Week Four

Last Week

- Bayesian Inference for Binomial & Multinomial Distributions
- Bayesian Inference vs. Maximum Likelihood Estimation
- Contingency Table Primer Video
  - Contingency Table Overview: multiple categorical variables
  - joint, marginal, and conditional probabilities

This Week: Contingency Tables

Today:

- Activity
  - Comparing Proportions: Relative Risk & Odds Ratios
  - Chi-Squared Tests for Independence
- Thursday: Lab

Next Week: More Contingency Tables

Comparing Proportions

Consider two binary variables. As an example, participants are given ice cream from either sweet peaks or genuine and asked whether it was delicious (5 stars). Note: this could be displayed in a 2 X 2 contingency table.

	Yes	NO	
SP	$\pi_{sp,y}$	$\pi_{sp,n}$	π_{sp}
Gen	$\pi_{g,y}$	$\pi_{sp,n}$	π_g

We may be interested in comparing the proportion of respondents that rated ice cream as delicious given the ice cream shop that made it. Note these are conditional probabilities: $\pi_{y|sp} = \frac{\pi_{sp,y}}{\pi_{sp}}$.

Here we can directly compare $\pi_{y|sp}$ and $\pi_{y|g}$. Recall for binomial data ($Y \sim Binomial(n,p)$) that:

- E[Y]

- Var[Y]

We can use the MLE estimates of $\pi_{y|sp}$, $\pi_{y|g}$, and $\pi_{y|sp} - \pi_{y|g}$

- MLE of $\pi_{y|sp}$: $p_{y|sp}$

- MLE of $\pi_{y|g}$: $p_{y|g}$

- MLE of $\pi_{y|sp} - \pi_{y|g}$:$

We can use a large sample approximation to construct a confidence interval.

- The SE for $p_{y|sp} - p_{y|g}$ is $\sqrt{\frac{p_{y|sp}(1-p_{y|sp})}{n_{sp}} + \frac{p_{y|g}(1-p_{y|g})}{n_g}}$

$p_{y|sp} - p_{y|g} \pm z_{\alpha/2}(SE)$

Example: Construct an uncertainty interval for the difference in proportions when,

- $n_{sp,y} = 80$

- $n_{g,y} = 65$

- $n_{sp} = n_g = 100$

Sometimes we are interested in other comparisons of binary proportions, consider the ratio of the success probabilities $\frac{\pi_{sp}}{\pi_g}$.

- The ratio of probabilities is referred to as relative risk and is fairly common in medical settings. Consider the two sets of probabilities: 0.410 and 0.401 versus 0.010 and 0.001. Calculate the difference and relative risk

- difference:

- relative risk:

- Is it a good thing or bad thing that differences and relative risks can have such contrasting implications?

Another comparison, which we will see in much more detail later involves odds.

- odds $= \pi/(1-\pi)$,

so consider the following probabilities and odds:

- $\pi = .5$, odds $=$
- $\pi = \frac{2}{3}$, odds $=$
- $\pi = .75$, odds $=$
- $\pi = .8$, odds $=$

- While odds are useful for looking at a single event, we are often interested in comparing two binary events. In addition to differences and relative risk, we can also consider the odds ratio.

- The odds ratio is defined as $\frac{odds_1}{odds_2} = \frac{\pi_1(1-\pi_1)}{\pi_2(1-\pi_2)}$

- We will return to odds ratios in the context of logistic regression in coming weeks.

χ^2 test for independence

Recall the table created in the video lectures

	NO	EB	
SP	76	47	123
Gen	35	42	77
	111	89	200

Our question was whether circadian rhythm and ice cream preference were independent.

Generically we can write this table as

	1	2	
1	π_{11}	π_{12}	π_{1+}
2	π_{21}	π_{22}	π_{2+}
	π_{+1}	π_{+2}	

where the π values that include a **+** are marginal values.

- Then independence can be stated as $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}\ \forall i, j$. Describe this in words, To test this hypothesis (of independence), we can use the Pearson χ^2 statistic,

$$\chi^2_{df} = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

where the degrees of freedom is (I -1) $\times$ (J-1).

The μ_{ij} values can be calculated as $n * \pi_{i+} * \pi_{+j}$.

	NO	EB
SP	68.265	54.735
Gen	42.735	34.265

for comparison, the observed data counts were

	NO	EB
SP	76	47
Gen	35	42

Recall, we know the true values

	1	2	
1	.4	.2	.6
2	.2	.2	.4
	.6	.4	

and $\pi_{11} = .4 \neq .36 = \pi_{1+}\pi_{+1}$

Finally, let's think about this problem in the context of estimation. We can directly estimate the probabilities (joint, marginal, or conditional) associated with this data.

Assuming we use a uniform Dirichlet prior, estimate posterior means and intervals for the four joint probabilities.