UNIVERSITY OF SAN FRANCISCO | College of Arts and Sciences

BACHELOR OF SCIENCE IN DATA SCIENCE

ASSESSMENT REPORT
ACADEMIC YEAR 2018 – 2019

I. LOGISTICS & PROGRAM LEARNING OUTCOMES

1. **Please indicate the name and email of the program contact person to whom feedback should be sent (usually Chair, Program Director, or Faculty Assessment Coordinator).**

   Outgoing Director (F19): James D. Wilson (jdwilson4@usfca.edu)

   Incoming Director (S20): Daniel O'Connor (doconnor@usfca.edu)

2. **Were any changes made to the program mission statement since the last assessment cycle in October 2018? Kindly state "Yes" or "No." Please provide the current mission statement below. If you are submitting an aggregate report, please provide the current mission statements of both the major and the minor program.**

   No

   To deliver a high-quality data science program that instructs students in the theory and practice of mathematical and computational analysis of applied data driven problems, and to graduate students with appropriate experience in industry-standard data science tools.

3. **Were any changes made to the program learning outcomes (PLOs) since the last assessment cycle in October 2018? Kindly state "Yes" or "No." Please provide the current PLOs below. If you are submitting an aggregate report, please provide the current PLOs for both the major and the minor programs.**

   No

   - [PLO1] Analyze information critically and logically in a mathematical setting.
   - [PLO2] Reformulate and solve problems in an abstract framework.
   - [PLO3] Express mathematical results verbally, working individually and in collaborative groups.
   - [PLO4] Apply mathematical techniques to specific problem domains
   - [PLO5] Demonstrate competence with programming concepts, including software development techniques and data structures
   - [PLO6] Apply mathematical and computational techniques to real-world problems involving large, complex data sets.
   - [PLO7] Visualize, present and communicate analytical results.

4. **Which particular Program Learning Outcome(s) did you assess for the academic year 2017-2018?**

   PLO1, PLO4, PLO5

II. METHODOLOGY

5. **Describe the methodology that you used to assess the PLO(s).**

   For example, "the department used questions that were inputted in the final examination pertaining directly to the <said PLO>. An independent group of faculty (not teaching the course) then evaluated the responses to the questions and gave the students a grade for responses to those questions."

   We directly assessed all graduating seniors with and end-of-degree exam given in the Spring 2019 semester. This exam consisted of 14 multiple choice questions spanning topics from the required curriculum. This exact exam, with the same 14 questions, was also given to the previous cohort of graduating seniors in the Spring 2018 semester as well as (for the first time) to the cohort of graduating seniors in Spring 2017. It is our intention to continually assess our students and, by extension, the program by annually giving the graduating seniors the same exam. This will provide objective and comparable year-over-year data with which we can evaluate the effectiveness of the program. We presently have three years of data whose results and findings I will discuss in the next section. Note that the exit exam is attached as a separate document.

III. RESULTS & MAJOR FINDINGS

6. What are the major takeaways from your assessment exercise?

This section is for you to highlight the results of the exercise. Pertinent information here would include:

Having only used this exit exam as a direct assessment tool for three years, it is difficult to draw strong year-over-year conclusions, especially with such small sample sizes. That said, strictly speaking, graduating seniors performed slightly worse (on average) on the exam in 2019 than the ones in 2017 and 2018, as is evidenced by the data presented in Figure 1. However, a two-sample *t*-test does not find these averages (9.00 vs. 7.18) to be statistically significantly different (*p*-value = 0.6510). There was also no significant difference between the test averages from 2017 to 2018's graduating seniors.

Because there is no significant difference in exam scores between consecutive years we pool all years together, increasing our effective sample size to $n = 27$. In Figure 1 we see that the scores range from a minimum of 4 to a maximum of 12. We distinguish among different levels of student mastery based on these scores. In Table 1 we define four levels of mastery, map those to ranges of test scores and identify the percentage of students achieving each level. This information is also depicted in Figure 2.

Level of Mastery	Exam Scores	No. Students	% Students
Poor	0-6	9	33%
Satisfactory	7-9	8	30%
Good	10-11	9	33%
Excellent	12-14	1	4%

Table 1: BSDS Levels of Mastery (2017 - 2019 aggregate)

Overall, we were pleased with the performance of our graduating seniors though we recognize there is still room for improvement. For example, we would prefer to have a smaller percentage of students scoring in the "Poor" category – though this distribution should be interpreted cautiously as there are so few data points. Nonetheless, we intend to provide additional emphasis on problem areas identified by the exam in our degree-required classes. This is elaborated upon in the next section.

As mentioned previously, we also plan to give this exam to every graduating student on an annual basis allowing us to evaluate year-over-year improvements.

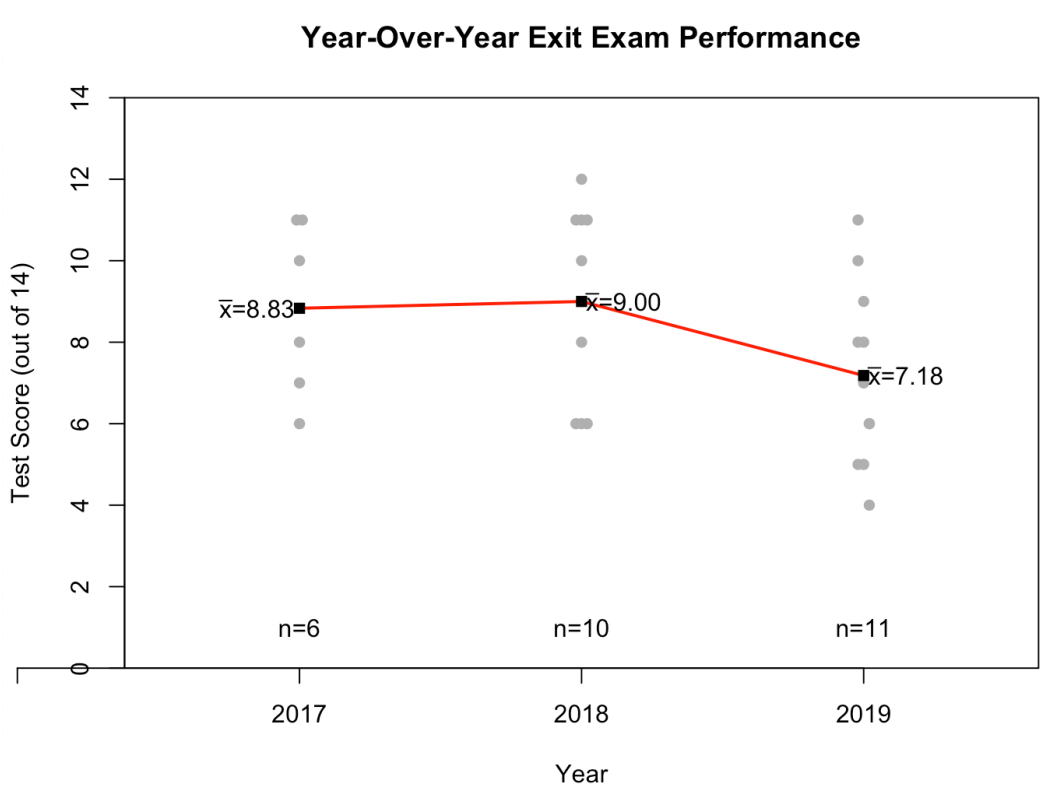

Figure 1: BSDS Exit Exam Scores Year over Year

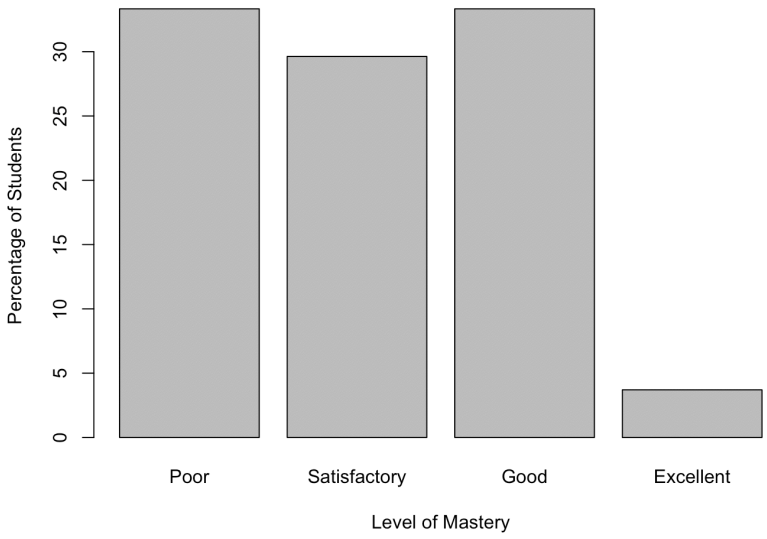

Figure 2: BSDS Levels of Mastery (2017 – 2019 aggregate)

IV. CLOSING THE LOOP

7. **Based on your results, what changes/modifications are you planning in order to achieve the desired level of mastery in the assessed learning outcome? This section could also address more long-term planning that your department/program is considering and does not require that any changes need to be implemented in the next academic year itself.**

   By carefully scrutinizing the exam results it has become apparent that Questions 1, 3, 8, 11 and 13 are the ones that students most often struggle with. The students' difficulty with these questions indicates a struggle specifically with PLO4. The topics being tested by these questions are:

   - Conditional probability
   - Eigenvalue calculation
   - *p*-value interpretation
   - Likelihood estimation

   All of these topics are tied to specific classes (MATH 230, MATH 370, MATH 371). Based on these findings we plan to ensure these topics (in these classes) are clearly emphasized and that their importance beyond the classroom is highlighted. This should help to improve student performance on these questions and the level of mastery associated with this learning outcome.

8. **What were the most important suggestions/feedback from the FDCD on your last assessment report (for academic year 2017-2018, submitted in October 2018)? How did you incorporate or address the suggestion(s) in this report?**

   The feedback that we received last year was quite positive. One area of suggestion was that we continue to improve the students' ability to accomplish PLO4, which was mentioned as an area of struggle for students in question 7 of the 2017 - 2018 assessment. This past year, to help ensure sufficient student progress and engagement in the BSDS major, we voted as a department to require a C- in all requisite courses in the major. This action requires students to obtain at least a C- in any prerequisite courses before given the ability to take upper level courses. This is a small change from the previous D requirement; however, we believe that this change will have a significant positive impact on the students' ability to achieve our course and program learning outcomes.

V. ADDITIONAL MATERIALS

Below we include the following additional materials:

- End of the degree exam
- Curricular maps

BSDS EXIT EXAM Instructions:

Don't consult with any outside materials (e.g., notes, web sites, R documentation, etc.) or individuals as you work through this examination. Feel free to use scratch paper and a calculator.

1. Suppose $P(A) = 0.60$, $P(B) = 0.47$ and $P(A \cap B) = 0.19$. Calculate $P(A|B^C)$.

(a) 0.36
(b) 0.41
(c) 0.68
(d) 0.77
(e) 0.87

2. Suppose a store has 100 light bulbs in stock. Assume 40 light bulbs are from Distributor A and the remainder of the light bulbs are from Distributor B. Assume 5.0% of the light bulbs from Distributor A are defective and 10.0% are defective from Distributor B.

If a consumer purchases 3 light bulbs, what is the probability that exactly 2 of the light bulbs are defective? Choose the correct expression.

(a) $\binom{2}{3}0.08^3(1-0.08)^2$
(b) $0.08^2(1-0.08)$
(c) $\binom{3}{2}0.08^2(1-0.08)$
(d) 0.08
(e) $0.08(1-0.08)$

3. Suppose that you know that exactly 10% of all emails sent to you are spam. Moreover, you know that 80% of the emails that you have received in the past that were not spam contained the word "analytics" and that 40% of the emails that you received that were spam contained that word. Suppose that you receive a new email that contains the word "analytics." What is the probability that the email is spam?

(a) 0.40
(b) 0.10
(c) 0.053
(d) 0.056

4. Consider the following Python code:

```
def f(x):
    ind=0
    val=x[0]
    for i in range(len(x)):
        if x[i]>val:
            val = x[i]
            ind = i
    return ind
l=[1,22,13,194,5,-4,0]
f(l)
```

The result of running the above code is:

(a) 3
(b) 4
(c) 194
(d) 0

5. Consider the following R code:

```
mat <- matrix(rnorm(200000,50,7),10000,20)
rms <- rowMeans(mat)
par(mfrow=c(2,1))
hist(rnorm(10000,50,7),breaks=50,xlim=c(30,70))
hist(rms,breaks=50,xlim=c(30,70))
```

The result of running the above code is best described as illustrating:

(a) The standard normal distribution
(b) A confidence interval for a population mean
(c) The central limit theorem
(d) A chi-square distribution

6. A researcher has a dataset consisting of, among other things, the anual income for a large sample of individuals in the US. If the researcher were to make a histogram of this income data, it is likely that the distribution would be:

(a) Exponential
(b) Symmetric
(c) Skewed to the left
(d) Skewed to the right

7. A researcher studying a sample of U.S. cities plots, for each city, the number of churches per thousand people on the x-axis and the number of violent crimes per thousand people on the y-axis. The researcher notices that the points in the scatterplot could be reasonably described as moving up and to the right in a roughly linear fashion. Which of the following statements is most correct based on this information?

(a) Increasing the number of churches in a city will increase the crime rate.
(b) $r > 0$ where r is the correlation coefficient.
(c) The number of churches and the crime rate are independent.
(d) $\beta_1 = 0$ where β_1 is the slope of the population regression line.

8. A clinician runs a double blind experiment to assess the effectiveness of a potential new drug for migraine headaches. She records the mean improvement in migraine severity in both a treatment group and an independent placebo group, and uses them to run a t-test. If she obtains a p-value of 0.03, and assuming that this test is appropriate, which conclusion is correct?

(a) The experiment shows that the drug reduces migraine severity by 3%.
(b) 97% of people in the treatment group improved.
(c) There is a 3% chance that the drug is ineffective.
(d) If the drug were ineffective, the researcher would expect data like her's 3% of the time.

9. Which technique could plausibly be used by a marketer to determine the probability that a customer will make a purchase based on several pieces of demographic information?

(a) Analysis of variance
(b) Simple linear regression
(c) Logistic regression
(d) Chi-squre test

10. Suppose $\mathbf{X} = [X_1, X_2, \ldots, X_n]^{\mathrm{T}}$ is a vector where each X_i is a random variable with mean $E[X_i]$ and finite variance σ_i^2. Define Σ to be the matrix $E\left[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^{\mathrm{T}}\right]$. The (i, j) entry of the matrix Σ is:

(a) The joint density of X_i and X_j.
(b) The conditional expectation of X_i given X_j.
(c) 0
(d) The covariance between X_i and X_j.

11. Which of the following is an eigenvalue for the matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 4 \\ 5 & 0 & -2 \end{bmatrix}$$

(a) 3
(b) -2
(c) 2
(d) 0

12. Suppose X is an $n \times k$ matrix with linearly independent columns, and $\mathbf{b}$ is a vector in $\mathbb{R}^n$ that is not in the column space of X. The best approximation to $\mathbf{b}$ in the column space of X is given by:

(a) $X^{-1}\mathbf{b}$
(b) $X(X^TX)^{-1}X^T\mathbf{b}$
(c) Does not necessarily exist
(d) $(X^TX)^{-1}X^T\mathbf{b}$

13. Let $Y_1, Y_2, \ldots Y_n$ be a random sample of observations from a uniform distribution with probability density function given, for each $i = 1, 2, \ldots, n$, by $f(y_i) = \frac{1}{\theta}$, $0 \leq y_i \leq \theta$. The maximum- liklihood estimator of θ is:

(a) $\frac{1}{n}\sum_{i=1}^{n} Y_i$
(b) $\min\{Y_1, Y_2, \ldots, Y_n\}$
(c) $\max\{Y_1, Y_2, \ldots, Y_n\}$
(d) 1

14. Consider a classification problem involving only two classes $\{0, 1\}$ using a p dimensional predictor X. The Bayes classifier sets:

$$P(Y = 1|X = x) = \frac{P(Y = 1)P(X = x|Y = 1)}{P(Y = 0)P(X = x|Y = 0) + P(Y = 1)P(X = x|Y = 1)}.$$

Suppose we take three new observations x_a, x_b, and x_c, and find that $P(Y = 1|X = x_a) = 0.55$, $P(Y = 1|X = x_b) = 0.87$, $P(Y = 1|X = x_c) = 0.32$. How should these three observations (x_a, x_b, x_c) be classified?

(a) $(1, 1, 1)$
(b) $(1, 1, 0)$
(c) $(1, 0, 0)$
(d) $(0, 1, 1)$
(e) $(0, 0, 1)$