

STAT 100 Lecture 25:
Analysis of Categorical Data
Part 1:
Pearson's χ^2 Test for Goodness of Fit

Nate Strawn

December 3

- 1 Matched Pairs Design
- 2 Examples

Today's Agenda

- 1 Introduction
 - Categorical data. Examples.
- 2 Pearson's χ^2 test for goodness of fit
- 3 Examples

Definition

Categorical data refers to observations that are only classified into categories so that the data set consists of frequency counts for the categories.

Example

- 1 Blood type (O,A,B,AB)
- 2 A shipment of assorted nuts (walnuts, hazelnuts, almonds pistachios)
- 3 Number of births by day of the week (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday)

Example 1. One Sample Classified in Several Categories

Example

The offspring produced by a cross between two given types of plants can be any of the three genotypes denoted by A, B, and C. A theoretical model of gene inheritance suggests that the offspring of types A, B, and C should be in the ratio 1 : 2 : 1. For experimental verification, 100 plants are bred by crossing the two given types. Their genetic classifications are recorded in the table below.

<i>Genotype</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>Total</i>
<i>Observed Frequency</i>	18	55	27	100

Example 2. Independent Samples Classified in Several Categories

Example

A survey is undertaken to determine the incidence of alcoholism in different professional groups. Random samples of the clergy, educators, executives, and merchants are interviewed, and the observed frequency counts are given in the table below.

Contingency Table of Alcoholism versus Profession

	<i>Alcoholic</i>	<i>Nonalcoholic</i>	<i>Sample Size</i>
<i>Clergy</i>	32	268	300
<i>Educators</i>	51	199	250
<i>Executives</i>	67	233	300
<i>Merchants</i>	83	267	350
<i>Total</i>	233	967	1200

Example 2. Independent Samples Classified in Several Categories

Example

Relative Frequencies

	<i>Alcoholic</i>	<i>Nonalcoholic</i>	<i>Sample Size</i>
<i>Clergy</i>	.11	.89	1
<i>Educators</i>	.20	.80	1
<i>Executives</i>	.22	.78	1
<i>Merchants</i>	.24	.76	1

Example 3. One Sample Simultaneously Classified According to Two Characteristics

Example

A random sample of 500 persons is questioned regarding political affiliation and attitude toward a tax reform program. From the observed frequency table given in the table below

Political Affiliation and Opinion

	<i>Favor</i>	<i>Indifferent</i>	<i>Opposed</i>	<i>Total</i>
<i>Democrat</i>	138	83	64	285
<i>Republican</i>	64	67	84	215
<i>Total</i>	202	150	148	500

Unlike Example 2, here we have a single random sample, but each sampled individual elicits two types of responses: political affiliation and attitude.

Contingency Table

Definition

*Frequency count data that arise from a classification of the sample observations according to two or more characteristics are called **cross-tabulated data** or a **contingency table**.*

Pearson's χ^2 test for goodness of fit

Back to Example 1.

Genotype	A	B	C	Total
<i>Observed Frequency</i>	18	55	27	100

Let us denote the population proportions or the probabilities of the genotypes A , B , and C by p_A , p_B , and p_C , respectively. Since the genetic model states that these probabilities are in the ratio $1 : 2 : 1$, our object is to test the null hypothesis

$$H_0 : p_A = 1/4 \quad p_B = 2/4 \quad p_C = 1/4$$

The data consist of frequency counts observed from a random sample and the null hypothesis specifies the unknown cell probabilities. Our primary goal is to test if the model given by the *null hypothesis* fits the data, and this is appropriately called **testing for goodness of fit**.

Definition

- 1 *Null hypothesis*

$$H_0 : p_1 = p_{10}, \dots, p_k = p_{k0}$$
$$p_{10} + p_{20} + \dots + p_{k0} = 1$$

- 2 *Test statistic*

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{\text{cells}} \frac{(O - E)^2}{E}$$

- 3 *Rejection region*

$$\chi^2 \geq \chi_{\alpha}^2$$

where χ_{α}^2 is the upper α point of the χ^2 distribution with $d.f = k - 1 = (\text{Number of cells} - 1)$

χ^2 Goodness of Fit for Genetic Model

Back to Example 1.

Test the goodness of fit of the genetic model. Take $\alpha = .05$. The computations for the χ^2 statistic are exhibited in the table below.

Cell	A	B	C	Total
<i>Observed Frequency O</i>	18	55	27	100
<i>Probability under H_0</i>	.25	.50	.25	1.0
<i>Expected frequency E</i>	25	50	25	100
$\frac{(O-E)^2}{E}$	1.96	.50	.16	$2.62 = \chi^2$ <i>d.f.</i> = 2

The rejection region $R : \chi^2 \geq 5.99$ since $\chi_{0.05}^2 = 5.99$ with *d.f.* = 2 (Appendix B, Table 5). Because the observed $\chi^2 = 2.62$ is smaller than this value, the null hypothesis is **not rejected** at $\alpha = .05$. We conclude that the data in the Example 1 do not contradict the genetic model.

Fact

- 1 The χ^2 statistic measures the overall discrepancy between the observed frequencies and those expected under a given null hypothesis.
- 2 **Additivity:** If χ^2 statistics are computed from independent samples, then their **sum** is also a χ^2 **statistic** whose d.f. equals the sum of the d.f.'s of the components.
- 3 **Loss of d.f. due to estimation of parameters:** If H_0 does not completely specify the cell probabilities, then some parameters have to be estimated in order to obtain the expected cell frequencies. In that case, the d.f. of χ^2 is reduced by the number of parameters estimated.
$$\text{d.f. of } \chi^2 = (\text{No. of cells}) - 1 - (\text{No. of parameters estimated})$$

For Next Time

- Read Section 13.3 from Johnson and Bhattacharyya