STAT 520 – Fall 2025 – Test 2

Note: For this midterm exam, **you are not allowed to receive help from *anyone except me* on the exams.** For example, you may not talk to other students about the exam problems, and you may not look at other students' exams. **Violations of this policy may result in a 0 on the exam, an F for the course, and/or punishment by the USC Office of Academic Integrity.**

In addition, the data analysis and writing of the mini-reports on this exam must be done entirely by you --- not with the help of any other individual or any AI program such as ChatGPT. You are welcome to use the textbook, course website, and other STAT 520 materials as aids in doing the problems. If you use other background sources (I'm not saying that this is necessary to do so), then you must cite the sources you used. IMPORTANT: To get credit for your answers, you must use the **specific methods, techniques, and notation** that we have studied in this STAT 520 class! Do **not** use time series methods, techniques, and notation that may be out there on the internet but do not reflect the way we have analyzed time series data in OUR class.

Problems 1-3 below involve using R to analyze some real time series. You can run the R code given at the following web site to input the data and turn each vector into a time series object:
`http://people.stat.sc.edu/hitchcock/STAT520DataEntryRcodeTest2Fall2025.txt`

Your mini-reports should be typed in paragraph form and should include relevant graphs where necessary. While you can and should include graphs to supplement your analysis, **please do not clutter up your reports with R code and unedited R output**. For each report, the amount of actual text (not counting plots and graphs) does not need to be more than about one page in length. If you would like to include your R code, please put it in an appendix at the end of all the mini-reports.

Note that there may be several ways to satisfactorily answer these questions. In addition, since these are real data, it is possible that *no* model may *perfectly* describe the time series behavior. Your reports will be graded partly on the quality of the statistical analysis that you do, and partly on your ability to communicate your conclusions clearly and concisely. Specifically, each problem will be worth 20 points, for a total of 60 points:

Writing (out of 10 points): How organized, clearly written, comprehensible, and grammatically correct is the report? Would the client reading this report be confident that it was written by an educated, well-trained statistical scientist?

Analysis (out of 10 points): Were the graphs and data analyses appropriate for the problem? Were the analyses carried out correctly? Were your statistical conclusions about the data set sensible and clearly justified by numerical or graphical evidence?

For each of the following data sets, you will conduct a complete analysis, including model specification, parameter estimation, model checking/diagnostics, assessing model fit, and relevant forecasting. Note that for some data sets, more than one model might be reasonable, so how you provide evidence to justify your choice of model is as important as which specific model you choose. You should consider aspects such as whether the time series process is stationary, and if not, whether it can be made stationary by some procedure, such as differencing. Also consider whether a transformation of the response variable is needed. It is recommended that for each analysis, you write the model equation of the model you chose, with estimated parameters plugged in.

1. Baseball playing styles and equipment have changed over time, but how have pitchers' performances changed, if at all? In this problem, we will analyze the National League's baseball pitching performance over time. The data object `NLWHIP.ts` contains the WHIP values for National League baseball (all National League teams combined) for each year between 1876 and 2015. [WHIP stands for "Walks and Hits per Inning Pitched" and is a measure of pitching performance. The lower the number, the better the pitching performance.] Conduct and summarize a full analysis of the data. Augment your report with relevant graphics or plots, and be sure to comment clearly about what the graphs tell us.

Use your chosen model to obtain forecasts and **90% prediction intervals** for the forecasted WHIP values for the next 10 years: 2016, 2017, …, 2025. About how many of these 10 prediction intervals would you expect to contain the true WHIP value for the corresponding year? Note that in fact, the WHIP values for the major leagues for years 2016, 2017, …, 2025 are: 1.327, 1.351, 1.296, 1.321, 1.332, 1.293, 1.279, 1.332, 1.288, 1.296. For your model and your prediction intervals, how many intervals contained the true value for that year? NOTE: Do not use these forecasts to calibrate your choice of model; the model selection should be done strictly based on the 1876-2015 data.
[Data from baseball-reference.com]

2. Wool sweaters were nearly a necessity for denizens of the British Isles, especially in the years before central heating, requiring a robust population of sheep to shear. The object `sheep.ts` gives the annual sheep population (in millions) of England and Wales, between 1867–1938. Conduct and summarize a full analysis of the data. Augment your report with relevant graphics or plots, and be sure to comment clearly about what the graphs tell us. Use your chosen model to obtain forecasts and **95% prediction intervals** for the forecasted values for the next 3 years, specifically 1939, 1940, and 1941. A hypothetical question you should answer: Based on the observed data through 1938 and your model, find the approximate predicted probability that the sheep population in 1939 will be *more than* 1.8 billion. Similarly, find the approximate predicted probability that the sheep population in 1940 will be *more than* 1.8 billion. Do the relative sizes of these two probabilities make sense?

3. The psychological condition of schizophrenia combined with the treatment of a drug called chlorpromazine can cause a slowdown in thinking among those afflicted by the condition. The object `schiz.ts` contains a set of perception speed scores for a schizophrenic patient, for the 60 days following the decision that the patient would be administered chlorpromazine for the remainder of the study. Conduct and summarize a full analysis of the data. Augment your report with relevant graphics or plots, and be sure to comment clearly about what the graphs tell us. Use your chosen model to obtain forecasts and **95% prediction intervals** for the forecasted perception speeds for the next three days (days 61, 62, 63).

The midterm exam will be due Thursday, November 13 by 11:59 p.m.