Chapter 11—Multinomial Experiments and Contingency Tables

GOODNESS-OF-FIT TESTS

We will ask: "how are my data distributed?" Towards this aim, we consider goodness-of-fit tests. Goodness-of-fit tests compare the observed frequency distribution of a sample with a hypothesized frequency distribution. We wish to discern whether or not, apart from sampling error, the observed sample conforms to the hypothesized distribution.

Goodness-of-fit tests are based on count data, i.e., the number of individuals falling in a given category. The attribute observed on each unit sampled determines which category that individual is counted in.

Pearson's χ^2 Goodness-of-Fit Test

Example -- Yule (1923)

We wish to determine whether crossing a yellow-round (YR) pea with a green-wrinkled (gw) pea follows classical Mendelian genetics. Yellow is dominant to green; round is dominant to wrinkled. Theory predicts a relative frequency of

9 (*YR*) : 3 (*Yw*) : 3 (*gR*) : 1 (*gw*)

in the second generation cross F_2.

THEORY

Parents *YYRR* x *ggww*

↓

First Generation (F_1) *YgRw*

↓

Second Gen. (F_2) 9 *YR* : 3 *Yw* : 3 *g*R : 1 *gw* phenotypes

Gametes	***YR***	***Yw***	***gR***	***gw***
YR	*YYRR*	*YYRw*	*YgRR*	*YgRw*
Yw	*YYRw*	*Yyww*	*YgRw*	*Ygww*
gR	*YgRR*	*YgRw*	*ggRR*	*ggRw*
gw	*YgRw*	*Ygww*	*ggRw*	*ggww*

Statistical Hypotheses

H_0: The distribution of phenotypes in the F_2 generation follows a 9:3:3:1 distribution, or specify as multinomial probabilities

$p_{YR} = 9/16$ and $p_{Yw} = 3/16$ and $p_{gR} = 3/16$ and $p_{gw} = 3/16$

H_1: The distribution of phenotypes in the F_2 generation does not follow a 9:3:3:1 distribution.

$p_{YR} \neq 9/16$ or $p_{Yw} \neq 3/16$ or $p_{gR} \neq 3/16$ or $p_{gw} \neq 3/16$

Multinomial Distribution

$P(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } \ldots \text{ and } X_k = x_k) =$

$$\begin{cases} \dfrac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} & \text{when } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise.} \end{cases}$$

Each of the *k* components separately has a binomial distribution with parameters *n* and p_i.

Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{obs}_i - \text{exp}_i)^2}{\text{exp}_i} \sim \chi^2_{df=k-1}$$

where obs_i = observed count in the i^{th} category

exp_i = expected count in the i^{th} category under H_0

k = number of categories

and χ^2 is distributed as a *chi-square random variable with k - 1 degrees of freedom* (*df*).

Degrees of freedom = number of categories (*k*) minus the number of sample constants used to calculate the expected frequencies.

In this example, only one sample constant is used

$\rightarrow n$ (total sample size $n = \sum_{i=1}^{k} \text{obs}_i$)

When *n* is a known, only *k* - 1 frequencies can be specified; the last is determined by subtraction and therefore "fills itself in automatically".

Assumptions

- Sample observations are a random sample of the population.
- Sample observations are independent.

Constraints

- Data are categorical.
- For chi-square approximation:
  (i) none of the expected frequencies < 1.0
  (ii) no more than 20% of the expected frequencies < 5.0
  (i) is more important than (ii).

Set a significance level such as $\alpha = 0.05$

H_1 is always two-tailed -- we consider only general alternatives.

Critical value $\chi^2_{0.05,3} = 7.815$ (Table A-4, page 775)

Decision rule: H_0 rejected if $\chi^2_{obs} > \chi^2_{crit}$. If test statistic is 7.815 or greater, then reject H_0.

Step 7: Data:

Category	Obs.	Expected	Obs - Exp	$\frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$
YR	2504	4530 * 9/16 = 2548.1	-44.1	.763
Yw	853	4530 * 3/16 = 849.4	3.6	.015
gR	881	4530 * 3/16 = 849.4	31.6	1.176
gw	292	4530 * 1/16 = 283.1	8.9	.280
	$n =$ 4530	4530		$\chi^2_{obs} = 2.234$

$\chi^2_{obs} = 2.234 < 7.815$

0.10 < P-value < 0.90

Therefore, do not reject H_0.

No significant difference between observed frequency and the frequency predicted by Mendelian genetics.

Testing Categorical Probabilities: Two-Way tables

Now consider multinomial experiments where data are classified according to two criteria--- ***classification wrt two qualitative factors***

Study, based on a survey of 300 TV viewers, looking at relationship between gender of a viewer and the viewer's brand awareness

		Gender		
		Male	**Female**	**Totals**
Brand Awareness	**Could identify product**	95	41	136
	Could not identify product	55	109	164
	Totals	150	150	300

This is an example of a ***two-way contingency table***

		Gender		
		Male	**Female**	**Totals**
Brand Awareness	**Could identify product**	n_{11}	n_{12}	r_1
	Could not identify product	n_{21}	n_{22}	r_2
	Totals	c_1	c_2	n

		Gender		
		Male	**Female**	**Totals**
Brand Awareness	**Could identify product**	p_{11}	p_{12}	p_{r1}
	Could not identify product	p_{21}	p_{22}	p_{r2}
	Totals	p_{c1}	p_{c2}	1

p_{r1}, p_{r2}, p_{c1}, and p_{c2} are the called ***marginal probabilities*** for each row ar respectively-- $p_{r1} = p_{11} + p_{12}$

Suppose we want to know if the two classifications are independent?

Does knowing the gender of the TV viewer provide information about the viewer's brand awareness?

H_0 : Brand awareness and gender are statistically independent.

H_1 : Brand awareness and gender are statistically dependent.

Remember, if two events A and B are independent, $P(AB) = P(A)P(B)$

In the analysis of contingency tables, if the two classifications are independent,

$$p_{11} = p_{r1}p_{c1} \quad p_{12} = p_{r1}p_{c2}$$

$$p_{21} = p_{r2}p_{c1} \quad p_{22} = p_{r2}p_{c2}$$

To test the hypothesis of independence, first need the ***expected or mean count in each cell***

$$E(n_{11}) = np_{11}$$

When the null hypothesis is true

$$E(n_{11}) = np_{r1}p_{c1}$$

Since p_{r1} and p_{c1} are unknown, they are estimated by $\hat{p}_{r1} = \frac{r_1}{n}$ and $\hat{p}_{c1} = \frac{c_1}{n}$

By substitution, the estimate of the expected value $E(n_{11})$ is

$$\hat{E}(n_{11}) = n\left(\frac{r_1}{n}\right)\left(\frac{c_1}{n}\right) = \frac{r_1 c_1}{n}$$

The general formula is

$$\hat{E}(n_{ij}) = \frac{(\text{Row total})(\text{Column total})}{\text{Total sample size}}$$

For the TV viewer example:

$$\hat{E}(n_{11}) = \frac{r_1 c_1}{n} = \frac{(136)(150)}{300} = 68$$

$$\hat{E}(n_{12}) = \frac{r_1 c_2}{n} = \frac{(136)(150)}{300} = 68$$

$$\hat{E}(n_{21}) = \frac{r_2 c_1}{n} = \frac{(164)(150)}{300} = 82$$

$$\hat{E}(n_{22}) = \frac{r_2 c_2}{n} = \frac{(164)(150)}{300} = 82$$

Use the χ^2 statistic to test the null hypothesis of independence---compare observed and expected counts in each cell

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left[n_{ij} - \hat{E}(n_{ij})\right]^2}{\hat{E}(n_{ij})}$$

$$\chi^2 = \frac{(95-68)^2}{68} + \frac{(41-68)^2}{68} + \frac{(55-82)^2}{82} + \frac{(109-82)^2}{82} = 39.22$$

What do large values of χ^2 imply?

To determine the cutoff value, we are making use of the fact that the sampling distribution of the χ^2test statistic is approximately a χ^2 probability distribution when the classifications are independent (under the null hypothesis)

The appropriate degrees of freedom in a 2-way table are:

$$(r-1)(c-1)$$

where r is the number of rows and c is the number of columns

The methods we have learned for a 2x2 two-way table can be generalized to problems with more categories. Could have a 3x2, 4x4, etc.

For the TV viewer example,

$$\text{df} = (2-1)(2-1) = 1$$

For $\alpha = .05$, reject the hypothesis of independence when $\chi^2 > \chi^2_{.05,1} = 3.841$

Since $\chi^2 = 39.22$ exceeds 3.841 conclude that viewer gender and brand awareness are statistically dependent events at $\alpha = .05$ significance level.

Using Table A-4 we can also conclude that $p < 0.005$

Good software can be found at:
http://faculty.vassar.edu/lowry/VassarStats.html

Choose Frequency Data from the menu on the left

Choose either Chi-Square "Goodness of Fit" Test, or Chi-Square, Cramer's V, and Lambda

A Word of Caution About Chi-square Tests

- Should be avoided when expected cell counts are too small (< 5)
- If $\chi^2 < \chi^2_{\alpha,\text{df}}$ ***do not accept the hypothesis of independence***---risking a Type II error and the probability, β of committing such an error is unknown
- If $\chi^2 > \chi^2_{\alpha,\text{df}}$ avoid inferring that a ***causal*** relationship exists between the classifications
    - the alternative hypothesis states that the two classifications are statistically dependent
    - statistical dependence does not imply causality