Week Three: Video Lecture

This Week: Contingency Tables (Part 1)

Tuesday:

- Watch Week 4 videos and submit video notes
- Week 4 activity

Thursday:

- Lab 3

Primer for Contingency Tables

- We've previously explored modeling outcomes of a single categorical variable- both binary and multicategory.
- Contingency tables can be used for comparing outcomes across two or more categorical variables. (Ex. favorite ice cream shop and night owl / early bird)

- With contingency tables there is generally interest in whether two variables are related.

Contingency Table Overview

Let's assume we are now interested in comparing MSU students' preference in Genuine vs. Sweet Peaks ice cream based on whether the students identify as night owls or early birds.

A contingency table can be used to visualize outcomes across the combined categories.

Table 1: 2 by 2 contingency table

	NO	EB	
SP	$n_{sp,no}$	$n_{sp,eb}$	n_{sp}
Gen	$n_{g,no}$	$n_{g,eb}$	n_g
	n_{no}	n_{eb}	n

where $n_{i,j}$ denotes the number of observed counts with the i^{th} *and* j^{th} category, n_i is the total number of counts for the i^{th} category (ignoring the other variable), and n is the total sample size.

- If there is no relationship between ice cream preference and circadian rhythms, what would we expect for the $n_{sp,no}$, $n_{sp,eb}$, $n_{g,no}$, and $n_{g,eb}$ values?

The values in the contingency tables can be described as being generated from joint, marginal, and conditional probabilities.

Table 2: 2 by 2 contingency table with probabilities

	NO	EB	
SP	$\pi_{sp,no}$	$\pi_{sp,eb}$	π_{sp}
Gen	$\pi_{g,no}$	$\pi_{g,eb}$	π_{g}
	π_{no}	π_{eb}	1

- joint probabilities relate to multiple outcomes both occuring, so $\pi_{sp,no}$ requires sweet peaks **and** night owl.

- marginal probabilities marginalize out (integration or summation) one variable and are only concerned with a singe outcome pi_{sp}

- conditional probabilities are not explicitly listed in the table above, but allow us to answer questions like, given a student is a night owl, what is the probability they prefer sweet peaks. Mathematically, this can be represented as $\pi_{sp|no} = \frac{\pi_{sp,no}}{pi_{no}}$.

- statistical independence implies that knowing one category does not change the expected value for another category.

Inferences from the contingency tables will be based on the $n_{i,j}$ values; however, we do have an underlying assumption that the count values can be generated from binomial or multinomial distributions with the underlying probability values (π).

For example, let's assume there are 200 students in the study. We can state the 4 marginal probabilities and generate a contingency table.

- $\pi_{sp,no} = .4$
- $\pi_{sp,eb} = .2$
- $\pi_{g,no} = .2$
- $\pi_{g,eb} = .2$

```
set.seed(09042025)
pi_vals <- c(.4, .2, .2, .2)
n_values <- rmultinom(1, 200, pi_vals)

output_table <- tibble(group= c('SP', 'Gen', ''),
                       NO = c(n_values[c(1,3)], sum(n_values[c(1,3)])),
                       EB = c(n_values[c(2,4)], sum(n_values[c(2,4)])),
                       marg = c(sum(n_values[1:2]), sum(n_values[3:4]), sum(n_values)))

output_table |> kable(col.names = c('', 'NO', 'EB', ''))
```

	NO	EB	
SP	76	47	123
Gen	42	35	77
	118	82	200

Given the data (and known probabilities), do we believe there is a relationship between circadian rhythm and ice cream preference?