STAT 515 -- Chapter 13: Categorical Data

Recall we have studied binomial data, in which each trial falls into one of 2 categories (success/failure).

Many studies allow for more than 2 categories.

Example 1: Voters are asked which of 6 candidates they prefer.

Example 2: Residents are surveyed about which part of Columbia they live in. (Downtown, NW, NE, SW, SE)

Multinomial Experiment

(Extension of a binomial experiment → from 2 to k possible outcomes)

(1) Consists of n identical trials
(2) There are k possible outcomes (categories) for each trial
(3) The probabilities for the k outcomes, denoted p_1, p_2, …, p_k, are the same for each trial
(and $p_1 + p_2 + \ldots + p_k = 1$)
(4) The trials are independent

The cell counts, n_1, n_2, …, n_k, which are the number of observations falling in each category, are the random variables which follow a multinomial distribution.

Analyzing a One-Way Table

Suppose we have a single categorical variable with *k* categories. The cell counts from a multinomial experiment can be arranged in a one-way table.

Example 1: Adults were surveyed about their favorite sport. There were 6 categories.

p_1 = proportion of U.S. adults favoring football
p_2 = proportion of U.S. adults favoring baseball
p_3 = proportion of U.S. adults favoring basketball
p_4 = proportion of U.S. adults favoring auto racing
p_5 = proportion of U.S. adults favoring golf
p_6 = proportion of U.S. adults favoring "other"

It was hypothesized that the true proportions are $(p_1, p_2, p_3, p_4, p_5, p_6) = (.4, .1, .2, .06, .06, .18)$.

95 adults were randomly sampled; their preferences are summarized in the one-way table:

Favorite Sport

Football	Baseball	Basketball	Auto	Golf	Other	n
37	12	17	8	5	16	95

We test our null hypothesis (at $\alpha = .05$) with the following test:

Test for Multinomial Probabilities

H_0: $p_1 = p_{1,0}, p_2 = p_{2,0}, \ldots, p_k = p_{k,0}$
H_a: at least one of the hypothesized probabilities is wrong

The test statistic is:

where n_i is the observed "cell count" for category i and $E(n_i)$ is the expected cell count for category i <u>if H_0 is true</u>.

<u>Rejection region</u>: $\chi^2 > \chi^2_\alpha$ where χ^2_α based on $k - 1$ d.f. (large values of χ^2 => observed n_i very different from expected $E(n_i)$ under H_0)

<u>Assumptions</u>: (1) The data are from a multinomial experiment. (2) Every expected cell count $E(n_i)$ is at least 5. (large-sample test)

<u>Finding expected cell counts</u>: Note that $E(n_i) = np_{i,0}$.

For our data,

i	$\boldsymbol{n_i}$	$\mathbf{E}(\boldsymbol{n_i})$

Test statistic value:

From Table VII:

Analyzing a Two-Way Table

Now we consider observations that are classified according to <u>two</u> categorical variables.

Such data can be presented in a <u>two-way</u> table (contingency table).

Example: Suppose the people in the "favorite-sport" survey had been further classified by gender:

Two categorical variables: Gender and Favorite Sport.

<u>Question</u>: Are the two classifications independent or dependent?

For instance, does people's favorite sport depend on their gender? Or does gender have no association with favorite sport?

Observed Counts for a $r \times c$ Contingency Table
(r = # of rows, c = # of columns)

Column Variable

		1	2	…	c	Row Totals
Row Variable	1	n_{11}	n_{12}	…	n_{1c}	R_1
	2	n_{21}	n_{22}	…	n_{2c}	R_2
	⋮	⋮	⋮		⋮	⋮
	r	n_{r1}	n_{r2}	…	n_{rc}	R_r
Col. Totals		C_1	C_2		C_c	n

Probabilities for a $r \times c$ Contingency Table:

Column Variable

		1	2	…	c	
Row Variable	1	p_{11}	p_{12}	…	p_{1c}	$p_{row\ 1}$
	2	p_{21}	p_{22}	…	p_{2c}	$p_{row\ 2}$
	⋮	⋮	⋮		⋮	⋮
	r	p_{r1}	p_{r2}	…	p_{rc}	$p_{row\ r}$
		$p_{col\ 1}$	$p_{col\ 2}$		$p_{col\ c}$	1

Note: If the two classifications are <u>independent</u>, then:
$p_{11} = (p_{row\ 1})(p_{col\ 1})$ **and** $p_{12} = (p_{row\ 1})(p_{col\ 2})$**, etc.**

So under the hypothesis of independence, we expect the cell probabilities to be the product of the corresponding <u>marginal probabilities</u>.

Hence the (estimated) expected count in cell (*i, j*) is simply:

χ^2 test for independence

H_0: The classifications are independent
H_a: The classifications are dependent

Test statistic:

where the expected count in cell (*i, j*) is $\hat{E}(n_{ij}) = \dfrac{R_i C_j}{n}$

Rejection region: $\chi^2 > \chi^2_\alpha$,
where χ^2_α is based on $(r - 1)(c - 1)$ d.f.
and r = # of rows, c = # of columns.

Note: We need the sample size to be large enough that every <u>expected</u> cell count is at least 5.

Example: Does the incidence of heart disease depend on snoring pattern? (Test using $\alpha = .05$.) Random sample of 2484 adults taken; results given in a <u>contingency table</u>:

		Snoring Pattern			
		Never	Occasionally	≈Every Night	
Heart	Yes	24	35	51	110
Disease	No	1355	603	416	2374
		1379	638	467	2484

Expected Cell Counts:

Test statistic: