Day 25 - Goodness of fit

Learning outcomes

In this section, we explore a special kind of Chi-square test: the goodness of fit test.

Goodness of fit

Goodness of fit tests

When doing statistics, we often find ourselves wondering whether observed count data are consistent with the counts you would expect under some hypothetical distribution. In the most basic case, we can compare observed counts to some expected distribution. In higher-level statistics, we often want to determine whether our data could reasonably come from a probability distribution. We can use the Chi-square test for this purpose, using the same test statistic as before:

$$X^2 = \sum_i \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2$$

Where E_i is the expected count according to the hypothetical probability distribution. The null and alternative hypotheses are:

H_0 : the population distribution across the response categories matches the claim

H_a : the population distribution across the response categories does not match the claim

Under the null hypothesis,

$$X^2 \sim \chi^2_{df}$$

where df is the number of categories minus one. To run the goodness of fit test in R, we use the following code:

```
chisq.test(
  x = data,
  p = c(null hypothesized proportions)
)
```

It is important to recognize that a goodness of fit test is used when **you only collect a single categorical variable**. Otherwise, the test functions identically to the test of independence.

Examples:

- In genetics: the Mendelian ratio (9:3:3:1) describes the ratio of genetic phenotypes resulting from crossing of two independent traits
- In Earth science: we can model whether earthquake arrival times follow a hypothesized distribution by binning the arrival times and comparing to the expected counts
- We could use the goodness of fit test to determine whether birthdays are evenly distributed across days of the week

Skittles

According to the Mars website[1], the five original skittles colors are equally distributed during production (i.e. the probability of obtaining any single color is 1/5). Do the data support this claim? Let's find out!

1. Before sampling the bag of skittles, write out the null and alternative hypothesis for this test in words.

2. Now collect your data. Use the Dixie cup to sample the bag of skittles and count the total number of each color that is present. Create a vector that represents your observed counts.

3. What is the expected count for each color?

4. Is the independence assumption reasonable for these data? Why or why not?

5. Is the large enough sample assumption met for these data? Why or why not?

6. Use R to calculate the test statistic.

7. What distribution does the test statistic follow under the null hypothesis?

8. Use R to plot the distribution of the X^2 statistic under the null hypothesis. Plot the test statistic as a line on the distribution. *Hint:* see `?dchisq`

9. Use `R` to find the p-value.

10. Interpret the p-value as a conditional probability in the context of the problem.

11. Interpret the p-value as strength of evidence in the context of the problem.

12. Provide a mosaic plot of the standardized residuals, and discuss what you see.

[1] Link to the website.