Read M&M Chapter 9. Chi square approximations and tests, with two-way tables as an example.

1. Counts and cross-classification

Many data sets come in the form of counts of individual cases classified into one of a finite set of disjoint categories; the data consist of the numbers of observations falling into each category. For example, persons may be classified as male or female; summonses may be classified as undeliverable or not; cups of coffee may be classified as small, medium or large (or as regular, grande, and humungo); expensive consumer items may be classified as deluxe, super-deluxe, and really good.

Sometimes the classification is determined by two (or more) factors. For example, individuals may be classified as either male or female, and as either dead, unhealthy, or healthy. Such data are often presented in the form of a two-way table (for two classifying factors), or as a set of two-way tables (for three classifying factors). Counts for greater numbers of classifying factors are more conveniently displayed in other ways.

For example, the following table shows (a subset of) some Census Bureau estimates of population:

year	state	county	R/S/E	under5	5–9	10–14	15–19	20–24	25–29	30–34	...
90	09	000	1	91295	81828	75729	84252	100951	116271	123865	...
90	09	000	2	86206	77762	72284	79673	97065	114231	123980	...
90	09	000	3	11392	9623	9078	8888	10035	10037	8576	...
...	...	...	...	...	...	...	...	...	...	...	...
90	09	001	1	21532	18342	17529	19034	21306	25717	28196	...
90	09	001	2	20022	17499	16982	18102	21222	25899	28665	...
...	...	...	...	...	...	...	...	...	...	...	...
91	09	000	1	91301	83866	77055	79237	96535	110792	123593	...
91	09	000	2	86500	79553	73700	74619	93187	108942	123560	...
...	...	...	...	...	...	...	...	...	...	...	...
94	09	015	11	377	309	266	223	226	204	193	...
94	09	015	12	336	300	256	210	231	226	227	...

The first column indicates the year (90 = 1990), the second column gives a state code (09 = CT), the third column gives a code for county (000 = whole of CT, 009 = New Haven County), the fouth column gives a "Race/Sex/Ethnicity Indicator" (1 = White Non-Hispanic Male, 2 = White Non-Hispanic Female, etc), and the remaining columns give estimated counts for various age ranges (Under 5 years, 5 to 9 years,...). The data represent a five-way (or is it seven-way?) cross-classification.

Sometimes we have probabilistic models for how individuals end up in one of the categories, and we are interested in checking how well the observed counts conform to the theoretical model. Sometimes we try to build probabiltiy models to 'explain' the distribution of counts across the categories.

2. Summonses to towns

When Judicial Information Systems (JIS) compiles the master list of potential jurors it is supposed to select fixed proportions from each town: the Connecticut General Statutes require that the numbers taken from each town in the judicial district be in proportion to the total population of the town, as determined by the 1990 Census.

At one stage during my analysis of the JIS data (the records of the summonses actually mailed), I developed serious doubts about whether JIS was actually following the statute. I was able to check one aspect of JIS procedure by comparing the proportions in their summary files for each court year with the Census figures. Knowing that

summonses were drawn from the master file in the order of a randomly allocated juror ID number, I reasoned that the proportions in the summary file should have reflected the proportions in the master file.

Even after taking the sampling variability into account, I was not expecting to find close agreement between summary file proportions and Census proportions, because:

(i) I was not allowed to see a small proportion of the juror records, for confidentiallity reasons

(ii) The address in the summary file did not always reflect the original address to which a summons was sent. JIS policy was unclear regarding the procedure taken when a change of address was needed, but it was clear that some addresses were being overwritten by updated information. In particular, a small proportion of potential jurors (4537 out of 88984, for 1992-93) were excused because they no longer lived in the judicial district.

The following table shows the story for the court year 1992-93, after exclusion of summonses with an address ouside the judicial district.† Note: the table is not a cross-classification on two factors.

	pop90	pct90	observed	expected	obs. − exp.	chi
AVON	13937	1.6	1311	1365.8	−54.8	−1.5
BERLIN	16787	1.9	1690	1645.1	44.9	1.1
BLOOMFIELD	19483	2.3	1930	1909.3	20.7	0.5
BRISTOL	60640	7	5999	5942.5	56.5	0.7
BURLINGTON	7026	0.8	707	688.5	18.5	0.7
CANTON	8268	1	787	810.2	−23.2	−0.8
EAST GRANBY	4302	0.5	427	421.6	5.4	0.3
EAST HARTFORD	50452	5.9	4946	4944.1	1.9	0
EAST WINDSOR	10081	1.2	986	987.9	−1.9	−0.1
ENFIELD	45532	5.3	4319	4462	−143	−2.1
FARMINGTON	20608	2.4	2044	2019.5	24.5	0.5
GLASTONBURY	27901	3.2	2731	2734.2	−3.2	−0.1
GRANBY	9369	1.1	942	918.1	23.9	0.8
HARTFORD	139739	16.2	13537	13693.9	−156.9	−1.3
MANCHESTER	51618	6	5078	5058.4	19.6	0.3
MARLBOROUGH	5535	0.6	535	542.4	−7.4	−0.3
NEW BRITAIN	75491	8.8	7406	7397.8	8.2	0.1
NEWINGTON	29208	3.4	2880	2862.3	17.7	0.3
PLAINVILLE	17392	2	1742	1704.3	37.7	0.9
PLYMOUTH	11822	1.4	1150	1158.5	−8.5	−0.2
ROCKY HILL	16554	1.9	1630	1622.2	7.8	0.2
SIMSBURY	22023	2.6	2093	2158.2	−65.2	−1.4
SOUTH WINDSOR	22090	2.6	2215	2164.7	50.3	1.1
SOUTHINGTON	38518	4.5	3653	3774.6	−121.6	−2
SUFFIELD	11427	1.3	1109	1119.8	−10.8	−0.3
WEST HARTFORD	60110	7	6099	5890.5	208.5	2.7
WETHERSFIELD	25651	3	2519	2513.7	5.3	0.1
WINDSOR	27817	3.2	2809	2726	83	1.6
WINDSOR LOCKS	12358	1.4	1173	1211	−38	−1.1
total	861739	100	84447	84447.1	0	

$X^2 = 32.2$, p-value = 0.27

The first column gives the population of each town, according to the 1990 Census. The second column expresses the population as a percentage of the total population for the whole judicial district. The third column (headed "observed") gives the number of summonses that were mailed to each town, according to the JIS records. Don't try just yet to figure out what the other columns mean.

† By excluding the 4537 from the comparison, I am tacitly modelling address changes to have the same effect on each town.

Consider what happened for the City of Hartford. According to the Census, the City made up a fraction $p = 139739/861739 \approx 16.2\%$ of the total population of the judicial district. About 16.2% of the persons named on the master list should have come from Hartford. In a sample of size $n = 84447$ drawn from the master list, the mean number of Hartford adresses would have been $np \approx 13693.9$, the value given in the column headed "expected". (I write the values with one figure after the decimal point so that you do not confuse the values with any actually observed count.) This mean value is not too far from the the observed number from Hartford, 13537. Of course there would have been some difference due to sampling variability. As a rough approximation we could treat the observed number as an observation from a Bin(n, p) distribution, treating the selection of each name from the master like the toss of a coin that lands heads with probability $p \approx 16.2\%$. The standard deviation for the Binomial is $\sqrt{np(1-p)}$. To be on the conservative side, and allow JIS the benefit of more potential variability around the mean value, I will treat the observed count X as approximately $N(np, \sqrt{np})$. That is,

$$\frac{\text{observed count} - \text{expected count}}{\sqrt{\text{expected count}}}$$

should be roughly like an observation from a standard normal distribution (or perhaps from a normal distribution with a slightly smaller variance).‡

The value for Hartford in the column headed "chi" equals

$$\frac{13537 - 13693.9}{\sqrt{13693.9}} \approx -1.3$$

It would not be too suprising to see a standard normal taking such a value. (How likely is it that $|Z| \geq 1.3$ if Z has a standard normal distribution?) The figure for Hartford appears consistent with the model (proportions as prescribed by the statute).

Similar reasoning applies to each of the other towns. The values in the "chi" column should behave roughly like standard normals, if the proportions in the master file were as directed by statute. What do you think? Do they look reasonable?

It is traditional to assess the overall fit of the model by calculating the ***Pearson chi-square*** statistic $X^2 =$ the sum of the squares of the chi values. The reason for the name is that the statistic should behave roughly like a random variable with a chi-squre distribution if the model is correct.

In general, the ***chi-square distribution on k degrees of freedom*** is defined as the distribution of the sum of squares of k independent standard normal random variables. M&M (page 630) denote the distribution by $\chi^2(k)$.

The chi values are not quite independent, because they are derived from the differences between observed counts and expected values under a model that forces the sum of all the differences to be zero. Once we know 28 of the chi values we can figure out the value of the 29th by simple algebra. We lose one ***degree of freedom*** due to the constraint. It can be shown that the statistic X^2 is approximately distributed not like the sum of squares of 29 independent standard normals, but rather like the sum of squares of $28 = 29 - 1$ such variables. If the model is correct, the statistics X^2 should have approximately a $\chi^2(28)$ distribution.

As noted near the bottom right-hand corner of the table, X^2 actually takes the value 32.2. If a random variable T has a $\chi^2(28)$ distribution, $\mathbb{P}\{T \geq 32.2\} \approx 0.27$. We would see a value as large as, or larger than, the observed X^2 with about a 27% probability. That is, we have a p-value of about 0.27. The observed value of X^2 is not so unusual.

‡ Actually, even the Binomial overestimates the variability in a single count for a simple random sample. I wanted to use the larger standard deviation $\sqrt{np}$ to fit with the definition used for the chi-square statistic. In effect, the increase from $\sqrt{np(1-p)}$ to $\sqrt{np}$ compensates for a dependence between towns in the chi values. Also, there are other ways to argue for the larger standard deviation.

The observed counts are in reasonable agreement with the proportions that would be expected from the statute.

3. Sourcelists versus disqualification status

For the period covered by the challenge, the names for the master lists were drawn from two sources: the voter lists for each town, and the DMV lists of licensed drivers. For each record in the summary file, JIS attached a code indicating whether the person was on the DMV list alone, the voter list alone, or on both lists.

The possible effect of the different sourcelists is an important issue, particularly in view of recent moves to require JIS to use more sourcelists.

One way to compare the quality of the different sourcelists is to look at outcomes that might suggest poor address quality. The first of the following four tables (the one labelled [counts]) shows a cross-classification of 1993-94 summonses, according to the sourcelist and according to various types of outcome:

undel = initial summons returned by Post Office as undeliverable
undel2 = follow-up mailing returned by Post Office as undeliverable
NS = did not show up at court house
other = all other possibilities for a summons

[counts]	undel	undel2	NS	other	total
DMV	5546	583	1901	37612	45642
voter	2974	294	785	8670	12723
both	552	55	252	8599	9458
total	9072	932	2938	54881	67823

[row %]	undel	undel2	NS	other	total
DMV	12.2	1.3	4.2	82.4	100
voter	23.4	2.3	6.2	68.1	100
both	5.8	0.6	2.7	90.9	100

[table %]	undel	undel2	NS	other	total
DMV	8.2	0.9	2.8	55.5	67.4
voter	4.4	0.4	1.2	12.8	18.8
both	0.8	0.1	0.4	12.7	14
total	13.4	1.4	4.4	81	100

[col %]	undel	undel2	NS	other
DMV	61.1	62.6	64.7	68.5
voter	32.8	31.5	26.7	15.8
both	6.1	5.9	8.6	15.7
total	100	100	100	100

The bottom left table expresses counts as percentages of the total. You could regard the percentages in the body of the table as estimates of probabilities of events $S_i \cap D_j$, where S_1, S_2, and S_3 denotes the event that the person was listed on DMV, voter, or both sourcelists, and D_1 through D_4 correspond to the four outcomes. The marginal percentages estimate the probabilities of events S_i or D_j. For example, the event {name only on DMV list AND summons undeliverable} is estimated as having probability 0.082. The event {name only on DMV list} is estimated as having probability 0.674. The event {summons undeliverable} is estimated as having probability 0.134.†

The two tables on the right express the counts as row- and column-percentages, which have an interpretation as estimates of conditional probabilities.

If the two classifications were independent, we would have factorizations

$$\mathbb{P}(S_i \text{ AND } D_j) = \mathbb{P}(S_i) \times \mathbb{P}(D_j) \qquad \text{for each } i \text{ and } j$$

The fractions in the body of the [table%] table should then be close to the products of the marginal percentages. For example, the "voter AND undel" cell should contain a value close to $18.8\% \times 13.4\% \approx 2.52\%$. It actually contains the value 4.4%. Similarly, we could calculate the expected counts that would appear in the body of the count table, by multiplying products of marginal percentages by the total, 67823. That would give a

† The interpretation of the percentages is complicated by the fact that JIS worked with only a sample from each voter list. The actual overlap between the two sourcelists was much larger than suggested by the figures for 'both'.

table of 'estimated counts', under the model ('no association') that the row and column factors were acting independently:

[expected]	undel	undel2	NS	other	total
DMV	6105.1	627.2	1977.1	36932.6	45642
voter	1701.8	174.8	551.1	10295.2	12723
both	1265.1	130	409.7	7653.2	9458
total	9072	932	2938	54881	67823

Notice that the [counts] table and the [expected] table have the same marginal totals. The differences between corresponding entries in the tables would be attributed to random noise, under the model. If we rescale, dividing each difference by the square root of the value in the [expected] table, we get the [chi] table, whose entries should be (roughly) standard normals under the model of no association.

[chi]	undel	undel2	NS	other
DMV	-7.2	-1.8	-1.7	3.5
voter	30.8	9	10	-16
both	-20	-6.6	-7.8	10.8

$X^2 = 2080.5$

Clearly the no-association model is not doing a good job at predicting the observed counts. Can you see any pattern in these standardized differences (which I am calling chi values)?

The statistic X^2 is defined as the sum of the squares of the chi values. Under the independence model, it should behave roughly like an observation on a chi-square distribution, with $(3-1)\times(4-1) = 6$ degrees of freedom. The p-value for 2080.5 is so close to zero that it is not worth displaying.

In general, under the no-association model applied to a two-way table with r rows and c columns, the X^2 should have an approximate chi-square distrbution on $rc - (r+c-1) = (r-1)\times(c-1)$ degrees of freedom. The matching up of the marginal totals between the table of observed counts and the table of expected counts places $r+c-1$ constraints on the chi values.

It is not surprising that there is some association between the sourcelist and outcome. For example, only citizens can be on the voter list, and only noncitizens are able to claim the noncitizenship disqualification, which is one of the outcomes included in the 'other' category.

We could try to eliminate the obvious forms of dependence between sourcelist and outcome by restricting attention to only those summonses that were "problematic". That is, set aside the summonses in the 'other' category. Under the no-association model, calculate expected counts, subtract them from observed counts, then divide by the square root of the expected counts to get a new table of standardized differences (chi values). The observed value of 52.3 X^2 is again far too large to have come from an approximate $\chi^2(4)$ distribution, but the no-association model appears to be a much better explanation than before.‡

[chi]	undel	undel2	NS
DMV	-1.1	0.2	1.8
voter	2.5	0.1	-4.5
both	-2	-0.9	4.1

As a final offering, I restrict attention to just those summonses that were in one of the two undeliverable categories. The no-association model gives a table of fairly small standardized differences (chi values), with a sum of squares X^2 equal to 0.7. A random variable T with a $\chi^2(2)$ distribution has about a 70% chance of exceeding 0.7. It appears that a model of no-association between sourcelist and type of undeliverability problem is consistent with the observed counts.

[chi]	undel	undel2
DMV	-0.2	0.5
voter	0.2	-0.6
both	0.1	-0.2

‡ Chi-square tests with large numbers of counts can detect very small departures—which may not be of great practical significance—from reasonable models.