DELTA Testing Services

go.ncsu.edu/testing

Campus Box 7555
1730 Varsity Dr.
Venture IV, Suite 236
Raleigh, NC 27695-7113

919.515.1560 phone
919.515.7180 fax

delta-testing@ncsu.edu

NC STATE
UNIVERSITY

# DELTA Testing Services

**Student Name:** David Grant **Date:** 3/6/2025

**Student's NCSU Email Address:** dgrant@ncsu.edu

**Course:** ST 563 601 **Exam #:** 2

**Start Time:** 10:00 **End Time:** 11:15

**Proctor's Name (Print):** Dylan Switala

**Proctor's Signature:** Dylan Switala

**Institution:** Ramapo College of New Jeresey

**PLEASE SIGN & DATE THIS SHEET AND RETURN ALONG WITH THE EXAM**

## Proctoring Guidelines

If you are unable to comply with the following, please destroy the exam and have the student submit the name of another proctor for approval.

1. Please ask student for their photo ID.
2. **Have the student put their name on the exam and exam answer sheet.**
3. The test should be conducted in an atmosphere conducive to good concentration (quiet, good lighting, etc.).
4. The student must take the exam without outside help. Have the students leave all materials (except blank paper, pen or pencil, or calculator, as needed) outside the testing room. This includes notes, books, calculators, phones, etc. (excluding materials required for the exam).
5. Close and constant supervision must be provided.
6. Please scan and email the proctoring form, completed exam, and any formula sheets permitted for the assessment to delta-testing@ncsu.edu or fax to 919-515-7180.
7. Not including exams that permit all notes or textbooks, students should not be permitted to leave the testing room with formula sheets or scrap paper unless explicitly stated.
8. DO NOT GIVE THE EXAM TO THE STUDENT TO MAIL BACK

If you have any questions, please contact DELTA Testing Services at our main Venture IV location via phone: (919)-515-1560 or e-mail: delta-testing@ncsu.edu.

Thank you for assisting our students.
DELTA Testing Services
NC State University

# ST 563 601 – SPRING 2025 – POST
# Exam #2

**Student's Name:** David Grant

**Date of Exam**: Thursday, March 6, 2025 - Friday, March 7, 2025
**Time Limit**: 75 minutes
**Allowed Materials**: None (closed book & closed notes)

**Student – NC State University Pack Pledge**

I, David Grant have neither given nor received unauthorized aid on this exam or assignment. I have read the instructions and acknowledge that this is the correct exam.

*Student's Printed Name*

3/6/25

*Student Signature* *Date*

**Exam must be turned in by:** 10: 45

*Exam End Time* *Student's Initial Agreement*

**NOTE: Failure to turn in exam on time may result in penalties at the instructor's discretion.**

# Exam 2

Please write your answers below each question. You should not have access nor use any materials during this exam.

A reminder that, by taking this exam, you are required to uphold the NC State honor pledge:

"I have neither given nor received unauthorized aid on this test or assignment."

1. In doing a classification task, we discussed the idea of classification and the idea of discrimination. What are these and what is the difference between the two? (8 pts)

Classification is using the conditional probabilities of $Y|X$ to find the most likely class, whereas discrimination finds the conditional probabilities of $X|Y$ and multiplies that by the likelihood and divided by the total probability to obtain the most likely class (Bayes' Theorem)

2. Suppose we have a categorical response with four levels. We could label those four levels with numeric values, say $Y = 1,2,3$, or 4. Explain the implications of treating our problem as a regression task with these values for $Y$. Could it ever make sense to do this? (6 pts)

No, it doesn't make sense to do this, for a couple of reasons. The first is because regression is continuous and so we could predict values outside of that range. The other issue is that the differences between each level shouldn't be quantified to be the same. i.e. even if we obtained an expected value of 2.5, it isn't practical to interpret that as being the same "distance" away from level 2 and level 3, especially if the levels aren't ordinal.

3. Select true or false for each classification method. (3 pts each)

True — a) We can never use the Bayes classifier in a real scenario.

True — b) LDA is a special case of QDA.

False — c) Logistic Regression provides a discriminant for classifying our observations.

False — d) Binary logistic regression generally requires a larger sample size than multinomial logistic regression.

4. We discussed the idea of the Bayes' error rate. Can we ever do better than this rate? Explain. (5 pts)

No, we can never do better than this rate because this rate is based on using the true population conditional distributions of $Y|X$ and using that to predict the most likely class for each $X$.

5. One measure of the quality of a classification model is accuracy. Define the no information rate and describe how interpreting the accuracy of a model is related. (6 pts)

No information rate (NIR) is the accuracy of simply predicting the most prevalent class for every observation. When we build a model, we want to make sure the accuracy is better than the NIR, otherwise we're saying it's better to predict the most likely class regardless.

6. Define the terms sensitivity and specificity. (6 pts)

Sensitivity - is how well the model predicts observations that are supposed to be in the desired class

Specificity - is how well the model predicts observations that are not supposed to be in the desired class.

7. When using a generative model for classification, we need to estimate the *prior probabilities* for each class. What is the most basic way we discussed for estimating these probabilities? (6 pts)

Simply take the number of observations in a specific class and divide by the total number of observations. Do this for each class.

8. Suppose we have a categorical response with $m$ categories and a single predictor variable $X$. When fitting an LDA model, we use normal distributions. What quantities do we model with a Normal distribution? Are those normal distributions related in anyway? (6 pts)

We model the conditional distributions of $X|Y$. No, they're not related. Each normal distribution has its own mean and standard deviation.

9. When trying to use LDA or QDA with $p = 10$ predictors, we can note that LDA is a special case of QDA. Why might we still prefer LDA to QDA even though QDA is more general? (6 pts)

QDA has more variance than LDA, because for QDA, there's a different variance-covariance matrix for each set of predictors. With 10 predictors, that could be a lot, therefore QDA runs a greater risk of overfitting to the training data.

10. We discussed the Naive Bayes classifier. This is a generative model. What simplifying assumption do we make when using the Naive Bayes classifier? (6 pts)

The assumption is that the predictor variables are all independent.

11. What is the difference between a cubic spline model and a natural cubic spline model? (6 pts)

A cubic spline model in a way tunes where the knots should be based on changes of patterns in the data, whereas a natural cubic spline sets the knots at equal-length intervals across the data.

12. Suppose we have data on whether or not someone has heart disease (No = 0, Yes = 1) and a number of predictors such as Age (quantitative), ExerciseAngina (Y or N), and Cholesterol (quantitative). We fit a logistic regression model with 'main effects' for each of these predictors. Relevant output is given below.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -4.4039 | 0.6501 | -6.7742 | 0.0000 |
| Age | 0.0530 | 0.0100 | 5.2905 | 0.0000 |
| ExerciseAnginaY | 2.4644 | 0.1925 | 12.8046 | 0.0000 |
| Cholesterol | 0.0024 | 0.0015 | 1.6052 | 0.1085 |

a) What is the fitted equation for those without Exercise Angina? Be careful how you write the left hand side of the model! No need to simplify. (6 pts)

$$p(Y=1) = \frac{e^{-4.4039 + 0.053X_1 + 0.0024X_3}}{1+e^{-4.4039 + 0.053X_1 + 0.0024X_3}}$$

b) How would we use this fitted equation to find a decision boundary for those without exercise angina? This isn't something you can solve! Just write down how you would use the equation to find the boundary for values of Age and Cholesterol. (6 pts)

We first set the L.H.S. of the equation to 0.5, meaning it's equally likely to have heart disease or not have it. Then, we can backsolve by multiplying the denominator of our equation by 0.5. What we'd hale to do though is plug in each age value for $X_1$ to solve for $X_3$ (the cholesterol value). Each $X_1$, $X_3$ pairing (which is continuous) represents our decision boundary.

c) How do we interpret the meaning of the intercept coefficient for this model? Be sure to use the context of the data. (5 pts)

The intercept coefficient represents the associated log odds of having heart disease given someone of age 0, has no exercise angina, and no cholesterol.

d) How do we interpret the meaning of the age slope coefficient for this model? Be sure to use the context of the data. (5 pts)

The age slope coefficient represents the change in the log odds of success (having heart disease), while holding the exercise angina and cholesterol values constant, for each unit increase in age.

e) How do we interpret the meaning of the ExerciseAnginaY coefficient for this model? Be sure to use the context of the data. (5 pts)

The exercise angina coefficient represents the log odds of someone having heart disease, given that they have exercise angina, and holding the values of age and cholesterol constant.