E_i = expected frequency

Step 3: Find the p-value.

p-value = P ($\chi^2 > \chi^2_{data}$)

Step 4: State the conclusion and the interpretation.

Ex (3) The following tables show figures on the market share of cable modem, DSL, and wireless broadband from a 2002 survey and a 2006 survey which was based on a random sample of 1000 home broadband users. Test whether the population proportions have changed since 2002, using the p-value method, and level of significance is 0.05.

2002 broadband adoption survey

Cable modem	DSL	Wireless/other
67%	28%	5%

2006 broadband adoption survey

Cable modem	DSL	Wireless/other
410	500	90

`