Categorical Data

Categorical Data

Problem: Instead of assuming numerical values, data may fall into categories. Such data is called ***categorical data***.

23.1. Example. Mars Corporation's ***M&Ms*** are produced in different colors: red, green, blue, brown, yellow and orange. If we pick a random M&M, it will randomly have one of these colors.

Approach: Each member of a population falls into one of k given categories with probability p_k, $0 < p_k < 1$, and

$$p_1 + p_2 + \cdots + p_k = 1.$$

Our goal is to make inferences on the values of these p_i.

Categorical Random Variables

We suppose that a random variable X is given, where X can take on the values $1, \ldots, k$ with respective probabilities $p_1 \ldots, p_k$ as above. We say that X is a ***categorical random variable***

A random sample of size n from X is collected and the results are expressed as a random vector

$$(X_1, X_2, \ldots, X_k) \qquad \text{with} \qquad X_1 + X_2 + \cdots + X_k = n.$$

For example, a packet containing $n = 14$ M&M's will yield a random vector $(X_{\text{red}}, X_{\text{green}}, \ldots, X_{\text{orange}})$.

When $k = 2$, then the distribution governing the probability of an item falling into category 1 ("success") or category 2 ("failure") is the binomial distribution. For $k > 2$, we need to develop a new distribution.

Multinomial Trials

23.2. Definition. A ***multinomial trial*** with parameters $p_1, \dots, p_k$ is a trial that can result in exactly one of k possible outcomes. The probability that outcome i will occur on a given trial is p_i, for $i = 1, \dots, k$.

23.3. Remark. It is clear from the definition that $0 \leq p_i \leq 1$, $i = 1, \dots, k$, and $p_1 + \cdots + p_k = 1$. To avoid unnecessary trivial cases, we assume $0 < p_i < 1$, $i = 1, \dots, k$.For $k = 2$, $p_1 = p$ and $p_2 = q = 1 - p$, we regain the classic Bernoulli trial.

A ***multinomial random variable*** now counts the number of times that outcome i occurs when a fixed number of n i.i.d. multinomial trials is performed. It therefore generalizes the binomial random variable.

The Multinomial Distribution

23.4. Definition. A random vector $((X_1, \dots , X_k), f_{X_1 X_2 \cdots X_k})$ with

$$(X_1, \dots , X_k) \colon S \to \Omega = \{0, 1, 2, \dots , n\}^k$$

and (joint) distribution function $f_{X_1 X_2 \cdots X_k} \colon \Omega \to \mathbb{R}$ given by

$$f_{X_1 X_2 \cdots X_k}(x_1, \dots , x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

$p_1, \dots , p_k \in (0, 1)$, $n \in \mathbb{N} \setminus \{0\}$ is said to have a ***multinomial distribution*** with parameters n and $p_1, \dots , p_k$.

23.5. Remark. Of course, it would be sufficient to consider a $k - 1$ dimensional random variable, as one of the X_i is wholly determined by the others. (The case $k = 3$ is handled by a bivariate, the case $k = 2$ by a simple random variable.) For reasons of symmetry it is, however, worth investing in the additional random variable.

Expectation and Variance of the Multinomial Distribution

23.6. Theorem. Let $((X_1, \dots, X_k), f_{X_1 X_2 \cdots X_k})$ be a multinomial random variable with parameters n and $p_1, \dots, p_k$.

(i) The (marginal) expectations of the individual random variables X_i are given by

$$\mathrm{E}[X_i] = np_i, \qquad\qquad i = 1, \dots, k.$$

(ii) $\mathrm{Var}[X_i] = np_i(1 - p_i)$, $i = 1, \dots, k$,

(iii) $\mathrm{Cov}[X_i, X_j] = -np_i p_j$, $1 \le i < j \le k$.

While results (i) and (ii) are easy to see, the proof of (iii) requires some work. Since we won't need that result, it is left to you.

The Pearson Statistic

Hypothesis testing and statistical analysis are based on the following result, which we will not prove:

23.7. Theorem. Let $((X_1, \dots, X_k), f_{X_1 X_2 \cdots X_k})$ be a multinomial random variable with parameters n and $p_1, \dots, p_k$. For large n the ***Pearson statistic***

$$\sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{23.1}$$

follows an approximate chi-squared distribution with $k-1$ degrees of freedom.

The Pearson Statistic

A good way to memorize this statistic is to see that the X_i are the "observed" frequencies and $np_i = \mathrm{E}[X_i]$ are the "expected" frequencies.

Writing $O_i := X_i$ and $E_i := \mathrm{E}[X_i]$, (23.1) becomes

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

23.8. Remark. The number of degrees of freedom in Theorem 23.7 is equal to the number of independent cells: given k cells and a total of n multinomial trials, the number of results in the first $k-1$ cells is random, while the number of results in the final cell is completely determined by these $k-1$ results.

One could say that there are $k-1$ independent cells, hence $k-1$ degrees of freedom.

Cochran's Rule

In the context of Theorem 23.7 we need to know how large n needs to be for the chi-squared distribution to be a good approximation to the true distribution of the statistic (23.1).

William G. Cochran (19091980). Statisticians in History, Amstat News (2016)

Cochran's Rule states that wee should require

$$\begin{aligned} \mathrm{E}[X_i] &= np_i \geq 1 && \text{for all } i = 1, \dots, k, \\ \mathrm{E}[X_i] &= np_i \geq 5 && \text{for 80\% of all } i = 1, \dots, k, \end{aligned}$$

Especially if the p_i are not known roughly beforehand, care needs to be taken to ensure that the sample size n is sufficiently large so that these criteria can apply.

Literature: Kroonenberg, P. M. and Verbeek, A. ***The Tale of Cochran's Rule: My Contingency Table has so Many Expected Values Smaller than 5, What Am I to Do?***, The American Statistician, 72:2 (2018)

Fisher Test for Multinomial Distribution

We can then develop a statistical test for the hypothesis that a set of data follows a given multinomial distribution: if data follows a given distribution, the number of observed values in each category will be close to the expected number and (23.1) will be small. Conversely, if (23.1) is large and the observed data deviates from the expected data significantly, then we have evidence that the data does not follow the presumed distribution.

This test will, by its nature, always be a Fisher test. Furthermore, it makes no sense to use terms such as "two-sided" or "ones-sided" for the test, since if some p_i are larger than their null values, then some other p_i will be smaller than their null values.

Test for Multinomial Distribution

23.9. Pearson's Chi-squared Goodness-of-Fit Test. Let $(X_1, \ldots, X_k)$ be a sample of size n from a categorical random variable with parameters $(p_1, \ldots, p_k)$ satisfying Cochran's Rule Let $(p_{1_0}, \ldots, p_{k_0})$ be a vector of null values. Then the test

$$H_0\colon p_i = p_{i_0}, \qquad\qquad i = 1, \ldots, k,$$

based on the statistic

$$X^2_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_{i_0})^2}{np_{i_0}}$$

is called an ***chi-squared goodness-of-fit test***.

We reject H_0 at significance level α if $X^2_{k-1} > \chi^2_{\alpha,k-1}$.

Multinomial Statistics

23.10. Example. A computer scientist has developed an algorithm for generating pseudorandom integers over the interval 0-9. He codes the algorithm and generates 1000 pseudorandom digits. The data is shown below:

i	0	1	2	3	4	5	6	7	8	9	n
O_i	94	93	112	101	104	95	100	99	94	108	1000
E_i	100	100	100	100	100	100	100	100	100	100	1000

We want to test whether these data conform to a discrete uniform distribution on $\Omega = \{0, 1, 2, \dots, 9\}$ at a level of significance $\alpha = 0.05$.

We test

$$H_0 : \text{The data follow a multinomial distribution}$$

$$\text{with parameters } (p_0, \dots, p_9) = \left(\frac{1}{10}, \dots, \frac{1}{10}\right).$$

Multinomial Statistics

The observed test statistic is

$$\sum_{i=0}^{9} \frac{(O_i - E_i)^2}{E_i} = \frac{(94-100)^2}{100} + \cdots + \frac{(108-100)^2}{100} = 3.72$$

This statistic follows a chi-squared distribution with $10 - 1 = 9$ degrees of freedom. Since $\chi^2_{0.05,9} = 16.92$, the P-value of the test is greater than 5%. There is not enough evidence to reject H_0.

We conclude that there is no evidence that the generated numbers are not random.

Goodness-of-Fit Test for a Discrete Distribution

The previous discussion centers on testing whether categorical data conforms to a ***completely determined*** distribution, i.e., we compare directly to a multinomial distribution. However, we can also use this method to see whether data conforms to an arbitrary discrete or continuous distribution.

Such a distribution will typically have one or more parameters, which we also estimate from the given data. We also need to divide our data into categories, so we can use our multinomial test. If we use our data to estimate parameters of the distribution, the statistic

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

will follow a chi-squared distribution with $k - 1 - m$ degrees of freedom, where m is the number of parameters that we estimate.

Goodness-of-Fit Test for a Discrete Distribution

23.11. Example. It is claimed that the number of defects in printed circuit boards follows a Poisson distribution with unknown parameter k. We want to determine if there is evidence that this claim is false.

A random sample of $n = 60$ printed boards has been collected and the following number of defects observed:

Number of Defects X	Observed Frequency
0	32
1	15
2	9
3	4

The parameter k (which is also the mean) of the assumed Poisson distribution is unknown and must be estimated from the data.

Goodness-of-Fit Test for a Discrete Distribution

From Example 12.5 we know that a maximum-likelihood estimator for k is the sample mean,

$$\widehat{k} = \overline{X} = \frac{1}{60}(32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3) = 0.75.$$

In order to apply the multinomial distribution, we first calculate

$$\begin{aligned}
P[X = 0] &= \frac{e^{-\widehat{k}}\widehat{k}^0}{0!} = 0.472 \\
P[X = 1] &= \frac{e^{-\widehat{k}}\widehat{k}^1}{1!} = 0.354 \\
P[X = 2] &= \frac{e^{-\widehat{k}}\widehat{k}^2}{2!} = 0.133 \\
P[X \geq 3] &= 1 - P[X = 0] - P[X = 1] - P[X = 2] = 0.041
\end{aligned}$$

Goodness-of-Fit Test for a Discrete Distribution

We can therefore replace the distribution of X with that of a categorical random variable with parameters

$$(p_0, p_1, p_2, p_3) = (0.472, 0.354, 0.133, 0.041).$$

We calculate the expected frequencies $E_i = np_i$ as follows:

Number of Defects X (Category i)	Expected Frequency E_i
0	$60 \cdot 0.472 = 28.32$
1	$60 \cdot 0.354 = 21.24$
2	$60 \cdot 0.133 = 7.98$
3	$60 \cdot 0.041 = 2.46$

We see that $E_3 < 5$ and since we have only four categories, this means that more than 1 in 5 categories have an expected frequency smaller than 5. Since (**??**) is not satisfied, we can not apply Pearson's test.

Goodness-of-Fit Test for a Discrete Distribution

The problem can be solved by combining the last two categories:

Category i	Exp. Frequency E_i	Obs. Frequency O_i
0	28.32	32
1	21.24	15
2	10.44	13

The test

H_0: the number of defects follows a Poisson distribution with parameter $k = 0.75$

is then equivalent to the test

H_0: the number of defects follows a multinomial distribution with parameters $(0.472, 0.354, 0.174)$

Goodness-of-Fit Test for a Discrete Distribution

For $N = 3$ categories, the statistic

$$X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

then follows a chi-squared distribution with $N - 1 - m = 3 - 1 - 1 = 1$ degree of freedom. We want to realize $\alpha = 0.05$ and therefore reject H_0 if $X^2 > \chi^2_{0.05,1} = 3.84$. Now

$$X^2 = \frac{(32 - 28.32)^2}{28.32} + \frac{(15 - 21.24)^2}{21.24} + \frac{(13 - 10.44)^2}{10.44} = 2.94 < 3.84,$$

so we are unable to reject H_0 at the 5% level of significance.

We can also test whether data fits a continuous distribution. In that case, the division of the data range into categories is essentially arbitrary, as illustrated in the following example.

Goodness-of-Fit Test for a Continuous Distribution

23.12. Example. A manufacturing engineer is testing a power supply used in a word processing work station. He wishes to determine whether output voltage is adequately described by a normal distribution. From a random sample of $n = 100$ units he obtains sample estimates of the mean and standard deviation $\overline{x} = 12.04\,\text{V}$ and $s = 0.08\,\text{V}$.

A common practice in constructing the class frequency distribution used in the chi-squared goodness-of-fit test is to choose the category boundaries so that the expected frequencies $E_i = np_i$ are equal for all categories. to use this method, we want to choose the category boundaries $a_0, \dots, a_k$ for the k categories so that all the probabilities

$$p_i = P[a_{i-1} \le X \le a_i] = \int_{a_{i-1}}^{a_i} f(x)\,dx$$

are equal.

Goodness-of-Fit Test for a Continuous Distribution

Suppose we decide to use $k = 8$ cells. For the standard normal distribution the intervals that divide the scale into 8 equally likely segments are

$$\begin{aligned} (a_0, a_1) &= (-\infty, -1.15), & [a_1, a_2) &= [-1.15, -0.675), \\ [a_2, a_3) &= [-0.675, -0.32), & [a_3, a_4) &= [-0.32, 0) \\ [a_4, a_5) &= [0, 0.32), & [a_5, a_6) &= [0.32, 0.675), \\ [a_6, a_7) &= [0.675, 1.15), & [a_7, a_8) &= [1.15, \infty) \end{aligned}$$

For the problem at hand, we need to transform these intervals to corresponding intervals for a normal distribution with mean $\overline{x}$ and standard deviation s. This is easily done by setting

$$a_i' := \overline{x} + s a_i, \qquad\qquad i = 0, \dots, 8.$$

For each interval, $p_i = 1/8$ so the expected cell frequencies are $E_i = np_i = 100/8 = 12.5$.

Goodness-of-Fit Test for a Continuous Distribution

The engineer observes 100 voltages as given below:

Category i	Exp. Frequency E_i	Obs. Frequency O_i
$x < 11.948$	12.5	10
$11.948 \leq x < 11.986$	12.5	14
$11.986 \leq x < 12.014$	12.5	12
$12.014 \leq x < 12.040$	12.5	13
$12.040 \leq x < 12.066$	12.5	11
$12.066 \leq x < 12.094$	12.5	12
$12.094 \leq x < 12.132$	12.5	14
$12.132 \leq x$	12.5	14
	100	100

We calculate

$$X^2 = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i} = 1.12$$

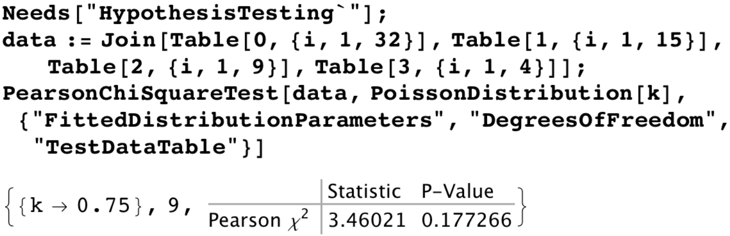
Goodness-of-Fit Test for a Continuous Distribution

We have $k = 8$ categories. Since two parameters in the normal distribution have been estimated, this statistic follows a chi-squared distribution with $k - 1 - m = 8 - 1 - 2 = 5$ degrees of freedom.

From Table IV we see that $\chi^2_{0.95,5} = 1.15$, so the P-value of the test is greater than 95%. We conclude that there is no reason to believe that output voltage is not normally distributed.

Goodness-of-Fit Tests with Mathematica

The goodness-of-fit test is also implemented in Mathematica. However, there is no control over the number of categories chosen; Mathematica will always choose about $2n^{2/5}$ categories and ignore the condition (**??**). Hence the test will not always be reliable. For instance, using the data from Example 23.11, we have



```
Needs["HypothesisTesting`"];
data := Join[Table[0, {i, 1, 32}], Table[1, {i, 1, 15}],
   Table[2, {i, 1, 9}], Table[3, {i, 1, 4}]];
PearsonChiSquareTest[data, PoissonDistribution[k],
 {"FittedDistributionParameters", "DegreesOfFreedom",
  "TestDataTable"}]
```

```
{{k → 0.75}, 9,              | Statistic  P–Value  }
                Pearson χ²   | 3.46021    0.177266 }
```

Mathematica uses $k = \lceil 2 \cdot 60^{2/5} \rceil = 11$ categories even though the data only occurs in four categories and Cochran's Rule isn't satisfied. The test statistic is the same as ours would have been had we used all four categories.

Independence of Categorizations

The multinomial distribution and the Pearson statistic are also very useful in another situation, best explained by an example:

23.13. Example. A researcher wants to study the relationship between nightly hours of sleep and academic performance of students. A test group of students fills out a questionnaire, giving their amount of sleep and their current GPA score. The test group can then be divided as follows

$$\{\text{test group}\} = \{< 6\text{h sleep}\} \cup \{6\text{-}9\text{h sleep}\} \cup \{> 9\text{h sleep}\},$$

$$\{\text{test group}\} = \{\text{low GPA}\} \cup \{\text{average GPA}\} \cup \{\text{high GPA}\}$$

If academic performance and nightly sleep are not related to each other, these categorizations will be independent, i.e., the likelihood of a student falling into any of the GPA categories will not depend on which sleep category the student is in.

Contingency Tables

The data from the test group can be summarized in a ***contingency table*** as follows:

	$<$ 6h sleep	6-9h sleep	$>$ 9h sleep
low GPA	n_{11}	n_{12}	n_{13}
average GPA	n_{21}	n_{22}	n_{23}
high GPA	n_{31}	n_{32}	n_{33}

Every member of the test group will count as 1 member of a specific ***cell*** (table entry) and the cells list the number of members with the corresponding properties. For example,

n_{23} = number of students with average GPA

and more than 9 hours of nightly sleep.

$r \times c$ Contingency Tables; Marginal Sums

In general, we will treat situations where the contingency table has r rows and c columns. We define the ***marginal row and column sums***

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij}, \qquad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}.$$

In our example,

	$<$ 6h sleep	6-9h sleep	$>$ 9h sleep	
low GPA	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
average GPA	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
high GPA	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Cell Probabilities and Independence

Suppose that

- p_{ij} is the probability of falling into the cell of the ith row and the jth column,
- $p_{i}.$ is the probability of falling anywhere in the ith row,
- $p_{.j}$ is the probability of falling anywhere in the jth column.

If the row and column categorizations are independent, then it should be the case that

$$H_0 \colon p_{ij} = p_{i}.p_{.j}. \tag{23.2}$$

We will therefore develop a test to determine whether there is statistical evidence that (23.2) is false.

Estimating the Probabilities

In principle, given n total sample elements, the number of elements in each of the $r \cdot c$ cells follows a multinomial distribution with $r \cdot c - 1$ independent probabilities p_{ij}. (Recall that the sum over all probabilities must equal 1, so there are one fewer than $r \cdot c$ independently selectable parameters.)

However, if we assume $p_{ij} = p_{i.} p_{.j}$, then the multinomial distribution only depends on the $r - 1 + c - 1$ parameters $p_{i.}$ and $p_{.j}$. We will exploit this for our test.

Natural estimates for the row and column probabilities are

$$\widehat{p_{i.}} = \frac{n_{i.}}{n}, \qquad \widehat{p_{.j}} = \frac{n_{.j}}{n}$$

so if (23.2) is assumed,

$$\widehat{p_{ij}} = \widehat{p_{i.}}\widehat{p_{.j}} = \frac{n_{i.} n_{.j}}{n^2}$$

Chi-Squared Test for Independence

Hence, if (23.2) is assumed, the expected number of elements in the (i, j)th cell is

$$E_{ij} = n \cdot \widehat{p_{ij}} = \frac{n_i.n._j}{n}$$

We can now compare the observed frequencies O_{ij} in the (i, j)th cell to the expected frequencies E_{ij}. We will again use the Pearson statistic

$$X^2_{(r-1)(c-1)} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

which follows a chi-squared distribution with

$$k - 1 - m = rc - 1 - (r + c - 2) = rc - r - c + 1 = (r - 1)(c - 1)$$

degrees of freedom. We reject H_0 if the value of $X^2_{(r-1)(c-1)}$ exceeds the critical value of the corresponding chi-squared distribution.

Testing for Independence

23.14. Example. A company has to choose among three pension plans. Management wishes to know whether the preference for plans is independent of job classification and wants to use $\alpha = 0.05$. The opinions of a random sample of 500 employees are shown below.

	Plan 1	Plan 2	Plan 3	Totals
Salaried Workers	160	140	40	$n_{1\cdot} = 340$
Hourly Workers	40	60	60	$n_{2\cdot} = 160$
Totals	$n_{\cdot 1} = 200$	$n_{\cdot 2} = 200$	$n_{\cdot 3} = 100$	$n = 500$

We want to test

H_0: there is no dependence between job classification and plan preference

Testing for Independence

We calculate the expected frequencies assuming that H_0 is true:

$$E_{11} = \frac{200 \cdot 340}{500}, \qquad E_{12} = \frac{200 \cdot 340}{500}, \qquad E_{13} = \frac{100 \cdot 340}{500}$$
$$E_{21} = \frac{200 \cdot 160}{500}, \qquad E_{22} = \frac{200 \cdot 160}{500}, \qquad E_{23} = \frac{100 \cdot 160}{500}$$

It is now a simple matter to calculate the value of the statistic

$$X^2_{(2-1)(3-1)} = X^2_2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 49.63.$$

Since the statistic follows a chi-squared distribution with 2 degrees of freedom and we want $\alpha = 0.05$, we compare this value with $\chi^2_{0.05,2} = 5.99$. As $49.63 > 5.99$, we may reject H_0. There is evidence that the pension plan preference is not independent of job classification.

Comparing Proportions

Finally, we note that a very similar, though subtly different approach can be taken when comparing multiple proportions. Suppose that we would like to compare the proportions students with little, average or much nightly sleep among the JI ECE, JI ME, SJTU EE and SJTU ME majors. We choose to randomly select (based on student IDs) $n_{1\cdot}$, $n_{2\cdot}$, $n_{3\cdot}$ and $n_{4\cdot}$ students, respectively, from each of these majors.

We obtain the following contingency table:

	$<$ 6h sleep	6-9h sleep	$>$ 9h sleep	
JI ECE	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$ (fixed)
JI ME	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$ (fixed)
SJTU EE	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$ (fixed)
SJTU ME	n_{41}	n_{42}	n_{43}	$n_{4\cdot}$ (fixed)
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n (fixed)

Comparing Proportions

We again suppose that the number of objects in each cell is governed by the multinomial distribution. However, since the row totals are now fixed, only the number of objects in the first $c-1$ columns can be independently chosen, so we have a total of $r \cdot (c-1)$ independent cells.

It is useful to rewrite the above table in terms of proportions:

	$<$ 6h sleep	6-9h sleep	$>$ 9h sleep	
JI ECE	p_{11}	p_{12}	p_{13}	$p_{1.} = 1$ (fixed)
JI ME	p_{21}	p_{22}	p_{23}	$p_{2.} = 1$ (fixed)
SJTU EE	p_{31}	p_{32}	p_{33}	$p_{3.} = 1$ (fixed)
SJTU ME	p_{41}	p_{42}	p_{43}	$p_{4.} = 1$ (fixed)

We test

$$H_0 : \begin{cases} p_{11} = p_{21} = p_{31} = p_{41}, \\ p_{12} = p_{22} = p_{32} = p_{42}, \\ p_{13} = p_{23} = p_{33} = p_{43}. \end{cases}$$

Comparing Proportions

Supposing that H_0 is true, we have the common proportions

$$p_j := p_{1j} = p_{2j} = p_{3j} = p_{4j}, \qquad j = 1, 2, 3,$$

where p_j is also equal to the proportion of all objects following into the jth column. An estimate for p_j is

$$\widehat{p_j} = \frac{n_{.j}}{n}, \qquad j = 1, 2, 3. \tag{23.3}$$

and $\widehat{p_j}$ also serves as an estimator for all of the p_{ij}, $i = 1, \dots, 4$. If H_0 is true, the expected frequency in each cell is given by

$$E_{ij} = n_{i.}\widehat{p_{ij}} = \frac{n_{i.}n_{.j}}{n}$$

and we can again apply the Pearson chi-squared test.

Comparing Proportions

For the general case of r rows and c columns, the test

$$H_0 \colon p_{1j} = p_{2j} = \cdots = p_{rj}, \qquad\qquad j = 1, \ldots, c.$$

is called a ***test for homogeneity***. When using the Pearson statistic, note that the degrees of freedom are

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

where $r(c-1)$ is the number of independent cells and $c-1$ is the number of independent parameters $\widehat{p}_j$ that are estimated in (23.3).

We note that the tests for independence and for homogeneity appear absolutely the same in practice. That is not very surprising, since the null hypotheses are logically equivalent (see the example on the next slide).

Comparing Proportions

23.15. Example. A study is to be conducted to consider the association between the sulfur dioxide (SO_2) level in air and the mean number of chloroplasts per leaf cell of trees in the area. Three regions are selected for study. One is known to have a high SO_2 concentration, one to have a normal level of SO_2, and the third to have a low SO_2 level.

Twenty trees are to be randomly selected from within each area, and the mean number of chloroplasts per leaf cell is to be determined for each tree. On this basis each tree will be classified as having a low, normal or high chloroplast count.

We want to test

H_0: the proportion of trees with low, normal or high chloroplast count is the same at low, normal or high SO_2 levels

Of course, this is the same as testing

H_0: the chloroplast count does not depend on the SO_2 level

Comparing Proportions

The following data is obtained:

SO_2 / Chloroplast	High	Normal	Low	Totals
High	3_{5}	$4_{8.33}$	$13_{6.67}$	$n_{1.} = 20$
Normal	5_{5}	$10_{8.33}$	$5_{6.67}$	$n_{2.} = 20$
Low	7_{5}	$11_{8.33}$	$2_{6.67}$	$n_{3.} = 20$
Totals	$n_{.1} = 15$	$n_{.2} = 25$	$n_{.3} = 20$	$n = 60$

The expected frequencies are noted within the above table. The statistic follows a chi-squared distribution with 4 degrees of freedom and gives

$$X_4^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 14.74.$$

Since $\chi^2_{0.01,4} = 13.3$ and $\chi^2_{0.005,4} = 14.9$, we can reject H_0 with $0.005 < P < 0.01$. Thus there is strong evidence of an association between chloroplast levels and SO_2 concentration.