Prof. H. Bozdogan	Fall Semester: September 29, 2025
Stat-563: Intro Math Stat	Due on or before Oct. 8, 2025

STAT 563 LAB PROJECT#1

INSTRUCTIONS:

- **SHOW ALL YOUR WORK ON SEPARATE PAGES FOR EACH PROBLEM. Please submit your write up with the source and pdf with computational modules and all your graphs. If you are typing your results in LyX or Latex. You can zip your files and submit your work by uploading to CANVAS under your NAME_LASTNAME_STAT563_PROJ#1_FALL_2025.**
- You can use MATLAB, R, or Python computational platform of your choice.

INTRODUCTION

- Please **SHOW ALL YOUR WORK FOR EACH PROBLEM.**
- Submit your work including the source of your write up, the data and Matlab modules in a ZIPPED folder under your name: LAST_NAME_and_NAME STAT 579 PROJ#1.zip**.**
- Use MATLAB, R, Python, or any other software to carry out your graphical visualization and interpret briefly your findings in a narrative style.

Project Goal

The objective of this project is to apply Maximum Likelihood Estimation (MLE) and Information Criteria (IC) to identify the most appropriate **Probability Density Function (PDF)** to model a given dataset. The simulated data is given in ***Raw_Project_Data.xlsx***. This exercise connects theoretical concepts of estimation and model complexity with computational practice.

Phase 1: Data Acquisition and Preparation

1. **Data Selection:** Obtain a real-world, non-normal, continuous dataset on your own (e.g., time-to-failure data, wait times, income levels). Record the sample size (n).
2. **Exploratory Data Analysis (EDA):**
   - Generate a histogram or use `histfit(Data,10,'kernel')` of the data.
   - Compute descriptive statistics (mean, variance, skewness, kurtosis).
   - Discuss the data's general shape and bounds (e.g., positive support, heavy tails).

Phase 2: Candidate Model Fitting

Select and fit a minimum of **five distinct, non-trivial continuous probability distributions** whose

characteristics match your data's EDA (e.g., Gamma, Weibull, Log-Normal, Inverse Gaussian, Beta).

Computational Steps (for each PDF)

1. **Define Log-Likelihood (log L):** Can you give the analytical expression for the log-likelihood function, log $L(\mathbf{x}|\boldsymbol{\theta})$, where **x** is the data and $\boldsymbol{\theta}$ are the parameters.
2. **Maximum Likelihood Estimation (MLE):** Matlab module provided uses numerical optimizer (e.g., **`optim`** in R or **`fminsearch`**) to find the parameter estimates $\hat{\boldsymbol{\theta}}$ that maximize log L. Provide the maximum value, log L_{max}.
3. **Count Parameters (m):** Record the number of free parameters, m, for the model.
4. **Visualize:** Overlay the fitted PDF (using $\hat{\boldsymbol{\theta}}$) onto the data histogram.

Phase 3: Model Selection and Information Criteria

Use the results from Phase 2 to calculate three information criteria for each of the models considered.

Information Criteria Formulas

The best model is the one that yields the **minimum value** for each criterion.

- **Akaike Information Criterion (AIC):**

$$AIC = -2\log L_{max} + 2m$$

- **Schwarz Bayesian Criterion (SBC/BIC):**

$$SBC = -2\log L_{max} + m\ln(n)$$

- **Information Complexity (ICOMP):** This penalizes both the number of parameters and the complexity of the estimated Fisher Information Matrix (FIM), $\mathcal{F}(\boldsymbol{\theta})$. We will use a simplified form based on the Hessian approximation of the FIM.

$$ICOMP = -2\log L_{max} + C_{1F}(\widehat{\mathcal{F}}^{-1})$$

Where C_{1F} is the penalty based on the eigenvalues (λ_i) of the estimated asymptotic covariance matrix $\hat{\boldsymbol{\Sigma}}_{\hat{\theta}} = \widehat{\mathcal{F}}^{-1}$:

$$C_{1F} = \frac{1}{4(\bar{\lambda})^2}\sum_{i=1}^{m}(\lambda_i - \bar{\lambda})^2$$

where $\bar{\lambda} = \frac{1}{m}\sum_{i=1}^{m}\lambda_i$ is the mean eigenvalue.

Phase 4: Kernel Density Estimation

Do the kernel density estimation of the simulated DGP (Data Generating Process) and your own real data set.
Report the optimal bandwidth value and the information criteria in a table. What do you observe?

Final Report and Discussion

The final report must include:

1. An Introduction and Data Summary (Phase 1).
2. A detailed presentation of the methodology, including the log L function for each model (Phase 2).
3. A comprehensive Results Table summarizing $\log L_{max}$, m, AIC, SBC, and ICOMP for all models fitted.
4. **Conclusion:** Identify the model selected by each criterion. Discuss the consistency (or lack thereof) among AIC, SBC, and ICOMP, and interpret the implications of the chosen model's parameters.

IF YOU HAVE QUESTIONS PLEASE DON'T HESITATE TO ASK ME OR DAWSON!