**UC Davis Graduate Group in Biostatistics M.S. Comprehensive Exam**

**Version 1.0 Assigned to Zhixin Tang**

**by Laurel Beckett, PhD (Chair of Zhixin Tang’s Exam Committee)**

The Centers for Disease Control and Prevention (CDC) tracks episodes of bacterial illnesses that are commonly transmitted through food, like *Salmonella* and *Escherichia coli (E. coli).* Their goal is to understand the patterns of occurrence and to decrease cases in the United States. Past reports have examined which pathogens are most common, where they occur, and whether there are trends over time. See, for example, the report from the CDC for 1998-2008 (Ref. 1, Gould *et al*.)

The CDC has made data available to researchers on the number of samples found to have bacteria, for each type of bacteria of interest to them, by month and year. They further give these numbers for each state where the bacteria were found. The characteristics of the state – population, for example – may influence the number of positive samples found. The number of cases may also vary seasonally.

Your assignment will have two parts: a descriptive summary of the kinds of pathogens reported and how common they are, then a more detailed examination of a pathogen of your choice.

*Part 1: Descriptive summary*: Each row of the dataset refers to one specific pathogen event: where and when did it happen, what was the pathogen, was it found in stool or urine, and how many individuals tested positive (number of isolates.) Your assignment is to determine for each pathogen how many “events” (rows of the table) were reported for that pathogen, and describe the number of isolates per event for that pathogen (e.g. mean or median, range). Include appropriate graphical summaries, e.g. box plots. At this stage, you do not need to look at differences by month or by state.

Then select a pathogen that interests you and is found frequently enough that you can learn something about its patterns of occurrence.

*Part 2: Study of patterns of occurrence*: For your selected pathogen, your next assignment is to characterize its patterns of occurrence, addressing the following questions:

* Is there seasonal variation in the presence or number of cases of this pathogen?
* Is the pathogen more common in some states than others, taking into consideration the population of the states?
* Taking population total of the state into account, is there any association between occurrence of your pathogen and how much of the population is rural (lives in or near farms or less developed areas)?

*Access to data files*: The CDC pathogen data and a data dictionary can be found and downloaded at <https://data.cdc.gov/Foodborne-Waterborne-and-Related-Diseases/BEAM-Dashboard-Report-Data/jbhn-e8xn> . Note that this is a “long” dataset, with almost 112,000 rows, but there are only 10 variables, and not all are of interest to us. Your project will focus on the following variables: Year, Month, State, Pathogen, and Number of Isolates. You may ignore the Serotype/species – for example, the first row of the table reports on *Campylobacter jejuni* and the second row on *Campylobacter upsaliensis*, but you can consider those as both belonging to the general class *of Campylobacter* occurrences. You will summarize all 111,622 rows for the first part of the assignment, characterizing the presence and number of cases of each class of pathogens.

For the second part of the assignment, you can reduce the dataset to just those rows containing data on the pathogen you have chosen to study. Then you will need data on the characteristics of each state. You can find tables of population for each state, based on US Census data, at <https://www.icip.iastate.edu/tables/population> . You can just use the 2020 US Census estimate of population, rather than worrying about changes in population during this time period. This website also has data on the percentage of the state population that is in urban areas. (The rural population is everyone else.) There may be some states that never had any cases of the pathogen of interest to you; you might consider omitting such states, but if so, please justify your decision.

*Organizing data for Part 2:* The dataset from CDC covers the period from January, 2018, to March, 2023, or 63 months. There are 50 states (plus District of Columbia). This suggests that your analytic dataset might consist of 63 x 50 (or 51) rows, one for each state for each month. The outcome of primary interest is the total number of positive isolates of the pathogen you chose, in that month, in that state. It might be 0, if there are no rows reported for that state in that month. The possible predictors are year and month, state, state population, and state percentage rural (100% - percentage urban.)

*Analysis advice/ suggestions*: The first part of your assignment should be relatively straightforward.

The second part will call on what you know about fitting regression models. Your outcome will be the number of cases, which might be zero. The standard assumptions of classical linear regression are likely to be violated here – you should have some idea of this already from the first part. What alternative regression models or approaches might be appropriate? What assumptions would your alternative approach make, and how could you check them? Do you think that successive months are likely to be independent? Your report should include a brief rationale for the model you have chosen and how you checked its assumptions. You may decide to recode some predictors; explain your choices if so.

*Suggestion*: Since the outcome is a count (number of cases), one natural model to consider is Poisson regression. If you decide to try this, a typical approach for incorporating state population would be to use log(population) as an offset in the model.

**Elements to address in your technical report**

The technical report should consist of an Abstract, Introduction, Methods, Results, and Discussion. The introduction should indicate your objective of describing the patterns and occurrence of food-borne pathogens in the US in the period from 2018 to the present. The Methods section can describe the data source and operationalization of study variables, and the statistical analysis methods you used. This section should be detailed enough that a person reading it should be able to reproduce the analyses, provided they had access to the data. The Results section, of course, presents the results of your analyses. When you interpret your findings, please explain in ways that make sense in this context. For example, incidence rates might be reported as something like number of cases per 100,000 people per month. Choose units that make your results easy to understand. The Discussion section highlights the main findings and puts them into context within the literature and can describe next steps.

The report is due to the three committee members by 5 pm on Wednesday, May 24. Please also send your abstract to Amy McFarland at that time for distribution to GGB faculty and students.

You may communicate directly by email with Dr. Beckett if you have questions on the data or the assignment.

**References:**

1. Gould LH, Walsh KA, Vieira AR, Herman K, Williams IT, Hall AJ, Cole D; Centers for Disease Control and Prevention. Surveillance for foodborne disease outbreaks - United States, 1998-2008. MMWR Surveill Summ. 2013 Jun 28;62(2):1-34. PMID: 23804024.

<https://www.cdc.gov/mmwr/preview/mmwrhtml/ss6202a1.htm>