

MS&E 125 Final (sample)

Name: _____

SUNet ID: _____@stanford.edu

Instructions

- Unless otherwise noted, each problem is self contained.
- Clearly mark your answer on the *answer sheet*, found on the last page.
- *Only the answer sheet will be graded*, with no partial credit.
- Questions with more than one answer marked on the answer sheet will be considered incorrect.
- There are a total of 40 questions, 10 true/false questions and 30 multiple choice questions.
- Each problem is identically worth 2.5 points, for a maximum total of 100, so don't spend too much time struggling on any single question.
- You will get zero points for any incorrectly answered question (no negative points).
- You may use all resources available (books, notes, homework solutions, and general Internet) for this exam. However, we recommend against using materials outside of this course, such as searching the Internet, since it will more likely result in over-complication, confusion and a waste of time.
- Good luck!

True/false questions

Problem 1.

For independent samples X_1, X_2, \dots, X_n from an exponential distribution with parameter θ , the sampling distribution of $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is approximately normal for large n .

Problem 2.

For a fixed set of covariates, a fitted L^2 regularized (ridge) regression model will typically have more non-zero coefficients than an L^1 regularized (lasso) regression model.

Multiple choice questions

As a data scientist at Webflix, you are given views data. Each row in the dataset indicates views of a movie for a given user during the first quarter of 2019. There are 70 movies and 10 users, for a total of 700 rows. You load the data in R as `views_df` and see that it has the following columns:

- **user**: a character column of unique user IDs for a total of 10 users
- **movie**: a character column of 70 unique movie IDs
- **viewed**: a logical column of TRUE/FALSE indicating whether each user watched the corresponding movie during the first quarter of 2019.

Answer problems 3 and 4.

Problem 3.

We would like to know how many users watched each movie. What are the correct values for each of the blanks in the following snippet

```
views_df %>%  
  group_by([ A ]) %>%  
  summarize(total_views = [ B ](viewed))
```

- (a) [A]: user, [B]: sum
- (b) [A]: user, [B]: max
- (c) [A]: movie, [B]: sum
- (d) [A]: movie, [B]: max

Problem 4.

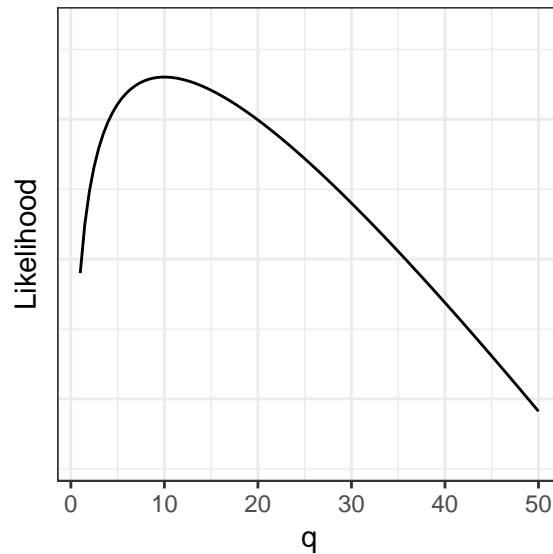
How many rows will the following code snippet result in?

```
views_df %>%  
  group_by(movie) %>%  
  summarize(p_view = mean(viewed)) %>%  
  filter(p_view > .6)
```

- (a) 10
- (b) 70
- (c) 700
- (d) Cannot determine without running the code

Problem 5.

The figure below shows the likelihood of some observed data X as a function of *all* possible values for the parameter q .



Which of the following statements is true?

- (a) The MLE for q is 10
- (b) The MLE for q is $\min(X)$
- (c) The MLE for q is $\max(X)$
- (d) The MLE for q cannot be determined from the plot

Problem 6.

You are given the model, `house_price ~ 1 + year_built + sq_feet + school_district_quality`, where `year_built` and `sq_feet` are continuous covariates, and `school_district_quality` is a categorical variable expressing the quality of the school in the area, with four possible values: A, B, C, and D. How many coefficients does the model have (including the intercept)?

- (a) 7
- (b) 6
- (c) 4
- (d) 3

Answer sheet

Name: _____

SUNet ID: _____@stanford.edu

True/false questions

Fill-in the circle of the correct answer. (T = true, F = false)

1 (T) (F)

2 (T) (F)

3 (T) (F)

4 (T) (F)

5 (T) (F)

6 (T) (F)

7 (T) (F)

8 (T) (F)

9 (T) (F)

10 (T) (F)

Multiple choice questions

11 (a) (b) (c) (d)

12 (a) (b) (c) (d)

13 (a) (b) (c) (d)

14 (a) (b) (c) (d)

15 (a) (b) (c) (d)

16 (a) (b) (c) (d)

17 (a) (b) (c) (d)

18 (a) (b) (c) (d)

19 (a) (b) (c) (d)

20 (a) (b) (c) (d)

21 (a) (b) (c) (d)

22 (a) (b) (c) (d)

23 (a) (b) (c) (d)

24 (a) (b) (c) (d)

25 (a) (b) (c) (d)

26 (a) (b) (c) (d)

27 (a) (b) (c) (d)

28 (a) (b) (c) (d)

29 (a) (b) (c) (d)

30 (a) (b) (c) (d)

31 (a) (b) (c) (d)

32 (a) (b) (c) (d)

33 (a) (b) (c) (d)

34 (a) (b) (c) (d)

35 (a) (b) (c) (d)

36 (a) (b) (c) (d)

37 (a) (b) (c) (d)

38 (a) (b) (c) (d)

39 (a) (b) (c) (d)

40 (a) (b) (c) (d)

Solutions

Problem 1: True

Problem 2: True

Problem 3: (c)

Problem 4: (d)

Problem 5: (a)

Problem 6: (b)