Chi-square and Fisher Exact tests

Contingency Tables

Contingency tables are just tables of counts of cross-tabulations such as those we saw for barplots and mosaicplots.

	0	1–150	151–300	>300
Married	652	1537	598	242
Prev.married	36	46	38	21
Single	218	327	106	67

For a goodness-of-fit test the null hypothesis is that the row and column categories are independent. Using that we can estimate the expected frequencies in each cell as $E_{ij} = np_ip_j$ where n is the total count, p_i is the proportion in row i and p_j is the proportion in row j. Then the chi-quared statistic is, for an $r \times c$ table,

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

at least approximately. In the example $P \approx 0$.

Validity of Chi-squared Test

The derivation of the null-hypothesis distribution of the chi-squared test relies on the counts being fairly large, and more detailed work suggest all counts should be at least 5: sometimes categories can be combined to achieve this.

What about this example, on whether parasites were present in a sample of crabs collected in Yaquina Bay, Oregon?

	Yes	No
Red Crab	5	312
Dungeness Crab	0	503

We need a test for which there is an 'exact' distribution, or we could simulate the null-hypothesis distribution of the chi-squared test.

Fisher's Exact Test

Fisher's exact test is of the difference in sample proportions between the rows of a 2×2 table, or equivalently of the odds ratio. The null-hypothesis distribution counts the number of tables which can be formed *with the same marginal counts* that have a more extreme statistic. Since we can only vary one cell of the table, let it be the upper left one: it has a hypergeometric distribution under the null hypothesis.

In our example the odds-ratio (of Dungeness *vs* Red Crabs) is estimated as being in $(0, 0.68)$, and the $P \approx 0.8\%$.

There are extensions to $r \times c$ tables, but enumerating the tables is a serious computational task even today.