DATS 6313 **Name (Print):** ______________________
Time series Analysis & Modeling
Summer 2022
Exam #2
6/24/2022
Time Limit: 120 Minutes

This exam contains 7 pages (including this cover page) and 4 problem(s). Check to see if any pages are missing. Enter all requested information on the top of this page, and put your initials on the top of every page, in case the pages become separated.

- Take a screen shot of figures/tables on the console and add to your solution manual. The .pdf version of the solution manual will only be graded.
- Submission after the deadline will be disregarded .
- Submit the supporting .py file that regenerates the results in the solution manual. A .py with error will subject to 50% penalty.
- This is a closed book test and you must *not* use your books, notes on this test.
- You should used python & pycharm IDE.
- You are *not* allowed to use cell phone during the test (except for uploading).
- This is an individual test and collaboration is *not* allowed.
- The exam is out of 85 points with 15 *bonus points* . Maximum grade attained is 100.
- Show all your work to receive full credit.

- **Organize your work**, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit.
- **Mysterious or unsupported answers will not receive full credit**. A correct answer, unsupported by calculations, explanation, or algebraic work will receive no credit; an incorrect answer supported by substantially correct calculations and explanations might still receive partial credit.
- If you need more space, use the back of the pages; clearly indicate when you have done this.
- **Question 4 is the bonus** question and weighs 15pts.

Problem	Points	Score
1	22	
2	33	
3	30	
4	0	
Total:	85	

Do not write in the table to the right.

1. Load the ”question1.csv” dataset on the blackboard. The tongue dataset contains 3 columns and 80 rows. The description of each column is a follows:

   - **type**: Tumor DNA profile (1= Aneuploid Tummor, 2 = Diploid Tummor)
   - **time**: Time to death or on-study time, weeks
   - **delta** : Death indicator (0=alive, 1= dead)

   (a) (14 points) Using ”lifelines” package and ”KaplanMeierFitter” write a python program that estimate the survival function for Aneuploid Tumor(type 1 Tumor) and Diploid Tumor (type 2 Tumor). Plot the survival functions for these two Tumor in one graph. Add title, legend x-label and y-label to your plot.

   (b) (4 points) What is the survival rate at week 50, for type 1 Tumor and Tumor 2?

   (c) (4 points) Which Tumor is deadlier? Justify your answer using the survival function graph. .

2. Load the "question2.csv" dataset from the blackboard. This is a non-seasonal time series dataset with 1000 observations $y(t)$. The data is collected daily starting from Jan 1^{st} 1981 . Using python program and the package of your interest, fit the dataset into an ARIMA model. Use all the data as the train set.

   (a) (2 points) Plot the time series dataset versus time. Make sure that the x-axis not be crowded. Add title, legend, x-label, y-label and grid to your graph.

   (b) (3 points) Plot the rolling mean and variance versus time and perform an ADF-Test & KPSS-test. Is the dataset stationary? Explain your answer.

   (c) (3 points) Perform a time series decomposition on the raw data and find the strength of the trend and strength of the seasonality. Is the dataset seasonal or trended? Explain your answer.

   (d) (4 points) If the raw dataset is not stationary, transform it to a stationary dataset and display the ADF-test & KPSS-test. Plot the rolling mean and variance to re-confirm stationarity. Explain your answer.

   (e) (3 points) Using the GPAC code display the GPAC table (7,7) and estimate a potential order for AR and MA. Highlight the potential pattern. Take a screen shot of the highlighted pattern and place it into your solution manual. *Hint: You may need to perform non-seasonal differencing before using GPAC.*

   (f) (3 points) Plot the ACF and PACF of the differenced dataset for 50 lags. Using the 'tail-off' & 'cut-off' scenario estimate the order of AR and MA for this dataset? Does the estimated order confirm the order from the GPAC table? Justify your answer.

   (g) (3 points) Using the estimated order from the previous sections, and using maximum likelihood estimation , estimate the parameter of the ARIMA model. Print the estimated parameters with the corresponding confident intervals. Hint: You can use your own LM algorithm or the Python package.

(h) (4 points) Using the estimated order and estimated parameters derived in the previous sections, develop the forecast function and predict $\hat{y}(t)$ (one-step prediction). Calculate the residual errors. Plot the ACF of residual errors for 20 lags.

(i) (4 points) Plot $y(t)$ versus $\hat{y}(t)$ in one graph for the first 200 samples. Perform a χ^2 test with $\alpha = 0.01$ to verify the accuracy of the derived model. Analyze χ^2 result on the validation of the derived model. Justify your answer. (χ^2 table is attached at the end). What is the final ARIMA model that best represent the data?

(j) (4 points) Calculate the correlation coefficient between $y(t)$ and $\hat{y}(t)$ and display the scatter plot between them. What does the plot and correlation coefficient tell about the accuracy of this prediction?

3. Load the "question3.csv" dataset from the blackboard. This is a seasonal time series dataset with seasonality order of 3. Using python program and the package of your interest, answer the following questions.

   (a) (5 points) Stationarity check: Plot the mean and variance versus time (rolling mean and variance) and perform an ADF-test & KPSS-test. Is the dataset stationary? Explain why.

   (b) (5 points) Perform a first order seasonal differing and check the stationarity by plotting the rolling mean and variance and an ADF-test & KPSS-test. Is the seasonally differenced dataset stationary? Explain why?

   (c) (5 points) Display the GPAC table (7,7) using the seasonally differenced dataset. Highlight the pattern. What is the estimated order of the SARIMA model? Take a screen shot and highlight the pattern. Include the screen shot into your solution manual.

   (d) (5 points) Plot the ACF and PACF of the seasonally differenced dataset for 50 lags. What is the estimated order of the SARIMA model?

   (e) (5 points) Using the estimated order from the previous section, and using maximum likelihood estimation , estimate the parameter of the SARIMA model. Print the estimated parameters with the corresponding confident intervals. Display the SARIMA model. Hint: You need to fed the seasonally differenced dataset into the LM algorithm or the python package for the parameter estimation.

   (f) (5 points) Plot the original raw data versus the seasonally difference dataset. Take a screen shot of the plot and add it to your solution manual.

4. (*15 pts) (Note: Non-python **bonus** question) Let the ARMA(1,1) process to be defined as :

$$y(t) + .25y(t-1) = e(t) + 0.25e(t-1)$$

where $e(t) \sim WN(1,1)$

(a) Write the corresponding fundamental equation of above process.

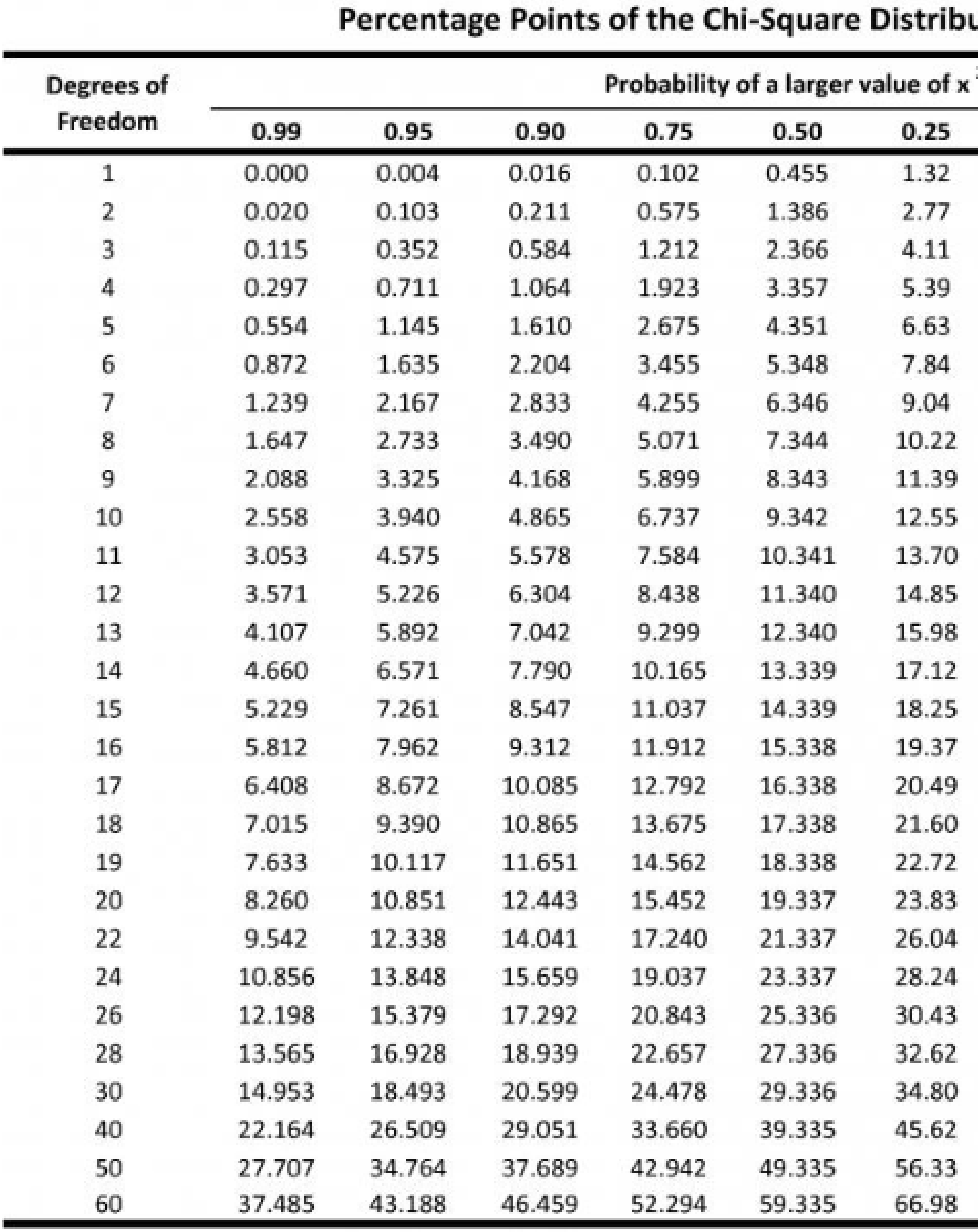
(b) Calculate $\mu_{y(t)}$. Show all your work.

(c) Calculate $\sigma^2_{y(t)}$. Show all your work.

(d) Calculate the theoretical autocorrelation function of $y(t)$ for $(R_y(\tau))$ for $\tau = 0, 1, 2$. Plot the result. Show all your work.

(e) Does this process represent a stationary process? Justify your answer.

(f) Is $y(t)$ correlated with $e(t)$? What is your expectation for the correlation coefficient of $y(t)$ (no need for mathematical calculations)? Justify your answer.



Percentage Points of the Chi-Square Distribu

Degrees of Freedom	Probability of a larger value of x					
	0.99	0.95	0.90	0.75	0.50	0.25
1	0.000	0.004	0.016	0.102	0.455	1.32
2	0.020	0.103	0.211	0.575	1.386	2.77
3	0.115	0.352	0.584	1.212	2.366	4.11
4	0.297	0.711	1.064	1.923	3.357	5.39
5	0.554	1.145	1.610	2.675	4.351	6.63
6	0.872	1.635	2.204	3.455	5.348	7.84
7	1.239	2.167	2.833	4.255	6.346	9.04
8	1.647	2.733	3.490	5.071	7.344	10.22
9	2.088	3.325	4.168	5.899	8.343	11.39
10	2.558	3.940	4.865	6.737	9.342	12.55
11	3.053	4.575	5.578	7.584	10.341	13.70
12	3.571	5.226	6.304	8.438	11.340	14.85
13	4.107	5.892	7.042	9.299	12.340	15.98
14	4.660	6.571	7.790	10.165	13.339	17.12
15	5.229	7.261	8.547	11.037	14.339	18.25
16	5.812	7.962	9.312	11.912	15.338	19.37
17	6.408	8.672	10.085	12.792	16.338	20.49
18	7.015	9.390	10.865	13.675	17.338	21.60
19	7.633	10.117	11.651	14.562	18.338	22.72
20	8.260	10.851	12.443	15.452	19.337	23.83
22	9.542	12.338	14.041	17.240	21.337	26.04
24	10.856	13.848	15.659	19.037	23.337	28.24
26	12.198	15.379	17.292	20.843	25.336	30.43
28	13.565	16.928	18.939	22.657	27.336	32.62
30	14.953	18.493	20.599	24.478	29.336	34.80
40	22.164	26.509	29.051	33.660	39.335	45.62
50	27.707	34.764	37.689	42.942	49.335	56.33
60	37.485	43.188	46.459	52.294	59.335	66.98

Figure 1: χ^2 table