56

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: Matthew Bray **Date:** 3/7/25

Student's NCSU Email Address: rmbray@ncsu.edu

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

Start Time: 1:25pm **End Time:** _____

Proctor's Name (Print): Alexander Khoury

Proctor's Signature: _____

Institution: Bridgewater State University

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

Updated March 2022

ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Matthew Bray

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, Matt Bray 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

Student Signature 07Mar25 *Date*

Exam must be turned in by: 2:40 pm *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 placing the observation into a category.

Discrimination is finding a function that creates the boundaries between the classes.

sort of -2

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)

If the categories can plausibly be thought of as being ordinal, say for a severity classification, then regression could possibly be used. If the levels are not ordinal (eg red, green, yellow, purple or truck, car, bike, train), then regression wouldn't make sense because any level could be described by any feature.

-2

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

   a) We can never use the Bayes classifier in a real scenario. true
   b) LDA is a special case of QDA. true
   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. true

-3

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

No, the Baye's error rate is analogous to the irreducible error (ε) from regression modeling.

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)

most prevalent

The no information rate is how the data are classified given no model (e.g. proportion of class 1 to total data size).

If the accuracy of the model is not better than the NIR, then the model is pointless.

-1

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

Sensitivity = ~~False~~ positive rate

Specificity = ~~False~~ negative rate

-6

-13

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)

model as $N(\hat{\mu}_k, \hat{\sigma}_k^2)$, where k is the class -6

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)

Each class is modeled with normal distributions. In LDA, they do share the same variance/covariance matrix. (prior probabilities of each class) -4

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)

If the boundaries are linear, then LDA will better predict on unseen data. -4

ok but big issue is # of params

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 prior probabilities are independent -6

-20

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

Natural cubic spline is smooth at the knots, cubic spline is not. -4

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)

D. $\ln\left(\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}\right) = -4.4039 + 0.0530 * Age + 2.4644 * ExerciseAngina + 0.0024 * Cholesterol$

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)

Solve the equation for [redacted] $\hat{P}(Y=1) = 0.5$, where exerciseAnginaY = 0/No

-4

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)

log-odd

where the log-odds of having heart disease = 0

-5

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 log-odds of having heart disease increase by 0.0530 for each unit increase in Age, holding Exercise Angina and Cholesterol constant.

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)

When ExerciseAnginaY = no, the slope coefficient becomes 0. When ExerciseAnginaY = Yes, the log-odds of having heart disease increase by 2.4644

-5