Categorical data

In 1912 the Titanic sank and more than 1,500 of the 2,229 people on board died. Did the chances of survival depend on the ticket class?

	First	Second	Third	Crew
Survived	202	118	178	215
Died	123	167	528	698

This is an example of **categorical data**: the data are counts for a fixed number of categories. Here the data are tabulated in a 2×4 **table.**

Such a table is called a **contingency table** because it shows the survival counts for each category of ticket class, i.e. contingent on ticket class.

We will look at three problems involving categorical data: testing goodness-of-fit, testing homogeneity, and testing independence.

Testing goodness-of-fit

In 2008 the producer of M&Ms stopped publishing the color distribution of M&Ms. Here are the last percentages published in 2008:

Blue	Orange	Green	Yellow	Red	Brown
24%	20%	16%	14%	13%	13%

Recently, students in my class inspected the contents of one bag of milk chocolate M&Ms and got the following counts:

Blue	Orange	Green	Yellow	Red	Brown
85	79	56	64	58	68

Are these counts consistent with the last published percentages? Or is there sufficient evidence to claim that the color distribution is now different?

This question requires a **test of goodness-of-fit** for the six categories.

Testing goodness-of-fit

H_0 = "nothing is extraordinary is going on" = the color distribution is given by the table from 2008.

H_A: the color distribution is different from the 2008 table

The idea is to compare the observed counts to the numbers one would expect if H_0 is true.

To compute the expected counts, note that we have a total of 410 M&Ms. So on H_0 we expect $24\% \times 410 = 98.4$ blue M&Ms. After doing this computation for the other colors, we get a table of observed and expected counts:

	Blue	Orange	Green	Yellow	Red	Brown
Observed	85	79	56	64	58	68
Expected	98.4	82	65.6	57.4	53.3	53.3

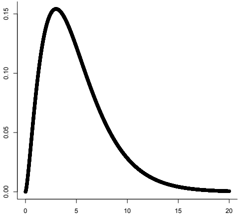
Testing goodness-of-fit

Next we combine the differences between observed and expected counts across all six categories:

$$\chi^2 = \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$= \frac{(85 - 98.4)^2}{98.4} + \frac{(79 - 82)^2}{82} + \ldots + \frac{(68 - 53.3)^2}{53.3} = 8.57$$

Large values of the **chi-square statistic** χ^2 are evidence against H_0. The p-value is the *right* tail of the χ^2 distribution with degrees of freedom = number of categories$-1 = 5$.



Testing goodness-of-fit

This χ^2-test is applicable if the observed data are counts that were obtained by drawing the M&Ms independently from a population of M&Ms - it is plausible that filling the M&Ms into bags follows approximately this process.

Blue M&Ms were not introduced until 1995. Initially some people suspected that there were fewer blue M&Ms in a bag than advertised. Note that testing the proportion of one category can be done with the z-test. The chi-square test provides an extension of the z-test to testing several categories.

Testing homogeneity

Are the color distributions the same for milk chocolate M&Ms, peanut M&Ms, and caramel M&Ms?

The χ^2-**test of homogeneity** tests the null hypothesis that the distribution of a categorical variable (color) is the same for several populations (milk, peanut, caramel).

It assumes that the samples are drawn independently within and across the populations.

To see how the test works, let's look at the survival data for the Titanic:

	First	Second	Third	Crew
Survived	202	118	178	215
Died	123	167	528	698

Testing homogeneity

Note that in this case we are not sampling from a population: The data are not a random sample of the people on board, rather the data represent the whole population.

So in this case the chance process resulting in survival or death is not in the sampling, but the result of the random events occuring when looking for a way out of the ship, getting into a life boat or into the water, being rescued out of the water in time etc.

Then the 325 observations about first class passengers represent 325 independent draws from a probability histogram that gives a certain chance for survival. The 285 observations about second class passengers are drawn from the probability histogram for second class passengers, which may be different. The null hypothesis says that the probability of survival is the same for all four probability histograms.

Testing homogeneity

	First	Second	Third	Crew
Survived	202	118	178	215
Died	123	167	528	698

In order to compute the table of expected entries, note that under H_0 the probability of survival is the same for all four populations. So we can estimate this probability by pooling the data: in total there were 713 survivors among the 2,229 people on board, so we estimate the probability of survival as $\frac{713}{2229} = 32\%$.

So the expected number of surviving first class passengers is $32\% \times 325 = 104.0$. Doing this computation for the other seven categories gives the table of expected values:

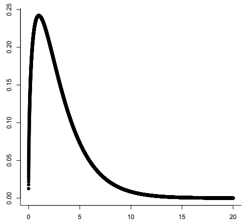
	First	Second	Third	Crew
Survived	104.0	91.2	225.8	292.1
Died	221.0	193.8	480.1	620.8

Testing homogeneity

Now we can compute the chi-square statistic over all 8 cells in the table:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$
$$= \frac{(202-104)^2}{104} + \ldots = 192.2$$

In this case the degrees of freedom
$= (\text{no. of columns} - 1) \times (\text{no. of rows} - 1) =$
$(4-1) \times (2-1) = 3.$



Testing independence

Is gender (male/female) related to voting preference (liberal/conservative)?

Now we have two categorical variables: gender and voting preference. The null hypothesis is that the two variables are independent. The alternative hypothesis is that there is some kind of association.

The chi-square test can be used to test this null hypothesis:
After sampling from the population the counts are recorded in a 2×2 table. The χ^2-statistic and p-value are computed exactly as in the case of testing homogeneity.

Comparing these two uses of the χ^2-test:

	Sample	Research question
χ^2-test of homogeneity	single categorical variable measured on several samples	Are the groups homogeneous (have the same distribution of the categorical variable)?
χ^2-test of independence	two categorical variables measured on one sample	Are the two categorical variables independent?