Analysis of count data.

Categorical response data are also called count data, comparing different counts for their categorical values. Here we examine the proportion or the number of observations for categorical response as opposed to the average value of some quantitative variable. A statistical method is similar to the one that occurs in analysis of variance, but utilize χ^2-distribution in place of F-distribution. We also deal with two-way tables, the rows of different groups and columns of categories, called **contingency tables**. They are summarized in a table with rows and columns representing the various categories where each cell in the table contains the number of observations for the given combination of categorical responses.

Goodness of fit. Assume that we have classified n individuals into k groups.

Group	Count	Expected
1	Obs_1	np_{01}
$\vdots$	$\vdots$	$\vdots$
k	Obs_k	np_{0k}
Total	n	n

We will compare the proportions of different groups with pre-specified probabilities

$$p_{01}, \ldots, p_{0k}$$

under the null hypothesis H_0. Then we can calculate the expected number of observations for each category. Thus, for group 1 we would expect np_{01} observations if the null hypothesis is true, and so forth.

We wish to test a null hypothesis H_0 that completely specifies the probabilities of the k groups.

$$H_0 : p_1 = p_{01}, \ldots, p_k = p_{0k}$$

Here p_i represents the population proportion that a randomly chosen individual will belong to group i. A statistical test is designed to describe how well the hypothesis H_0 fits a set of observations; thus, it is called "goodness of fit." In this case we are mostly interested in the failure to reject the null hypothesis. If we wish to test the null hypothesis, then we should compare the number Obs_i of observations with the expected number

$$\text{Exp}_i = np_{0i}$$

for each group $i = 1, \ldots, k$.

Chi-square test. To test a hypothesis for tabular data, we can use the chi-square statistic

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(\text{Obs}_i - \text{Exp}_i)^2}{\text{Exp}_i} = \sum_{i=1}^{k} \frac{(\text{Obs}_i - np_{0i})^2}{np_{0i}}$$

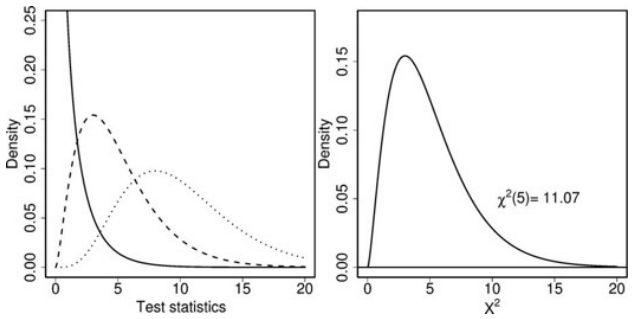
Here the summation is over all possible categories. The numerator in the i-th term contains the squared difference between the observed and expected number of observations in group i, so it must be small if the null hypothesis H_0 in goodness of fit is true, and it becomes large if there is a large discrepancy between the two (that is, if the null hypothesis is false).

How small χ^2_{obs} can be if H_0 is true?. It turns out that the chi-square test statistic χ^2_{obs} approximately follows a χ^2-distribution with the number

$$df = (\text{number of groups}) - 1$$

of degrees of freedom if H_0 is true. The critical value for χ^2_{obs} at significance level α is the value such that the right-most area under the χ^2-distribution is exactly α. Then the p-value is the probability (given that the null hypothesis is true) of observing a value that is more extreme (i.e., further away from zero) than what we have observed.

χ^2-distribution and critical value. The below left panel shows the density for the χ^2-distribution with $df = 1$ (solid), $df = 5$ (dashed), as well as $df = 10$ (dotted). The below right panel illustrates the density for the χ^2-distribution with $df = 5$. The critical value $\chi^2(5)$ is 11.07 at significance level $\alpha = 0.05$. The chi-square test statistic χ^2_{obs} is considered small if $\chi^2_{obs} < 11.07$.



Example: Mendelian inheritance. According to Mendel's theory, pea plant phenotypes should appear in the ratio $9:3:3:1$. Thus, Mendel's model specifies the probabilities, 0.5625, 0.1875, 0.1875, and 0.0625, of the four groups:

$$H_0 : p_{ry} = \frac{9}{16}, \quad p_{rg} = \frac{3}{16}, \quad p_{wy} = \frac{3}{16}, \quad p_{wg} = \frac{1}{16},$$

where "r", "w", "y", and "g" in the subscripts denote "round", "wrinkled", "yellow", and "green", respectively. Gregor Mendel published his results from his experiments on simultaneous inheritance of pea plant phenotypes (Mendel, 1866). In this study we wish to see that data do not contradict the law of Mendelian inheritance.

In the experiment, he examined 556 pea plants, and if we believe Mendel's model to be true (i.e., we look at the distribution under H0), then we would expect $(556)(9/16) = 312.75$ plants of round and yellow peas. If we do the same calculation for every group, we can summarize the observed data and the expected data in a table.

Class	Observed	Expected
Round, yellow	315	312.75
Round, green	108	104.25
Wrinkled, yellow	101	104.25
Wrinkled, green	32	34.75
Total	556	556

The χ^2-test statistic becomes

$$\begin{aligned}\chi^2_{obs} &= \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(101-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} \\ &= 0.470\end{aligned}$$

We compare this value to a χ^2-distribution with $df = 3$ because we have 4 groups, which results in 3 degrees of freedom. The critical value $\chi^2(3)$ at a significance level of $\alpha = 0.05$ is 7.81 (see, for example, the statistical table). Since $\chi^2_{obs} = 0.470$ is less than 7.81, we fail to reject the null hypothesis. In other words, these data do not contradict Mendel's hypothesis. Alternatively we can look up the exact p-value from computer package and we get that the p-value is 0.9254.

Contingency table. In a typical study observations are classified or tabulated according to two categorical variables and not just a single categorical variable. In the following table we have two categorical variables each with two categories.

	Response 1	Response 2	Total
Sample 1	y_{11}	y_{12}	$n_1 = y_{11} + y_{12}$
Sample 2	y_{21}	y_{22}	$n_2 = y_{21} + y_{22}$
Total	$c_1 = y_{11} + y_{21}$	$c_2 = y_{12} + y_{22}$	$n = n_1 + n_2 = c_1 + c_2$

It can be extended to more general $r \times k$ contingency tables, where we have two categorical variables; one with r groups and the other with k categories.

Test for homogeneity. For the first sample, we can look at the probability of observing response 1. If p_{11} is the probability of observing response 1 for an individual from sample 1, then we have

$$\hat{p}_{11} = \frac{y_{11}}{n_1}$$

We can estimate the corresponding probability for sample 2

$$\hat{p}_{21} = \frac{y_{21}}{n_2}$$

Then it is natural to be interested in the test for homogeneity

$$H_0 : p_{11} = p_{21}$$

Expected number of observations. If we assume that the probability of observing response 1 is the same for both samples, then we can use the data from both samples to estimate that probability. Our combined estimate of the probability of observing response 1, p, would be

$$\hat{p} = \frac{c_1}{n} = \frac{y_{11} + y_{21}}{n}$$

There are n_1 observations in sample 1. Thus, we would expect

$$\text{Exp}_{11} = n_1\hat{p} = \frac{n_1 c_1}{n}$$

From sample 1 to result in response 2, we would expect

$$\text{Exp}_{12} = n_1(1-\hat{p}) = \frac{n_1 c_2}{n}$$

Chi-square test for homogeneity. In general, the expected number of observations in cell (i, j) for a contingency table is given by

$$\text{Exp}_{ij} = \frac{(\text{Total for row } i)(\text{Total for column } j)}{n} = \frac{n_i c_j}{n}$$

Now that we have the expected number of observations for each cell under H_0, we can carry out a test of H_0 using the same type of chi-square test statistic

$$\chi^2_{obs} = \sum_{i,j} \frac{(\text{Obs}_{ij} - \text{Exp}_{ij})^2}{\text{Exp}_{ij}} = \sum_{i,j} \frac{(y_{ij} - (n_i c_j/n))^2}{(n_i c_j/n)}$$

to which χ^2-distribution with one degree of freedom should be compared.

Example: Avadex Study. The increasing risk of the fungicide Avadex on pulmonary cancer in mice was studied. Sixteen male mice were continuously fed small concentrations of Avadex (the treatment population), while 79 male mice were given the usual diet (the control population). We wish to see whether mice from the two groups develop tumors equally likely or not. After 85 weeks, all animals were examined for tumors, with the results shown below:

	Tumor present	No tumor	Total
Treatment group	4	12	16
Control group	5	74	79
Total	9	86	95

Under the hypothesis $H_0 : p_t = p_c$, the probability of developing a tumor is not affected by population (treatment group). The probability of observing a tumor is given by

$$\hat{p} = \frac{4+5}{16+79} = 0.0947$$

Then we would expect $(0.0947)(16) = 1.52$ of the mice from the treatment group and $(0.0947)(79) = 7.48$ from the control group to develop tumors. We can calculate all four expected values:

	Tumor present	No tumor	Total
Treatment group	1.52	14.48	16
Control group	7.48	71.52	79
Total	9	86	95

When testing H_0, we calculate the test statistic

$$\chi^2_{obs} = \frac{(4-1.52)^2}{1.52} + \frac{(12-14.48)^2}{14.48} + \frac{(5-7.48)^2}{7.48} + \frac{(74-71.52)^2}{71.52} = 5.4083$$

By comparing to a χ^2-distribution, we obtain the p-value of 0.020, and reject the null hypothesis. Thus, we find that the proportions of mice developing tumors are different in the two populations. Furthermore, we found

$$\hat{p}_t = \frac{4}{16} = 0.25; \quad \hat{p}_c = \frac{5}{79} = 0.0633.$$

Therefore, we are able to conclude that Avadex appears to increase the cancer rate in mice.

Comparison of two proportions. If we let p_t and p_c denote the probabilities that a mouse from the treatment population and the control population, respectively, develops a tumor, then we can estimate those probabilities by

$$\hat{p}_t = \frac{4}{16} = 0.25; \quad \hat{p}_c = \frac{5}{79} = 0.0633.$$

The 95% confidence interval for $p_t - p_c$ becomes $(-0.0321, 0.4056)$ Thus, regarding the null hypothesis

$$H_0 : p_t = p_c$$

we fail to reject it. But this finding is not consistent with the conclusion by chi-square test.

Conclusion for Avadex study. We conclude that the result is significant, and that Avadex appears to increase the cancer rate in mice. However, this conclusion from p-value is different from the finding from the 95% confidence interval for $p_t - p_c$. We get different results from the two methods because the hypothesis test and the confidence interval are not consistent for small sample sizes (which requires rather technical discussions). However, we found that zero was barely in the 95% confidence interval and got a p-value of 0.020 for the chi-square test statistic. These two results are really not that different; thus, you should report both results though they are inconsistent. When we get a p-value close to 0.05 or get a confidence interval where zero is barely inside or outside, we should be careful with the conclusions.