STAT 212: Principles of Statistics II

Lecture Notes: Chapter 3

Categorical Data Analysis

Patricia Ning

Dept. of Statistics

Texas A&M University

Categorical Data Analysis

What are categorical data?

Categorical data are specified by the *number* of cases that fall into various *categories (or levels)* of a categorical variable (a.k.a. trait, factor etc.). They are often also referred to as *count data*. (Since we count how many cases fall into each category.)

Example 15: *Drug usage in a community*

Category 1: Never used drug

Category 2: Have used drug on at least one occasion but not more than 5

Category 3: Have used drug on more than 5 occasions

Suppose we randomly select 100 persons and have them fill out a questionnaire on drug use. The data may be summarized as follows:

Category	1	2	3	Total
Count	72	18	10	100

We may also *cross-classify* data, which means we categorize according to two different traits. In the current example, *level of drug use* is one trait and *person's age* might be another.

Cross-classified data may be summarized with a table as follows:

		Drug use			
		1	2	3	Total
Age	Under 18	18	1	1	20
	18-25	17	6	2	25
	25-40	11	5	3	19
	Over 40	26	6	4	36
	Total	72	18	10	100

We'll discuss various ways of analyzing categorical data. A fundamental model for categorical data is the *multinomial experiment*.

A multinomial experiment is a generalization of the *binomial experiment*.

Binomial experiment

- n independent trials
- Each trial is a success (S) or failure (F).
- Probability of S on any given trial is p. $P(F) = 1 - p$.

If N is the number of successes in a binomial experiment with n trials, then

$$P(N = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, \ldots, n, \\ 0, & \text{otherwise.} \end{cases}$$

Multinomial experiment – exactly the same as binomial experiment except that there are r possible outcomes on each trial, where $r \geq 2$.

Each trial results in one and only one of the r possible outcomes.

$$P(\text{a trial results in category } i) = p_i$$

$$\sum_{i=1}^{r} p_i = 1$$

In n independent trials, let N_i be the number of trials resulting in category i.

$$\sum_{i=1}^{r} N_i = n$$

What is the joint probability distribution of $N_1, N_2, \ldots, N_r$*?*

Let $n_1, \ldots, n_r$ be any nonnegative integers that add up to n. Then

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_r = n_r) =$$

$$\left(\frac{n!}{n_1! n_2! \cdots n_r!}\right) p_1^{n_1} \cdots p_r^{n_r}.$$

For any other sort of $(n_1, \ldots, n_r)$,

$$P(N_1 = n_1, \ldots, N_r = n_r) = 0.$$

This distribution is called the *multinomial distribution*. Notice that when $r = 2$ it reduces to the binomial distribution.

It can be shown that for each $i = 1, \ldots, r$,

$$E(N_i) = np_i \text{ and } Var(N_i) = np_i(1 - p_i).$$

Often $p_1, \ldots, p_r$ are *population proportions*.

In Example 15 (pg. 172N)

p_i = proportion of community in category i, $i = 1, 2, 3$.

If we randomly select n individuals (without replacement) from the population and record how many fall into each category, then to a good approximation we have done a multinomial experiment.

Notes:

- The only reason this isn't *exactly* a multinomial experiment is that the sampling is not done *with* replacement.

- When sampling is done without replacment (the common practice), then the multinomial approximation is good so long as n is small relative to the size of the population. (Typically populations are large or infinite. So any n, large or small, should be fine.)

Goodness-of-fit tests for category probabilities

A fundamental problem in categorical data analysis is testing how well a model for the category probabilities fits the data.

For example, in genetics studies it is often of interest to test whether the offspring of certain parents are in agreement with what genetic theory predicts.

The theory might predict that there are three categories of offspring which should occur in a 1:2:1 ratio, meaning that the category probabilities would be

$$p_1 = \frac{1}{4}, \quad p_2 = \frac{1}{2} \quad \text{and} \quad p_3 = \frac{1}{4}.$$

The geneticist might do a breeding experiment leading to counts of offspring in each category.

She would then like to see if the discrepancy of counts from the 1:2:1 ratio is small enough to be attributed to chance.

There are two cases of interest:

- Category probabilities are completely specified by H_0.

- Category probabilities are constrained in some way by H_0, but not completely specified.

The genetics example on the previous page was of the first type.

We'll start out by discussing the first case. Suppose we want to test

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, \ldots, p_r = p_{r0},$$

where $p_{10}, \ldots, p_{r0}$ are *fixed and known* probabilities that add up to 1.

Now suppose that a multinomial experiment is done and the observed counts are to be used to test H_0. The test is based on a simple idea.

Compare observed category counts with expected counts assuming that the null hypothesis is true.

If we do n trials and if H_0 is true, then the expected count in category i is

$$E(N_i) = np_{i0}, \quad i = 1, \ldots, r.$$

For any given category, we can measure the discrepancy between *observed* and *expected* by using

$$\frac{(N_i - np_{i0})^2}{np_{i0}}.$$

An *overall* measure of discrepancy between observed and expected is

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - np_{i0})^2}{np_{i0}}.$$

When H_0 is true, χ^2 has, approximately, the χ^2 distribution with $r-1$ degrees of freedom.

H_0 is rejected at level of significance α if

$$\chi^2 \geq \chi^2_{r-1,\alpha}.$$

The χ^2 percentiles are in Table A.5, pg. 497.

Conditions required in order for the χ^2 approximation to be valid:

- $r \geq 3$
- The smallest np_{i0} should be at least $5(m/r)$, where m is the number of np_{i0} that are less than 5.

If $np_{i0} \geq 5$ for all i, then you're good to go!

Example 16: *A grand jury study*

A study of grand juries in a certain county compared the demographic characteristics of jurors with those of the general population to see if the jury panels were representative.

Here are the results for age:

Category	Age	County-wide percentage	No. of jurors
1	21-40	42	5
2	41-50	23	9
3	51-60	16	19
4	Over 60	19	33
		100	66

Let p_i = probability that a juror is selected from age category i.

We want to *test the hypothesis that the 66 jurors were randomly selected from the population in the county aged 21 and over*.

$$H_0: \quad p_1 = 0.42,\ p_2 = 0.23,\ p_3 = 0.16,\ p_4 = 0.19$$

$$np_{10} = 66(0.42) = 27.72$$
$$np_{20} = 66(0.23) = 15.18$$
$$np_{30} = 66(0.16) = 10.56$$
$$np_{40} = 66(0.19) = 12.54$$

$$\chi^2_{3,0.05} = 7.815$$

H_0 will be rejected if $\chi^2 \geq 7.815$.

$$\chi^2 = \frac{(5-27.72)^2}{27.72} + \frac{(9-15.18)^2}{15.18} + \frac{(19-10.56)^2}{10.56} + \frac{(33-12.54)^2}{12.54}$$

$$= 61.27$$

So, we reject H_0. In fact, $61.27 > \chi^2_{3,0.005}$, implying that the P-value is smaller than 0.005.

There is strong evidence of bias in the jury selection process. Older jurors are heavily overrepresented.

χ^2 test when category probabilities are not completely specified

When H_0 does not completely specify the category probabilities, some parameters will have to be *estimated from the data* in order to calculate the expected counts.

Example 17 *Testing whether count data follow a Poisson distribution*

It is of interest to know if the weekly number of traffic accidents at a busy intersection follows a Poisson distribution.

Suppose a category represents some no. of traffic accidents in a week.

0 1 2 3 4 more than 4

p_{i+1} = probability of i accidents in one week, $i = 0, 1, 2, 3, 4$

$$p_6 \;=\; \text{probability of at least 5 accidents in one week}$$

If the weekly number of accidents follows a Poisson(θ) distribution, i.e. a Poisson distribution with mean $\theta > 0$, then

$$p_{i+1} = \frac{e^{-\theta}\theta^i}{i!} = \pi_{i+1}(\theta), \quad i = 0, \ldots, 4,$$

and

$$p_6 = 1 - \sum_{j=0}^{4} \frac{e^{-\theta}\theta^j}{j!} = \pi_6(\theta).$$

The null hypothesis is

$$H_0 : \; p_1 = \pi_1(\theta), \ldots, p_6 = \pi_6(\theta).$$

This doesn't completely specify the p_is because θ can be any positive number.

A way around this difficulty is *to estimate θ from the data*.

Ideally, one should use the method of *minimum* χ^2 to estimate parameters. This means one computes a χ^2 statistic as discussed on pg. 180-181N *for each value of* θ, and then chooses θ to minimize the statistic.

In Example 17, θ represents the average number of accidents per week. Suppose we observe the intersection over a period of n weeks.

We will define $\widehat{\theta}$ to be the average number of traffic accidents over the n weeks. To a good approximation, $\widehat{\theta}$ is the minimum χ^2 estimator of θ.

How to do goodness-of-fit test when parameters are left unspecified:

$$H_0: \ p_1 = \pi_1(\theta), \ldots, p_r = \pi_r(\theta)$$

$\theta = (\theta_1, \ldots, \theta_\ell)$ is unknown.

- We do a multinomial experiment with n trials and observe the number, N_i, of occurrences in category i, $i = 1, \ldots, r$.

- Estimate $(\theta_1, \ldots, \theta_\ell)$ from the count data using the method of minimum χ^2. Call these estimates $(\widehat{\theta}_1, \ldots, \widehat{\theta}_\ell)$.

- Compute the estimated expected cell counts: $n\pi_i(\widehat{\theta})$.

- Compute the χ^2 statistic:

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - n\pi_i(\widehat{\theta}))^2}{n\pi_i(\widehat{\theta})}.$$

When H_0 is true, χ^2 has approximately the χ^2 distribution with $r - 1 - \ell$ degrees of freedom (the extra reduction by l in the df is because of estimating l many parameters in computing the χ^2 statistic). **Note:** *It must be true that* $r - 1 > \ell$.

Reject H_0 at level α if

$$\chi^2 \geq \chi^2_{r-1-\ell,\alpha}.$$

Example 17: continued

In our traffic example, the intersection was observed for $n = 100$ weeks. The number of accidents each week was recorded. The data are on the next page.

No. accidents	Count
0	10
1	31
2	25
3	20
4	7
5	3
6	2
7	1
8	0
9	1

H_0 : weekly number of accidents follows a Poisson distribution

Need to estimate θ. Use the sample mean number of accidents.

$$\begin{aligned} \widehat{\theta} &= (10 \cdot 0 + 31 \cdot 1 + \cdots + 1 \cdot 9)/100 \\ &= 212/100 \\ &= 2.12 \end{aligned}$$

Now we calculate estimated expected cell counts.

$$\pi_i(\hat{\theta}) = \frac{e^{-\hat{\theta}}\,\hat{\theta}^{i-1}}{(i-1)!}, \quad i = 1, \ldots, 5$$

$$\pi_1(\hat{\theta}) = e^{-2.12} = 0.1200$$

$$\pi_2(\hat{\theta}) = e^{-2.12}(2.12) = 0.2545$$

$$\pi_3(\hat{\theta}) = e^{-2.12}(2.12)^2/2 = 0.2697$$

$$\pi_4(\hat{\theta}) = e^{-2.12}(2.12)^3/6 = 0.1906$$

$$\pi_5(\hat{\theta}) = e^{-2.12}(2.12)^4/24 = 0.1010$$

$$\hat{\pi}_6(\hat{\theta}) = 0.0642$$

No. accidents	Expected count	Observed count
0	12	10
1	25.45	31
2	26.97	25
3	19.06	20
4	10.1	7
≥ 5	6.42	7

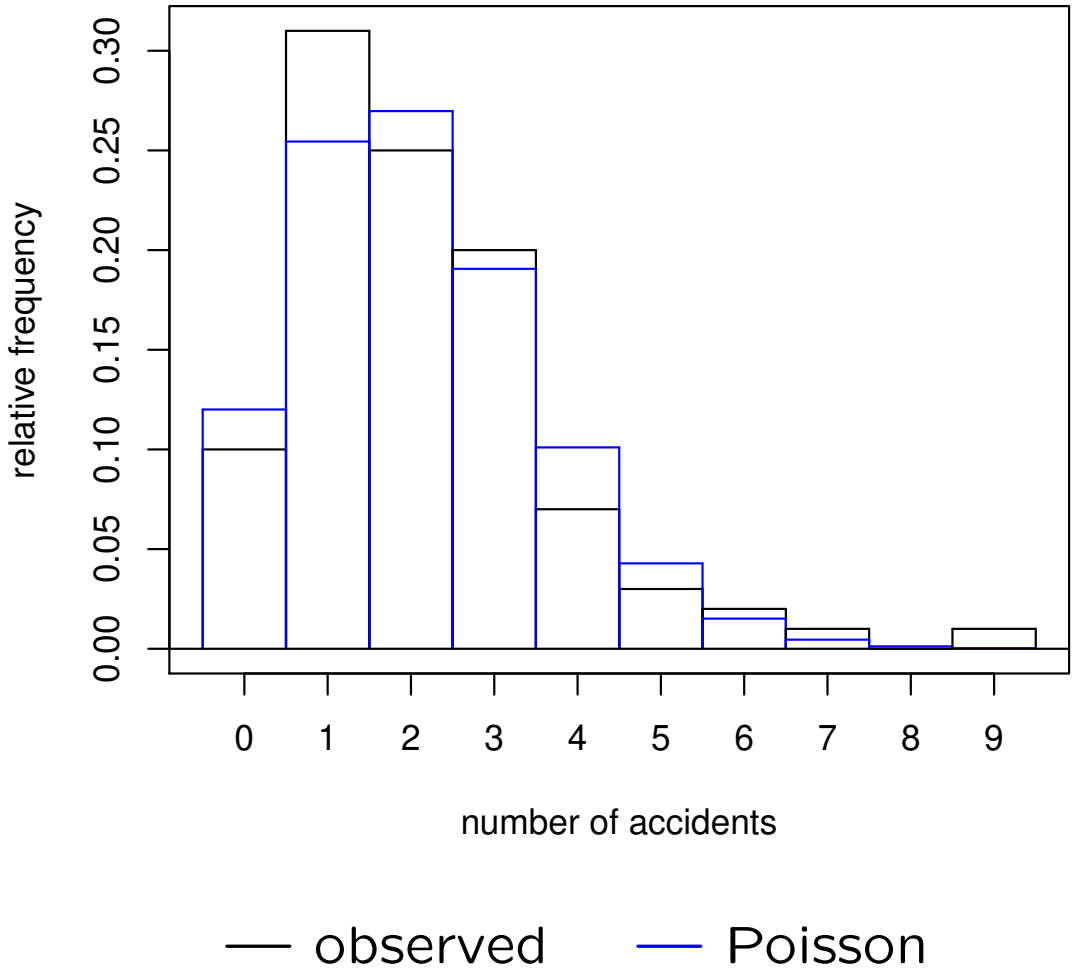
H_0 will be rejected at level of significance 0.05 if

$$\chi^2 \geq \chi^2_{6-1-1,0.05} = \chi^2_{4,0.05} = 9.488.$$

$$\begin{aligned} \chi^2 &= \frac{(10-12)^2}{12} + \frac{(31-25.45)^2}{25.45} + \cdots \\ &= 2.738 \end{aligned}$$

Since $2.738 < 9.488$, *we cannot reject H_0. The observed accident counts are consistent with what we'd expect if the number of accidents follow a Poisson distribution*.

Comparison of empirical and Poisson relative frequencies



Using a χ^2 goodness-of-fit test to check whether data are Normally distributed

Throughout 211 and 212, we've used the assumption that our data are Normally distributed.

An informal way to test the hypothesis of normality is to examine a Normal probability plot, as we have often done.

A more rigorous way to test Normality is to use the χ^2 goodness-of-fit test discussed on pg. 187-189N. We do so as follows:

- Divide the range of the data up into r class intervals (as when constructing a histogram).

- Count the number of data values that fall into the class intervals. Denote these counts as: $N_1, N_2, \ldots, N_r$.

- Compare these counts to what we'd *expect* in each class if the data were *really* Normally distributed.

To compute expected counts, one needs to know the mean and variance of the Normal distribution. Estimate these parameters by $\bar{X}$ (the sample mean) and s^2 (sample variance).

Given the mean and variance of a Normal distribution, one can determine the proportion of the distribution that falls into any particular class interval.

For example, suppose a class interval goes from 10 to 20, and $\bar{X}$ and s are 30 and 15, respectively.

A Normal distribution with mean 30 and standard deviation 15 has the same area between 10 and 20 as the *standard* Normal distribution has between

$$\frac{10-30}{15} = -1.33 \quad \text{and} \quad \frac{20-30}{15} = -0.67.$$

The area is $0.2514 - 0.0918 = 0.1596$.

So, the expected count in this class interval would be $n(0.1596)$, where n is the total sample size of the data.

Example 19: *Buffalo snowfall data*

The data are yearly snowfall totals (in inches) in Buffalo, NY over a 63 year period.

$$n = 63 \qquad \bar{X} = 80.30 \qquad s = 23.53$$

Class interval	Expected	Observed
≤ 40	2.733	3
40-60	9.498	11
60-80	18.949	18
80-100	19.143	16
100-120	9.793	11
> 120	2.884	4

Example of computing expected count:

Suppose that $X \sim N(80.3, 23.53^2)$. Then

$$P(40 < X < 60) =$$

$$P\left(\frac{40 - 80.3}{23.53} < Z < \frac{60 - 80.3}{23.53}\right) =$$

$$P(-1.71 < Z < -0.86) =$$

$$0.1949 - 0.0436 = 0.1513.$$

So, the expected count in the interval $(40, 60)$ is $63(0.1513) \approx 9.5$.

$$\begin{aligned} \chi^2 &= \frac{(2.733-3)^2}{2.733} + \frac{(9.498-11)^2}{9.498} + \cdots \\ &= 1.407785 \end{aligned}$$

This is a case where *H_0 does not completely specify the category probabilities*. H_0 just says the data are Normally distributed. It does not say what the mean and variance are, which we have to know to compute the probabilities.

So, the degrees of freedom in our example are $6 - 1 - 2 = 3$. The 2 is subtracted off since we estimated two parameters, the mean and variance.

$$\chi^2_{6-2-1,.05} = \chi^2_{3,.05} = 7.815$$

Since $1.407785 < 7.815$, we do not reject H_0. The Buffalo snowfall data are consistent with the assumption of Normality.

Contingency tables: testing for homogeneity of proportions

Recall the cross-classified data in *Example 15*. The table in which the data were displayed is called a *contingency table*, since the relative frequency of values in a column is contingent upon the row.

A fundamental contingency table problem is as follows:

- One is interested in several, say k, populations ($k \geq 2$).

- Each population can be divided into the same r categories ($r \geq 2$).

- One wishes to see if the category proportions are the same for all k populations.

Suppose we want to compare the distribution of traffic accidents at different intersections. We might have the following table:

		No. of traffic accidents						
		0	1	2	3	4	≥ 5	
	1	10	31	25	20	7	7	100
Intersec.	2	18	50	23	7	2	0	100
	3	13	45	30	8	3	1	100

The problem of interest is analogous to the ANOVA problem. *In both cases, we wish to see if the distributions of several different populations are the same.*

The main difference between the two is that one involves categorical data and the other continuous data.

Let p_{ij} be the proportion of population i that falls into category j, $i = 1, \ldots, k$ and $j = 1, \ldots, r$. Note that

$$\sum_{j=1}^{r} p_{ij} = 1, \quad i = 1, \ldots, k.$$

In other words, the probabilities for any given population add up to 1.

A multinomial experiment is done in each population.

$n_i =$ number of trials in sample i, $i = 1, \ldots, k$

n_{ij} = number of trials in sample i that fall into category j

The null hypothesis of interest is:

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{kj}, \quad j = 1, 2, \ldots, r.$$

Categories

Popn.		1	2	$\cdots$	r	
	1	p_{11}	p_{12}	$\cdots$	p_{1r}	1
	2	p_{21}	p_{22}	$\cdots$	p_{2r}	1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	k	p_{k1}	p_{k2}	$\cdots$	p_{kr}	1

Define

$$n_{\cdot j} = \sum_{i=1}^{k} n_{ij}, \quad j = 1, \ldots, r.$$

If H_0 is true, the k different multinomial experiments constitute one big multinomial experiment with number of trials equal to $n_1 + n_2 + \cdots + n_k = n$ and category probabilities $p_1, \ldots, p_r$.

For example,

$$p_1 = p_{11} = p_{21} = \cdots = p_{k1}.$$

We estimate p_j by $n_{\cdot j}/n$. The expected count in "cell" (i, j) is $n_i p_j$, which is estimated by $n_i n_{\cdot j}/n$.

We use a χ^2 statistic as before to test H_0.

- *Observed cell count:* n_{ij}

- *Expected cell count:* $\widehat{e}_{ij} = n_i n_{\cdot j}/n$

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - \widehat{e}_{ij})^2}{\widehat{e}_{ij}}.$$

When H_0 is true, χ^2 has the χ^2 distribution with degrees of freedom $(k-1)(r-1)$.

H_0 is rejected at level of significance α when

$$\chi^2 \geq \chi^2_{(k-1)(r-1),\alpha}.$$

Example 18: *Comparing cities in a market study*

The marketing division of a certain company wanted to determine if three cities differed with respect to how aware they are of a product. Random samples of 200, 150 and 300 consumers are taken from the three cities.

Category 1: "Never heard of product"

Category 2: "Heard about product but have never bought it"

Category 3: "Bought product at least once"

The following data were collected:

Category

	1	2	3	Total
City 1	36	55	109	200
City 2	45	56	49	150
City 3	54	78	168	300
Total	135	189	326	650

$$H_0: \quad p_{1j} = p_{2j} = p_{3j}, \quad j = 1, 2, 3.$$

Table of expected values

	1	2	3	Total
City 1	41.54	58.15	100.31	200
City 2	31.15	43.62	75.23	150
City 3	62.31	87.23	150.46	300
Total	135	189	326	650

For example,

$$\widehat{e}_{11} = 135(200)/650 = 41.54$$

and

$$\widehat{e}_{21} = 135(150)/650 = 31.15.$$

Suppose we test H_0 with $\alpha = .05$. The appropriate degrees of freedom are $(k-1)(r-1) = 2 \cdot 2 = 4$. H_0 is rejected if $\chi^2 \geq \chi^2_{4,0.05} = 9.488$.

$$\begin{aligned}\chi^2 &= \frac{(36-41.54)^2}{41.54}+\frac{(45-31.15)^2}{31.15}+\cdots \\ &\quad +\frac{(168-150.46)^2}{150.46} \\ &= 24.61\end{aligned}$$

The P-value is less than .005. We thus reject H_0. It appears that the category probabilities differ in some way between the cities.

The following table of proportions makes it more clear how the cities differ. Cities 1 and 3 are similar, while City 2 is substantially different from the other two.

	1	2	3
City 1	0.180	0.275	0.545
City 2	0.300	0.373	0.327
City 3	0.180	0.260	0.560

Contingency tables: testing for independence of row and column categories

Suppose now that we have a single sample of n individuals or items from a population and each item is classified according to two traits.

For example, suppose a sample of persons is selected from all registered voters in Brazos County. The two traits of interest are

- Political affiliation: Republican, Democrat, other

- Attitude on gun control: In favor of more stringent gun control, against more stringent gun control, no opinion

Let's say a sample of 500 voters is obtained, leading to the following table:

	Favor	Against	No opinion	Total
Republican	100	166	55	321
Democrat	68	27	12	107
Other	28	24	20	72
Total	196	217	87	500

Of interest is determining whether or not attitude on gun control is independent of political affiliation.

In other words, are the distributions of attitudes similar for Democrats, Republicans and others?

This situation differs in two ways from the previous:

- There is only one population and one sample and two different traits are observed.

- Neither row nor column totals are fixed in advance of sampling.

Define p_{ij} to be the proportion of the population in cell (i, j), where $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$.

Furthermore, let $p_{i\cdot}$ be the proportion of the population in row i and $p_{\cdot j}$ the proportion of the population in column j.

Note that

$$p_{i\cdot} = \sum_{j=1}^{J} p_{ij} \quad \text{and} \quad p_{\cdot j} = \sum_{i=1}^{I} p_{ij}.$$

The null hypothesis of interest is:

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \quad \text{for all } i \text{ and } j.$$

H_0 says that *the event of falling into row i is independent of the event of falling into column j for all i and j.*

Recall that events A and B are independent if and only if

$$P(A \cap B) = P(A)P(B).$$

Since the conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

we may also say that A and B are independent if and only if

$$P(A|B) = P(A).$$

So, H_0 is equivalent to saying that, *for each i and j, the conditional probability that a population member falls into row i given that the member is in column j, is the same as the unconditional probability that a member falls into row i.*

In terms of our political setting, this would mean, for example, that the proportion of all Republicans who favor more stringent gun control is the same as the proportion of all Democrats who favor more stringent gun control.

We test H_0 in exactly the same way as in the "homogeneity of proportions" setting. Let n_{ij} be the number of items in the sample that are in row i and column j. Define $\widehat{e}_{ij}$ as before:

$$\widehat{e}_{ij} = \frac{(i\text{th row total})(j\text{th column total})}{n}.$$

Also as before, let

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\widehat{e}_{ij})^2}{\widehat{e}_{ij}} \\ &= n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(\widehat{p}_{ij}-\widehat{p}_{i\cdot}\widehat{p}_{\cdot j})^2}{\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}},\end{aligned}$$

where $\widehat{p}_{ij} = n_{ij}/n$,

$$\widehat{p}_{i\cdot} = n_{i\cdot}/n \quad \text{and} \quad \widehat{p}_{\cdot j} = n_{\cdot j}/n.$$

Here, $n_{i\cdot}$ = ith row total = $\sum_{j=1}^{J} n_{ij}$, and $n_{\cdot j}$ = jth column total = $\sum_{i=1}^{I} n_{ij}$.

H_0 is rejected at level α if:

$$\chi^2 \geq \chi^2_{(I-1)(J-1),\alpha}.$$

In our Brazos county example, suppose we want to test independence of political affiliation and attitude at level of significance 0.05. We have

$$\chi^2_{(I-1)(J-1),\alpha} = \chi^2_{4,0.05} = 9.488.$$

The table of expected values is:

	Favor	Against	No opinion	Total
Republican	125.83	139.32	55.85	321
Democrat	41.94	46.44	18.62	107
Other	28.22	31.25	12.53	72
Total	196	217	87	500

The value of the χ^2 statistic is 43.25. So, we reject H_0 and conclude that attitude depends on political affiliation.

The following table of estimated conditional probabilities is useful for describing the nature of the dependence.

	Favor	Against	No opinion	Total
Republican	0.31	0.52	0.17	1
Democrat	0.64	0.25	0.11	1
Other	0.39	0.33	0.28	1

Obviously, Democrats are more in favor of more stringent gun control than Republicans. Opinions are fairly evenly divided between the three categories for "others."

Remarks

- The tests for homogeneity of proportions and independence are exactly the same. However, the sampling methods and hypotheses are different. *In contingency tables with more than two traits, these differences become more important since the appropriate tests are also different.*

- In all settings we have considered, validity of the χ^2 test is insured if all expected counts are at least 5.