# ST 518: Data Analytics II

## Multinomial Data and Distributions

Data in 2 by 2 Tables

- The Binomial Distribution

More General Contingency Tables

- The Multinomial Distribution

# Data in $2 \times 2$ Tables

The Vitamin C and Salk Vaccine examples both involve data that are recorded in $2 \times 2$ contingency tables, representing cross-classification according to two binary variables.

- In the Vitamin C example the explanatory variable, Vitamin C or Placebo, is binary, and the response variable, cold or no cold, is binary.
- In the Salk Vaccine Example, the explanatory variable, Salk Vaccine or Placebo, is binary, and the response variable, infant paralysis or no paralysis is binary.

In those cases, we treat the counts of cold getters and the counts of infant paralysis victims as binomial counts.

# Salk Vaccine Data

| | Infantile Paralysis Victim? | |
|---|---|---|
| | Yes | No |
| Placebo | 142 | 199,858 |
| Salk Vaccine | 56 | 199,944 |

For the placebo group and the Salk Vaccine group we take:

$$X_j \sim \text{bin}(200000, p_j)$$

for $j = 1, 2$, representing the placebo and Salk Vaccine groups, respectively.

# The Binomial Distribution

It's these statistical distributions that serves as the basis of our test about the difference in the two groups

- We're ultimately asking a questions about the difference between $p_1$ and $p_2$
- Or, equivalently, about the ratio of $\omega_1$ and $\omega_2$, the corresponding odds

This is an important point: it's by using a probability model (distribution) to represent data that we're able to proceed with statistical inference.

# More General Tables

What about when we move beyond the rather simple case of $2 \times 2$ tables? What probability models and statistical methods do we use in these more general situations:

- $3 \times 2$ or $8 \times 2$ tables
- $2 \times 4$ or $2 \times 5$ tables
- $3 \times 4$ or $6 \times 3$ tables
- $6 \times 2 \times 2$ or $3 \times 4 \times 4$ or even $2 \times 3 \times 4 \times 6$ tables

# The $I \times 2$ Case

In an $I \times 2$ contingency table with $I > 2$, we'll use the convention that the rows represent levels of an explanatory or grouping variable, and the columns represent the two categories of a binary response.

Here's an example schematic for a $4 \times 2$ table:

| | **Outcome** | |
|---|---|---|
| | Yes | No |
| Group 1 | | |
| Group 2 | | |
| Group 3 | | |
| Group 4 | | |

# The $I \times 2$ Case

Since the response variable (columns) has two categories, we'll use the binomial distribution to model it.

- We'll use logistic regression in this case, and that's the topic of the next two modules.
- For a simple test of homogeneity, we could use a chi-squared test provided that the sample sizes are large (all table entries are $\geq 5$).

# The $2 \times J$ Case

In a $2 \times J$ contingency table, with $J > 2$, there are two levels of the explanatory variable (rows) and $J$ levels of the response (columns).

- In the case the response variable is categorical with more than two categories.
  - We can use the multinomial distribution to model these responses, and turn to multinomial logistic regression—stay tuned for a later module.
  - You'll also see another method for this situation in this module's lab.
  - And, as before, if the sample sizes are large enough a chi-squared test for homogeneity can help you decide whether the two multinomial distributions are the same.

# The Multinomial Distribution

Remember that a multinomial random variable is represented by a vector of values, for example:

$$W_1, W_2, \ldots, W_J$$

- And, corresponding to each of those values is a category probability, $p_j$ for $j = 1, 2, \ldots, J$.
- Now, in terms of statistical inference, we might ask questions like: are any of the category probabilities different depending on levels of the explanatory variable(s)?
- For example: Are any of the Prechtl score probabilities (remember the categories are normal, dubious and abnormal) different depending on gestational age at birth?

# The General Case

In the case of an $I \times J$ contingency table and in cases where there are more than two dimensions to the table, we usually turn to regression models:

- When the response is binary, we'll use logistic regression
- When the response is categorical, we'll use multinomial logistic regression or log-linear regression (another upcoming topic)
- As always, it's important to consider the research questions of interest.