UNIVERSITY OF SAN FRANCISCO | College of Arts and Sciences

2023-2024 MS in Data Science Annual Assessment Report

Table of Contents

UNIVERSITY OF SAN FRANCISCO | College of Arts and Sciences

MS in Data Science

Name of the Program

MS in Data Science

Degree Type

Graduate

Contact Information

Shan Wang
Program Academic Director/Assistant Professor
MS in Data Science
swang151@usfca.edu

Mission Statement

The mission of our program is to produce graduates who possess a theoretical and practical understanding of many classical and modern statistical modeling and machine learning techniques; who use contemporary programming languages and technologies to scrape, clean, organize, query, summarize, visualize, and model large volumes and varieties of data; and who use their knowledge and skills to successfully solve real-world data-driven business problems and to communicate those solutions effectively.

Notes: The MSDS program’s mission statements were ratified by its faculty during a January 2016 vote over email

Program Learning Outcome

Our program learning outcomes (PLOs) were changed in December of 2018 and ratified by a vote of the faculty to the following:

PLO 1. Possess a theoretical understanding of classical statistical models (e.g., generalized linear models, linear time series models, etc.), as well as the ability to apply those models effectively.

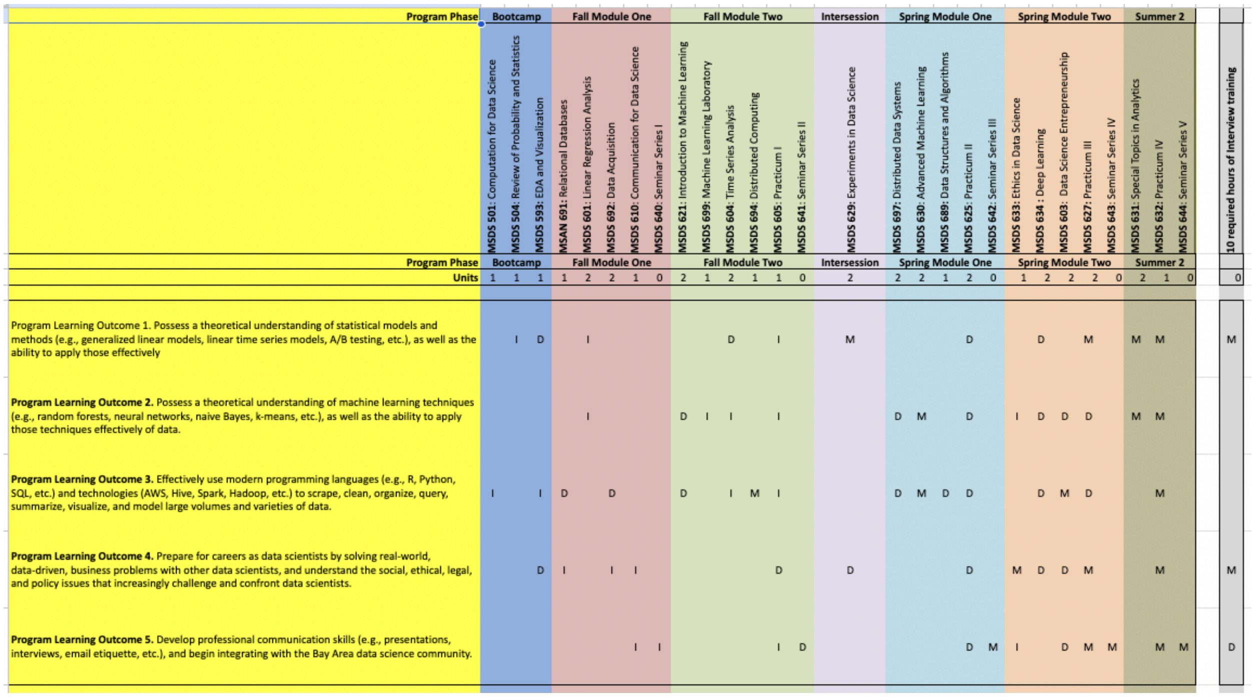
PLO 2. Possess a theoretical understanding of machine learning techniques (e.g., random forests, neural networks, naive Bayes, k-means, etc.), as well as the ability to apply those techniques effectively.

PLO 3. Effectively use modern programming languages (e.g., R, Python, SQL, etc.) and technologies (AWS, Hive, Spark, Hadoop, etc.) to scrape, clean, organize, query, summarize, visualize, and model large volumes and varieties of data.

PLO 4. Prepare for careers as data scientists by solving real-world, data-driven, business problems with other data scientists, and understand the social, ethical, legal, and policy issues that increasingly challenge and confront data scientists.

PLO 5. Develop professional communication skills (e.g., presentations, interviews, email etiquette, etc.), and begin integrating with the Bay Area data science community.

Current Curricular Map



Program Phase	Bootcamp			Fall Module One					Fall Module Two						Intersession	Spring Module One					Spring Module Two					Summer 2			
	MSDS 501: Computation for Data Science	**MSDS 504**: Review of Probability and Statistics	**MSDS 593**: EDA and Visualization	**MSAN 691**: Relational Databases	**MSDS 601**: Linear Regression Analysis	**MSDS 692**: Data Acquisition	**MSDS 610**: Communication for Data Science	**MSDS 640**: Seminar Series I	**MSDS 621**: Introduction to Machine Learning	**MSDS 699**: Machine Learning Laboratory	**MSDS 604**: Time Series Analysis	**MSDS 694**: Distributed Computing	**MSDS 605**: Practicum I	**MSDS 641**: Seminar Series II	**MSDS 629**: Experiments in Data Science	**MSDS 697**: Distributed Data Systems	**MSDS 630**: Advanced Machine Learning	**MSDS 689**: Data Structures and Algorithms	**MSDS 625**: Practicum II	**MSDS 642**: Seminar Series III	**MSDS 633**: Ethics in Data Science	**MSDS 634** : Deep Learning	**MSDS 603**: Data Science Entrepreneurship	**MSDS 627**: Practicum III	**MSDS 643**: Seminar Series IV	**MSDS 631**: Special Topics in Analytics	**MSDS 632**: Practicum IV	**MSDS 644**: Seminar Series V	**10 required hours of Interview training**
Program Phase	**Bootcamp**			**Fall Module One**					**Fall Module Two**						**Intersession**	**Spring Module One**					**Spring Module Two**					**Summer 2**			
Units	1	1	1	1	2	2	1	0	2	1	2	1	1	0	2	2	2	1	2	0	1	2	2	2	0	2	1	0	0
Program Learning Outcome 1. Possess a theoretical understanding of statistical models and methods (e.g., generalized linear models, linear time series models, A/B testing, etc.), as well as the ability to apply those effectively		I	D		I						D		I		M				D			D		M		M	M		M
Program Learning Outcome 2. Possess a theoretical understanding of machine learning techniques (e.g., random forests, neural networks, naive Bayes, k-means, etc.), as well as the ability to apply those techniques effectively of data.					I				D	I	I		I			D	M		D		I	D	D	D		M	M		
Program Learning Outcome 3. Effectively use modern programming languages (e.g., R, Python, SQL, etc.) and technologies (AWS, Hive, Spark, Hadoop, etc.) to scrape, clean, organize, query, summarize, visualize, and model large volumes and varieties of data.	I		I	D		D			D		I	M	I			D	M	D	D			D	M	D			M		
Program Learning Outcome 4. Prepare for careers as data scientists by solving real-world, data-driven, business problems with other data scientists, and understand the social, ethical, legal, and policy issues that increasingly challenge and confront data scientists.			D	I		I	I						D		D				D		M	D	D	M			M		M
Program Learning Outcome 5. Develop professional communication skills (e.g., presentations, interviews, email etiquette, etc.), and begin integrating with the Bay Area data science community.							I	I					I	D					D	M	I		D	M	M		M	M	D

Assessment Schedule between APRs

We are assessing 5 PLOs throughout the 3-year plan.

Year	Assessment Schedule
Year 1	Assessing PLO 1 based on statistics courses including MSDS 504, 601, 604, and 629
Year 2	Assessing PLO 2, 3 and 4 based on computational courses including MSDS 621, 630, 689, 691, 692, 694, 697, and 699
Year 3	Assessing PLO 5 based on communication and business courses including MSDS 610 and 603

This report presents the assessment of PLO 1. The last assessment on this PLO was done in 2016-2017.

Description of the Assessment Methodology

Two faculty (Shan Wang and Nathaniel Stevens) in MSDS contributed representative questions from one or more written exams and assignments taken from four data science courses related to PLO1.

These learning outcomes are loosely described as a theoretical understanding of classical statistical models and methods, and the ability to apply those models effectively in applications.

Here are the number of questions broken down by program learning objectives and courses:

Module	**Course**	**PLO1**
Bootcamp	MSDS 504: Review of Probability and Statistics	19
Fall 1	MSDS 601: Linear Regression Analysis	13
Fall 2	MSDS 604: Time Series Analysis	8
Intersession	MSDS 629: Experiments in Data Science	19
Total		

Note:

All exam questions used for this assessment report can be found in the Appendix section.

Rubrics

We decided on a scoring mechanism that could be used across all courses and all questions on those exams used in this report.

Score	Description
3	Student has mastered the material necessary to answer a specific question
2	Student did not give a perfect answer to the question but had a solid grasp on the concepts
1	Student did not achieve a minimum level of competency for a specific question
0	Student did not answer

For the purposes of this assessment report, faculty returned to their exams, sometimes months after initially grading the students, to select and score assessment questions as 3, 2, 1, or 0.

Description of Results

The averages of PLO 1 for all four courses ranged from 2.54 to 2.82 with the distribution in table 1.

Module	Course	No. of Students Assessed	Average Score
Bootcamp	MSDS 504: Review of Probability and Statistics	102	2.75
Fall 1	MSDS 601: Linear Regression Analysis	94	2.54
Fall 2	MSDS 604: Time Series Analysis	88	2.82
Intersession	MSDS 629: Experiments in Data Science	87	2.66
Total			

Table 1: Average Assessment Score of Four Courses related PLO 1.

1. What Do We Want Students to Learn?

The PLO 1 aims to develop a solid theoretical foundation in classical statistical models, including generalized linear models and time series models, with an emphasis on practical application. Specifically:

- **Conceptual Mastery**: Understanding core statistical theories and model frameworks, as covered in foundational courses like Probability and Statistics, Linear Regression, and Time Series Analysis.
- **Analytical and Applied Skills**: Using this theoretical knowledge to design and analyze experiments (e.g., A/B testing) and to develop effective forecasting models, as practiced in Experiments in Data Science and Time Series Analysis.

2. Are Students Learning These Skills?

Based on the evaluation scores, students are achieving these objectives at a consistently strong level:

- **High Mastery in Time Series Analysis (2.82)**: This indicates a particularly robust understanding and application of time series forecasting models, supported by hands-on assignments.
- **Solid Performance in Probability and Statistics (2.75)**: Students are successfully reviewing and applying foundational statistical concepts, which is essential for building advanced skills in linear regression and experimental design.
- **Slightly Lower Scores in Linear Regression (2.54)**: Though still above average, this suggests a need for reinforcing linear model applications or reviewing more complex aspects such as interaction terms or model diagnostics.
- **Competent Understanding in Experiment Design (2.66)**: Students demonstrate a sound grasp of A/B testing and other experimental design methods but may benefit from additional practical exercises.

3. How Do We Know They Are Learning It?

The assessment results are derived from sets of questions specifically targeting the key theoretical and applied skills taught in each course. The questions reflect critical topics within each course:

- **MSDS 504 (Probability and Statistics)** questions cover core statistical principles, hypothesis testing, and Bayesian estimation, showing that students can reliably apply these foundational skills.
- **MSDS 601 (Linear Regression Analysis)** focuses on regression diagnostics and variable selection techniques, with room for reinforcing certain complex techniques.
- **MSDS 604 (Time Series Analysis)** assesses forecasting methods like ARIMA and model selection, demonstrating students' high proficiency in this area.
- **MSDS 629 (Experiments in Data Science)** evaluates experimental design methods, confirming that students understand and can apply these principles effectively.

Conclusion

The program learning outcome is largely met, with students showing strong theoretical knowledge and practical application skills across statistical models. The slight variation in scores suggests some focus areas for continued improvement.

According to Fortune, even though the number of job postings in the tech industry has reached a record low this year for the third consecutive year, MSDS managed to achieve an 80% employment rate by the end of October, within three months of graduation (Note: Our students were awarded their degrees in August.). The median base salary was $128K, with a maximum base salary of $190K. Since data science interviews involve multiple stages, including technical rounds that require strong theoretical skills, the high employment rate indirectly reflects the outcome of PLO 1.

Discussion

The cohort assessed this year in the MSDS program largely achieved a high standard, with solid to mastery grades. This success can be attributed to the efforts of our faculty in providing high-quality education and adapting the curriculum to equip our students with the skills that the industry demands.

The evaluation of PLO 1 was conducted by two faculty members from four courses that emphasize statistics. Each instructor was asked to compile the questions and anonymize the students' answers in a shared folder for review. The final report was also shared with all other faculty members in the program to ensure an accurate assessment of learning objectives and outcomes. Going forward, the program has decided to incorporate a set of questions/assignments in each course to assess student learning.

The last evaluation on a similar PLO was completed in 2016, before the new PLOs and assessment methods were proposed in 2018. We will explore better metrics and methods for assessing students, including the development of new rubrics that can provide a more detailed evaluation of student learning.

During our curriculum meeting, we will strategize ways to balance the number of questions in each PLO to promote a more balanced learning experience and evaluate student progress. We may also consider consulting with FDCD, the Center for Teaching Excellence (CTE), and Educational Technology Services (ETS) to align the PLOs with each course.

MSDS constantly faces the challenge of the shortage of tenured faculty members, with only one currently in such positions. This places additional burdens on junior faculty members who have to take on administrative duties and lack of tracking on the program development overtime. Additionally, the program faces resource constraints, especially in terms of space at the downtown campus for our students to study and work.

While the overall salary in the tech industry has increased, the faculty salary has not been increased accordingly. We believe that faculty salaries should be higher than the median base salary of our new graduates to ensure retention. Many faculty members in the program left the industry for their passion, but this affects their quality of life and the lives of their families, which can affect retention.

Unlike faculty members on the main campus who have access to the gym and reasonably priced parking, our faculty and students do not have these amenities. We also lack support from most services on the main campus, like SDS proctoring support. This can pose challenges for recruitment and retention.

Appendix: Questions

MSDS 601

1. True or False: In the case of a multiple regression model, a rejection (p-value<=0.05) in the global F test determines that every coefficient is different from zero. (87%)
2. True or False: For the same MLR model with p>1, the R^2 value is always less than adj_R^2. (86%)
3. Which of the following is NOT a potential symptom of severe multicollinearity in a data? (89%)
4. True or False: If extremely influential outlying cases are detected in a data set, simply delete all those cases from the data set. (95%)
5. For the dataset \textcolor{blue}{KelleyBlueBookData.csv}, response= price against the following predictors: mileage, type, cylinder, liter, cruise, sound, and leather. Treat Leather (0 for not-leather, 1 for yes-leather), Type and Cylinder as categorical variables. Run partial anova, and interpret the F test result for the Cylinder. Combine the results of t test and F test for Cylinder, should we conclude that it's a significant predictor? (0.97)

6. A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:
   - price: price in US dollars
   - carat: the weight of the diamond (0.2–5.01)
   - cut: quality of the cut with 5 levels: Fair, Good, Very Good, Premium, Ideal
   - color: diamond color with 7 levels: from D (best) to J (worst)
   - clarity: a measurement of how clear the diamond is with 8 levels (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
   - x: length in mm (0–10.74)
   - y: width in mm (0–58.9)
   - z: depth in mm (0–31.8)
   - depth: total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)

   If I regress "price" against carat, cut, and depth, how many parameters are estimated in the regression model?

7. Continued from the previous question: If the part of the ANOVA(Typ=2) table is as below:

	SS	df	F	PR(>F)
Carat		?		0.01
Cut		?		0.000
Depth		1		0.04
Residuals				

(a) What is the df in the cell in the row of "Carat"?

(b) What is the df in the cell in the row of "Cut"?

(c) What is the H0 model and H1 model for the F test for "Cut"? (You can write the simple formula version of Y~X1+X2....). How does the p value suggest the significance of "Cut" with alpha=0.05?

8. For the dataset KelleyBlueBookData.csv, response= price against the following predictors: mileage, type, cylinder, liter, cruise, sound, and leather. Treat Leather (0 for not-leather, 1 for yes-leather), Type and Cylinder as categorical variables.

   (a) Report the estimated coefficient of "leather" and interpret the t test result for βleather =0. Combine the results to comment on the impact of "leather" to price.

   (b) By reading the coefficients with "Type", which type was used as the reference level? Which type seems to have the highest average price?

   (c) By reading the coefficient and t test result for "Cylinder=6", what conclusion you can make about the price when Cylinder=6?

   (d) Run partial anova, interpret the F test result for Cylinder. Combine the results of t test and F test for Cylinder, should we conclude that it’s a significant predictor?

Question 1 (Average: 97%)

YouTube is experimenting with two versions of its recommendation algorithm. Two thousand users were randomly selected and then randomly assigned one of the two versions of the algorithm for their session (i.e., 1000 users in each condition). Interest lies in determining whether or not the users choose to watch one of the recommended videos. The data scientists running the experiment are worried about a day-of-week effect and so the experiment is run only on a Wednesday. Ultimately the algorithm that achieves a higher recommendation acceptance rate will be put into full production.

In the sentences below, select from the dropdown menus the word that most appropriately fills in the blank. Note that there is exactly one correct answer for each blank space, and each word is used exactly once.

[*design*, *levels*, *sample size*, *unit*, *nuisance*, *response*, *MOI*, *observational*, *causal*]

(a) The YouTube experiment has two experimental conditions. These conditions are defined by the ________ factor *algorithm version,* which has two ________.

(b) The number 1000 in each condition is called the ________.

(c) Each of the users is considered an experimental ________.

(d) 'Day of week' in this experiment can be thought of as a ________ factor.

(e) The ________ variable, which indicates whether a user watches a recommended video, is binary.

(f) The *recommendation acceptance rate* is the ________.

(g) The benefit of such an experiment, relative to an ________ study, is that it facilitates ________ inference.

Question 2 (Average: 97%)

Consider a pregnancy test analogy with the following hypothesis:

H_0: the person is not pregnant vs. H_A: the person is pregnant

Which of the following corresponds to a Type I Error?

(a) The pregnancy test tells a pregnant person that they are pregnant
(b) The pregnancy test tells a non-pregnant person that they are not pregnant
(c) The pregnancy test tells a non-pregnant person that they are pregnant
(d) The pregnancy test tells a pregnant person that they are not pregnant

Question 3 (Average 84%)

Consider a pregnancy test analogy with the following hypothesis:

$$H_0\text{: the person is not pregnant vs. } H_A\text{: the person is pregnant}$$

In the context of such a hypothesis, the power of the test is:

(a) The probability that the pregnancy test tells a non-pregnant person that they are pregnant.
(b) The probability that the pregnancy test tells a pregnant person that they are not pregnant.
(c) The probability that the pregnancy test tells a non-pregnant person that they are not pregnant.
(d) The probability that the pregnancy test tells a pregnant person that they are pregnant.

Question 4 (Average: 80%)

Suppose an A/B test is performed resulting in the following data summaries:

- n_1 = 345; $\hat{\mu}_1 = \bar{y}_1$ = 108; $\hat{\sigma}_1 = s_1$ = 18
- n_2 = 345; $\hat{\mu}_2 = \bar{y}_2$ = 112; $\hat{\sigma}_2 = s_2$ = 15

Now suppose we wish to test the following null hypothesis:

$$H_0: \mu_1 = \mu_2 \text{ (assuming } \sigma_1 \neq \sigma_2\text{)}$$

Calculate the appropriate test statistic and enter its value into the box below. Be sure to round to four decimal places.

□

Question 5 (Average: 86%)

Consider the following hypothesis and suppose the appropriate Z-test statistic t is calculated. In the context of this hypothesis, what values of t are considered "extreme" and would give us evidence against H_0?

$$H_0: \pi_1 \geq \pi_2 \text{ vs. } H_A: \pi_1 < \pi_2$$

(a) Large positive numbers (i.e., $t \gg 0$)
(b) Large negative numbers (i.e., $t \ll 0$)
(c) Both of the above (i.e., $t \gg 0$ and $t \ll 0$)
(d) Values close to zero (i.e, $t \approx 0$)

Question 6 (Average: 86%)
Suppose that you are determining the sample size required to test a hypothesis that has a significance level of α and a power of $1-\beta$, and that is designed for a minimum detectable effect of δ. For each part below, fill in the blank space by choosing the appropriate word from the dropdown menu.

[*increase*, *decrease*]

(a) All else being equal, *increasing* δ will ________ the required sample size.
(b) All else being equal, *increasing* α will ________ the required sample size.
(c) All else being equal, *increasing* $1-\beta$ will ________ the required sample size.

Question 7 (Average: 73%)

Consider an experiment with two conditions (A vs. B) that gives rise to the following data:

A	B
{1,2}	{3,4}

And suppose interest lies in testing the following hypothesis:

$$H_0: \mu_A = \mu_B \text{ vs. } H_A: \mu_A \neq \mu_B$$

The 6 possible unique rearrangements of the data are shown in the table below.

A	B
{1,2}	{3,4}
{1,3}	{2,4}
{1,4}	{2,3}
{2,3}	{1,4}
{2,4}	{1,3}
{3,4}	{1,2}

Calculate the p-value associated with an exact permutation test of the hypothesis stated above. Round your answer to 4 decimals.

Question 8 (Average: 95%)

Suppose we conduct an experiment with a single design factor at 3 levels (and hence three conditions). We analyze the results with the following linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where:

- $x_{i1} = 1$ if unit i is in condition 1, and 0 otherwise.
- $x_{i2} = 1$ if unit i is in condition 2, and 0 otherwise.

With the observed data we estimate the regression coefficients as $\hat{\beta}_0 = 3.1; \hat{\beta}_1 = 1.8; \hat{\beta}_2 = -0.6$. Using these values, estimate the expected response in each of the three conditions. Be sure to round your answers to 1 decimal place.

- $\hat{\mu}_1 =$ ________
- $\hat{\mu}_2 =$ ________
- $\hat{\mu}_3 =$ ________

Question 9 (Average: 95%)

The *observed* 2x2 contingency table associated with a χ^2-test of independence is shown below.

	Condition 1	**Condition 2**	
y = 0	10	20	30
y = 1	40	130	170
	50	150	200

Fill in the missing cells for the *expected* 2x2 contingency table below. Be sure to round to 1 decimal place.

	Condition 1	**Condition 2**	
y = 0			30
y = 1			170
	50	150	200

Question 10 (Average: 88%)
Suppose that $M = 10$ independent hypothesis tests are performed, each at significance level $\alpha = 0.02$. Calculate the family-wise error rate in this case. Round your answer to four decimals.

Question 11 (Average: 92%)
In class we've discussed three methods for dealing with the *multiple comparison problem*. These methods vary in their likelihood of committing Type II Errors. Match each of the methods with the correct statement concerning their relative power.

Bonferroni's Method	Least powerful
Holm's Method	Somewhere in between
Šidák's Method	Most powerful

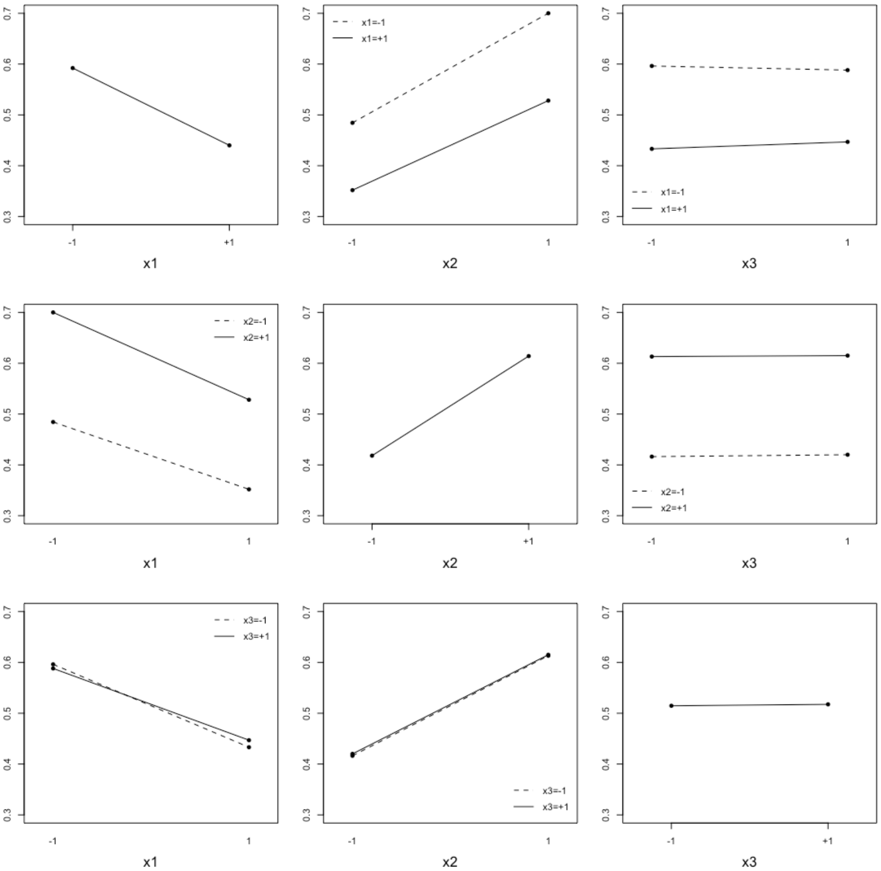
Question 12 (Average: 77%)
Suppose that a factorial experiment with $m = 64$ conditions, and $n = 100$ units in each condition, is conducted. Suppose also that these conditions arose by considering all possible combinations of the levels of factors A (4 levels), B (4 levels) and C (4 levels). Interest lies in determining whether the A:B:C interaction effect is significant. If the response is binary, a likelihood ratio test with a χ^2 null distribution would be used to determine the significance of the A:B:C interaction. In the box below, enter the degrees of freedom associated with this test.

Question 13 (Average: 91%)
Suppose that a factorial experiment with $m = 64$ conditions, and $n = 100$ units in each condition, is conducted. Suppose also that these conditions arose by considering all possible combinations of the levels of factors A (4 levels), B (4 levels) and C (4 levels). Interest lies in determining whether the A:B:C interaction effect is significant. If the response is binary, a likelihood ratio test with a χ^2 null distribution would be used to determine the significance of the A:B:C interaction. In such a likelihood ratio test, which of the following test statistic values would provide evidence to suggest that the A:B:C interaction *is not significant*?

(a) Positive values of t very close to 0.
(b) Positive values of t very far from 0.
(c) Both of the above.

Question 14

A 2^3 factorial experiment was performed to evaluate the influence of three factors on a click-through-rate. The main effect and two-factor interaction plots are shown below. Note that the binary variables x_1, x_2, x_3 respectively represent factors A, B, C.



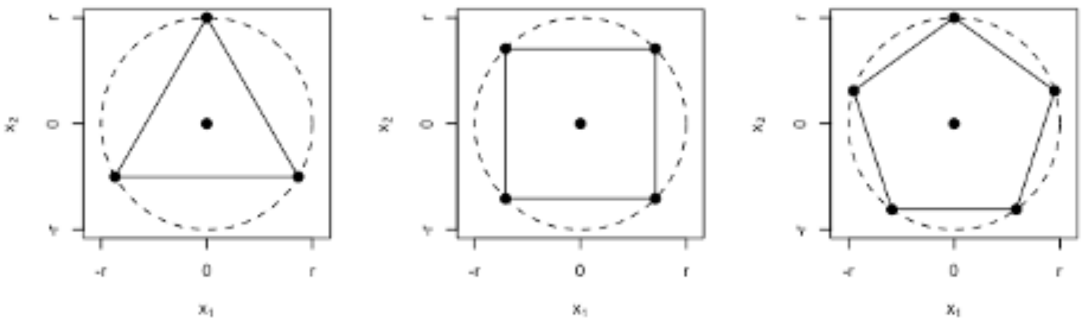
(a) Which factor's effect is the smallest, A, B, or C? **(Average: 100%)**

(b) Does there seem to be a strong A:C interaction effect? **(Average: 90%)**

(c) Does there seem to be a strong B:C interaction effect? **(Average: 94%)**

Question 15 (Average: 82%)

In class we discussed central composite designs as one particular type of response surface design. Another class of response surface designs are *equiradial designs*. An equiradial design of radius r that explores K' factors has 1 center point condition and n_s spherical conditions equidistant from the center point. When $K' = 2$, the spherical design points are dispersed on a circle of radius r and the designs constitute regular polygons. Examples for $n_s = 3,4,5$ are shown below



TRUE or FALSE: A central composite design with $a = \sqrt{2}$ is an equiradial design

Question 16 (Average: 83%)

Suppose that a response surface design is performed to investigate the relationship between a response variable and two factors, and the resulting data yields the following estimated second order linear predictor

$$5 - 2x_1 + 8x_2 + x_1^2 + 4x_2^2$$

Find the *stationary point* of this surface and input the x_1 and x_2 coordinates in the box below.

Question 17 (Average: 94%)

TRUE or FALSE: The optimal factor values, as determined by a second order response surface model, always correspond to one of the experimental conditions considered in the response surface design.

MSDS 504 Probability and Stats

1. Let $\mathbb{P}$ be a probability measure defined on a sample space S. For an event A from S, define $\mathbb{Q}(A) = \mathbb{P}(A)^2$ and $\mathbb{R}(A) = \mathbb{P}(A)/2$. Is $\mathbb{Q}$ a probability measure on S? Is $\mathbb{R}$ a probability measure on S? Why or why not?

2. Let A and B be two events. Use your own words and examples to illustrate the below questions

   (a) (99%)if the events A and B are mutually exclusive, are A and B always independent? if no, can they ever be independent? Explain.

   (b) (99%)if the events A and B are independent, are A and B always mutually exclusive? if no, can they ever be mutually exclusive? Explain.

   (c) (99%)if $A \subset B$, can A and B ever be independent events? Explain.

3. Let the random variable X have the p.d.f.

$$f(x) = \frac{x+1}{2}, -1 \leq x \leq 1$$

   (a) Sketch the graph of this p.d.f.

   (b) Find the c.d.f. of X and sketch the graph of the c.d.f

   (c) Find the expected value $E(X)$.

   (d) Find

      i. $\mathbb{P}(0 \leq X \leq 1/2)$

      ii. $\mathbb{P}(X \geq 3/4)$

      iii. $\mathbb{P}(X = 3/4)$

      iv. $q_1 = \pi_{0.25}$ (25% quantile, aka $\mathbb{P}(X \leq q_1) = 0,25$)

4. Let's use some True of False questions to summaries some results we have learned from the lecture and this homework. For each statement below, choose True or False; if False, briely state why.

   (a) Expected value is a linear operator for multiple random variables X_1, X_2, ...,X_n i.e.

$$E(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i E(X_i)$$

   .

(b) If X and Y are independent, they are uncorrelated;

(c) If X and Y are uncorrelated, they are independent;

(d) If X is normally distributed, a linear transformation of X (i.e. $aX + b$) is still normally distributed.

5. If X_1 and X_2 are independent with respective means μ_1, μ_2 and variances σ_1^2 and σ_2^2, and $Y = a_1X_1 + a_2X_2$

   (a) Prove $E(Y) = a_1\mu_1 + a_2\mu_2$

   (b) Prove $Var(Y) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2$. HINT: use $Var(Y) = Cov(Y, Y)$ and properties of covariance.

   (c) If X_1 and X_2 are **uncorrelated** instead of independent, are (a) and (b) still true?

   (d) If If X_1 and X_2 are **correlated**, are (a) and (b) still true?

6. This question will help you summarise all the distributions related to normal distribution. Let X_1, ..., X_n be a random sample of size $n = 16$ from $X \sim N(0, 1)$, and Y_1, ..., Y_m be a random sample of size $m = 9$ from $Y \sim N(3, 4)$. X and Y are independent, i.e. any X_i and Y_j are independent for $i = 1, ..., n$ and $j = 1, ..., m$.

   (a) What is the distribution of $\bar{X}$?

   (b) What is the distribution of $\bar{Y}$?

   (c) What is the distribuiotn of $\bar{X} - \bar{Y}$

   (d) Find the constant a and b such that $\frac{aS_x^2}{b}$ has a chi-square distribution. What is the degree of freedom of this chi-square distribution? What is $E(S_x^2)$?

   (e) Find the constant c and d such that $\frac{cS_Y^2}{d}$ has a chi-square distribution. What is the degree of freedom of this chi-square distribution? What is $E(S_Y^2)$?

   (f) What is the distribution of $\frac{\bar{X}}{S_x/\sqrt{16}}$?

MSDS 604 Time Series Analysis

---

1. In the lecture notes, we have proved that the Moving Average process with order 1, or $MA(1)$, is stationary. If the Moving Average process with order q, or $MA(q)$ is defined as

   $$X_t = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}, \{Z_t\} \sim WN(0, \sigma^2)$$

   Prove that $MA(2)$ is also stationary, by finding the mean, variance, autocovariance and autocorrelation functions.

2. For the time series given below,

   $$x_0 = 4, x_1 = 6.1, x_2 = 5.2, x_3 = 6.5, x_4 = 8.9, x_5 = 8, x_6 = 8.2, x_7 = 11.4, x_8 = 10$$

   use the classical decomposition method to calculate by hand, the estimates of trend component m_t, seasonal component s_t and random noise ϵ_t, using a seasonal lag $h = 3$ and an additive model. (Hint: for the observations that can not be estimated by this method, leave it blank; choose the centered MA filter with order =3)

3. For each of the AR processes below, determine whether they are stationary by checking the stationary condition. In each case $\{Z_t\} \sim WN(0, \sigma^2)$

   (a) $X_t = -0.2X_{t-1} + 0.48X_{t-2} + Z_t$

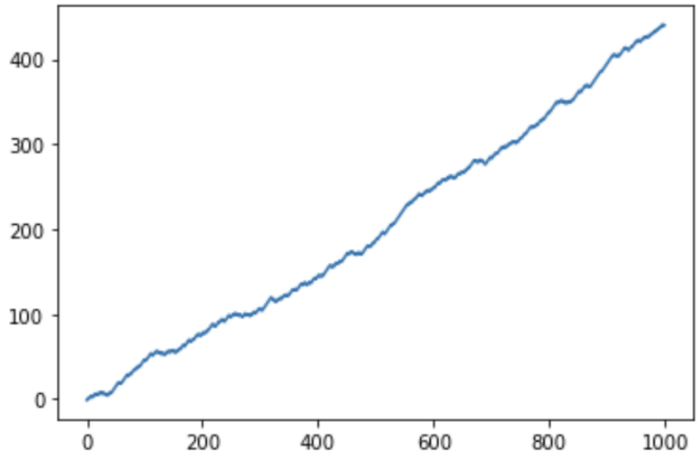
   (b) $X_t = X_{t-1} - 0.8X_{t-2} + Z_t$

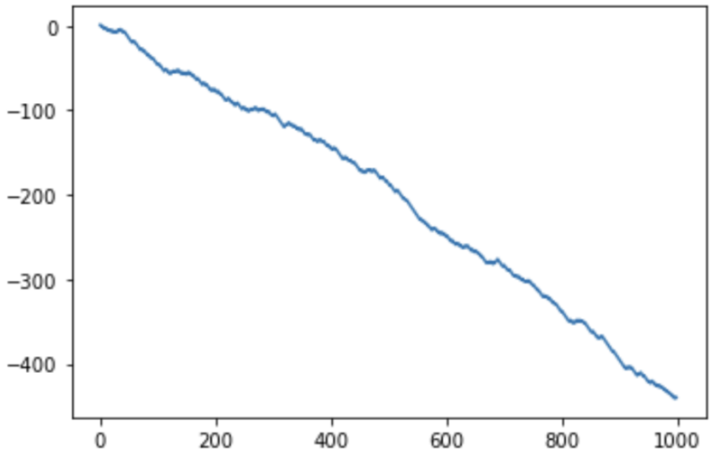
4. True or False: If $\{X_t\}$ is stationary, then $(1 - B)X_t$ (difference once) is still stationary.

5. You are given the history of monthly revenue data of a retail company from the past ten years in quiz2_history.csv Download quiz2_history.csv, and your goal is to provide a forecast of the monthly revenue for the following year.

   Use the historical data to select and fit a model and provide the forecast for the next 12 months. You can select mode from the methods of AIC/BIC, train-test split, cross-validation, or even manual differencing+plots. You are not limited to which selection method to choose or how you design the method.

6. True or False: Based on the **time series plots** below, the two time series data will have almost opposite (mirror-like) patterns in ACF plots.





A True

B False

7. The dataset *profit.csv* recorded the profits (in $*k*) of an investment product in 200 days (positive number shows increased price compared to original price, negative number shows dropped price from original price). Our goal is to give 3-step forecast $\hat{x}_{n+1}$, $\hat{x}_{n+2}$ and $\hat{x}_{n+3}$. For the given data, leave out the last 3 observations (index 197-199) as your test set (don't touch it until you are told to do so), and use the rest data (index 0-196) as history for ARMA model selection. For all questions, use max p, q=5. Based on the forecast goal, design a 5-fold cross-validation to select the order of ARMA model using the history dataset. In the answer, attach a screenshot of the definition of the python function for cross-validation, and fill out the table below. Hint: in method 2, you need to update the data to fit the model at each step to generate the forecast in the next step.

	1-step	2-step	3-step	avg of 3 steps
Best-RMSE				
Best-Model (p,q)				