Categorical Data Analysis: Basics

In our lab, we sometimes collect categorial data. Categorical variables are those which can be assigned to different groups, but the groups have no natural order to them. For example, hair color (brown, red, blonde, ….).

This document describes how to display and evaluate the relationships between categorical variables with examples relevant to the kinds of data we generate.

Example of a typical dataset

Consider an experiment in which we collect embryos from control or wild-type clutches and then assess the laterality of heart jogging. We score the embryos as having either left, middle or right heart jogging.

Genotype	**Heart jogging laterality**		
	Left	**Middle**	**Right**
WT	222	14	7
***pkd1l1* mutant**	30	92	29
***cfap298* mutant**	19	27	15

We must plot these data and perform a statistical test to assess potential differences between *pkd1l1* and *cfap298* mutants.

Hypotheses and the chi-square test of independence

To evaluate categorical data, we can use the chi-square test of independence. Before starting this test, consider the null hypothesis:

> Null hypothesis: There are no relationships between the variables.

Another way to think about the null hypothesis is that *knowing the value one variable in no way helps you predict the value of the other variable*. By contrast, support for the alternative hypothesis would suggest that knowing the value of one variable *would* help you predict the other. That is to say:

> Alternative hypothesis: There are relationships between the variables.

The chi-square test of independence compares the experimentally observed distribution to the distribution you would expect if the null hypothesis were true.

You can perform the chi-square test using Excel, R (using `chisq.test`), or statistics websites. You must use numerical values for each cell of the contingency table, not percentages.

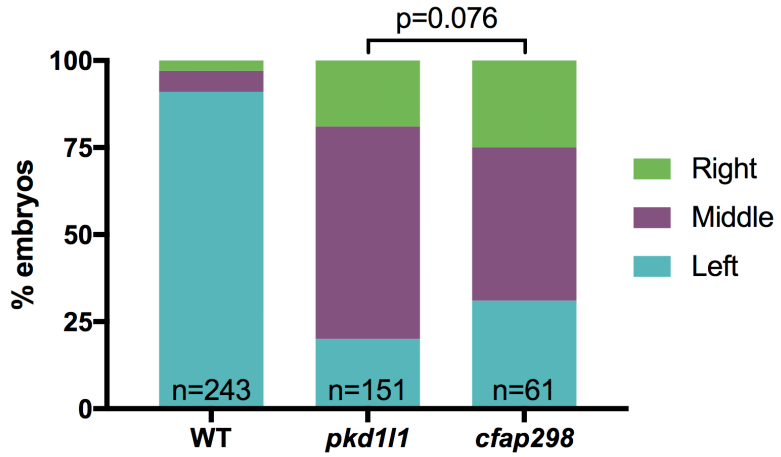
If the resulting p-value is less than your significance level (α), then there is reason to infer that a relationship exists between the variables. If, by contrast, the p-value is greater than α, then you

cannot reject the null hypothesis i.e. there is no reason to infer your observed distribution is any different to the distribution expected assuming there is no relationship between the variables.

In the example above, using the chi-square test on a 2x3 contingency table to compare *pkd1l1* and *cfap298* gives a p=0.076. Thus, we cannot reject the null hypothesis since the p value is greater than the significance level (0.05).

Plotting categorical data

The clearest way to plot data of this sort is with a stacked bar chart in which percentages are given for each category. For instance, 222 of 243 WT embryos exhibited left jogging. Thus, we plot (222/243) = 91.4%. Performing this calculation then plotting the data gives:



This bar chart communicates the three samples (WT, *pkd1l1* and *cfap298*), the three categories (left, middle, right), the total n of each sample and a p value for a statistical test between two of the samples.

Some critical information should be included in the figure legend: 1) the name of the statistical test used and 2) the number of replicates that, when combined, gave rise to the total n shown in the plot. In this case, replicates are likely to be clutches of embryos.

Combining data from biological replicates

We should always repeat our experiments. Imagine the data given above for *cfap298* actually came from three clutches:

Genotype	**Heart jogging laterality**		
	Left	**Middle**	**Right**
Clutch 1	8	8	6
Clutch 2	4	9	2

Clutch 3	7	10	7

It is a common mistake to calculate percentages for each clutch and then average them. For left jogging, this would give (36.7 + 26.7 + 29.2)/3 = 30.9%. Instead, when done correctly, we find that the percentage jogging left is 31.1%. We should not average percentages because that does not take into account the differences in the number of sampled embryos between clutches.

As such, plots showing individual clutches with means and standard errors or standard deviations are not typically appropriate for these kinds of data.