

# STAT 439 Midterm Exam: 2025

## Part 1

The first part of this exam focuses on a dataset with car sales in Florida and Washington. The original dataset is filtered to retain sales of cars categorized as convertibles, minivans, and hatchbacks.

```
cars <- read_csv('https://raw.githubusercontent.com/STAT439/Data/refs/heads/main/car_sales.csv')
head(cars)
```

```
# A tibble: 6 x 6
  year   make     model state sellingprice type
  <dbl> <chr>    <chr>  <chr>        <dbl> <chr>
1 2012 Scion    iQ      fl       8500 Hatchback
2 2012 Subaru   Outback fl       21200 Hatchback
3 2012 Toyota   Venza   fl       14300 Hatchback
4 2012 Toyota   Prius c fl      9700 Hatchback
5 2012 Volkswagen Beetle fl      12600 Hatchback
6 2012 Toyota   Prius v wa     15400 Hatchback
```

Our research question will be to explore how the distribution of car types (convertibles, minivans, and hatchbacks) potentially differs between the Sunshine State (Florida) and the Evergreen State (Washington).

### 1. Data Visualization (4 points)

Create a figure (or figures) to visualize the research question stated above. Include a summary paragraph describing your findings.

## **2. Contingency Table**

### **2.1 (2 points) Table Construction**

Create and print a contingency table for state by type

### **2.2 (4 points) Testing**

Run a test for independence between state and type

### **2.3 (4 points) Written Summary**

In a paragraph, write a summary of your findings in part 2.2.

## **3. Inference for Car Types**

For this question we will estimate the multinomial probabilities corresponding to vehicle type ( $\pi_{convertible}$ ,  $\pi_{minivans}$ , and  $\pi_{hatchbacks}$ ) across the two states.

### **3.1 (2 points) Bayesian Prior Specification**

State and justify a prior distribution, for the three probabilities ( $\pi_{convertible}$ ,  $\pi_{minivans}$ , and  $\pi_{hatchbacks}$ ), for each state.

### **3.2 (2 points) Bayesian Posterior**

What are the posterior distributions for each state?

### **3.3 (4 points) Visual Summary**

Create a figure that includes the uncertainty intervals for the probability of vehicle types for each state. You can use maximum likelihood or Bayesian methods.

### **3.4 (4 points) Written Summary**

Create an uncertainty interval for the difference in proportion of vehicles classified as convertibles between FL and WA. You can use maximum likelihood or Bayesian methods.

### 3.5 (4 points) Written Summary

Using the results from 3.2, 3.3, and 3.4, write a paragraph summary discussing the differences in car distribution between Florida and Washington - make sure to include uncertainty when discussing parameter estimates.

## Part 2

The second part of the exam will involve model fitting with logistic regression. Use the `midterm_data` and note that `y` is a single binary variable.

```
midterm_data <- read_csv('https://raw.githubusercontent.com/STAT439/Data/refs/heads/main/midterm_data')
```

```
# A tibble: 6 x 5
  y     x1      x2      x3 x4
  <dbl> <dbl>    <dbl>    <dbl> <chr>
1 1     -3     -1.46   -0.952 A
2 1     -2.99  0.0120  1.79   C
3 1     -2.99  -3.89   -0.498 C
4 0     -2.98  0.301   0.193   B
5 0     -2.98  -2.07   1.90    B
6 1     -2.97  -1.76   0.981   C
```

**4. (4 points) Create a set of EDA figures to explore the relationship between the response (success out of 1 trial) and the potential covariates.**

**5. (4 points) Summarize your findings in the figures**

Which variables and combinations of variables do you think are important?

**6. (4 points) Using residual diagnostics and AIC fit a series of models.**

You don't need to print out all of these results, but include a written summary of models you explored including the final model that you ended up selecting. You are welcome to use bullet points for this section.

**7. (4 points) Graphically summarize the final model you selected**

Include estimated model fits for all parameters or combinations of parameters included in your model.

**8. (4 points) Written summary the final model you selected**

Describe the final model you selected and discuss how each variable (or combination of variables) impact the probability of success.