STAT 563 LAB PROJECT#2

Prof. H. Bozdogan
Fall 2025
Due November 7, 2025

October 29, 2025

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.**

Overview

This project integrates large-sample theory, likelihood-based inference, and resampling techniques in modern computational statistics. You will:

- Examine the logistic distribution and its importance in statistical modeling and machine learning.
- Compute confidence intervals based on Fisher information
- Explore bootstrapping and the distribution of correlation matrices,
- Implement slice sampling for correlation uncertainty.

Q1 Visualization and Interpretation of the Logistic Distribution

Consider the logistic density given by

$$f(x \mid \mu, s) = \frac{\exp\left(-\frac{x-\mu}{s}\right)}{s\left(1+\exp\left(-\frac{x-\mu}{s}\right)\right)^2}, \quad x \in R.$$

(a) Plotting. Plot $f(x \mid \mu, s)$ for several parameter settings:

- $\mu = 0$ with $s = 0.5, 1, 2$ to study the effect of scale.
- $s = 1$ with $\mu = -2, 0, 2$ to study the effect of location.

Overlay these curves on the same axes with a clear legend and color scheme.

(b) Interpretation.

- The location parameter μ shifts the curve horizontally, serving as both mean and median.
- The scale parameter s controls spread: larger s produces heavier tails and flatter curvature.
- The variance of logistic distribution is $\mathrm{Var}(X) = \sigma^2 = \pi^2 s^2/3$.

(c) Comparison with the Normal Distribution. Overlay $f(x \mid 0, 1)$ with the standard normal $\mathcal{N}(0, 1)$. Discuss how the logistic distribution has heavier tails, offering more robustness to outliers.

(d) Discuss the importance in Statistics and Machine Learning.

- The **logistic CDF**

$$F(x \mid \mu, s) = \frac{1}{1 + \exp\left(-\frac{x-\mu}{s}\right)}$$

  defines the canonical *logit link* in generalized linear models.
- In **logistic regression**, this CDF maps linear predictors to probabilities in $(0, 1)$.
- The heavier tails make logistic likelihoods more robust than Gaussian ones.
- In **machine learning**, logistic loss (cross-entropy) and the softmax generalization underpin neural classification models.

Q2 Find the MLEs

Given observations $x_1, \ldots, x_n$, the likelihood function is

$$L(\mu, s) = s^{-n} \prod_{i=1}^{n} \frac{\exp\left(-\frac{x_i-\mu}{s}\right)}{\left(1 + \exp\left(-\frac{x_i-\mu}{s}\right)\right)^2}.$$

and taking logarithms we have the log-likelihood

$$\ell(\mu, s) = -n \log s - \frac{1}{s} \sum_{i=1}^{n} (x_i - \mu) - 2 \sum_{i=1}^{n} \log\left(1 + \exp\left(-\frac{x_i - \mu}{s}\right)\right).$$

Let $z_i = (x_i - \mu)/s$, we obtain the score equations:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{s}\sum_{i=1}^{n} \tanh\Big(\frac{z_i}{2}\Big),$$

$$\frac{\partial \ell}{\partial s} = -\frac{n}{s} + \frac{1}{s^2}\sum_{i=1}^{n}(x_i - \mu) - \frac{2}{s^2}\sum_{i=1}^{n}(x_i - \mu)\frac{e^{-z_i}}{1+e^{-z_i}}.$$

So, the score vector is:

$$S = \begin{pmatrix} \frac{\partial \ell}{\partial \mu} \\ \frac{\partial \ell}{\partial s} \end{pmatrix} = \begin{pmatrix} \frac{1}{s}\sum_{i=1}^{n} \tanh\big(\frac{z_i}{2}\big) \\ -\frac{n}{s} + \frac{1}{s^2}\sum_{i=1}^{n}(x_i - \mu) - \frac{2}{s^2}\sum_{i=1}^{n}(x_i - \mu)\frac{e^{-z_i}}{1+e^{-z_i}} \end{pmatrix}$$

(a) Show that the MLEs $(\hat{\mu}, \hat{s})$ are:

$$\sum_{i=1}^{n} \tanh\Big(\frac{X_i - \hat{\mu}}{2\hat{s}}\Big) = 0, \qquad \hat{s} = \frac{1}{n}\sum_{i=1}^{n} |X_i - \hat{\mu}|\, w_i.$$

(b) Fisher information Taking the second partial derivative, the observed information at $(\hat{\mu}, \hat{s})$ is:

$$\mathcal{F}_{\text{obs}}(\hat{\mu}, \hat{s}) = -\nabla^2 \ell(\hat{\mu}, \hat{s}) = -\mathcal{H}(\hat{\mu}, \hat{s}).$$

The expected Fisher information (**per observation**) is

$$\mathcal{F}(\mu, s) = \begin{pmatrix} \dfrac{1}{3s^2} & 0 \\ 0 & \dfrac{1}{s^2}\Big(\dfrac{\pi^2}{3} - 1\Big) \end{pmatrix}, \qquad \mathcal{F}_n(\mu, s) = n\,\mathcal{F}(\mu, s).$$

Generate random numbers from logistic distribution of sample size $n = 200$ observations and write a Fisher scoring iterative algorithm in Matlab, R, or Python to find the MLEs and estimate the parameters (μ, s) of the logistic distribution.

(c) Large-sample Wald intervals By inverting $\mathcal{F}_{\text{obs}}$ or using $\mathcal{F}_n$ and using the generated data in part (b), compute confidence intervals:

$$\hat{\mu} \pm z_{\alpha/2}\sqrt{\frac{3\hat{s}^2}{n}}, \quad \hat{s} \pm z_{\alpha/2}\hat{s}\sqrt{\frac{1}{n(\frac{\pi^2}{3} - 1)}}.$$

Q3 Wald CI and percentile bootstrap CI for the mean

Using the simulated data from Q2 from part (b) compute:

(i) the Wald CI based on $\mathrm{Var}(X) = \pi^2 s^2/3$, and

(ii) a percentile bootstrap CI for the mean.

Compare shapes and widths using the Matlab module provided.

Q4 Analytical and Bootstrap Distribution of the Correlation Matrix

Assume $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^\top \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, with correlation matrix $\mathbf{R} = \mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}$, where $\mathbf{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$.

Analytical Distribution. The sample covariance matrix

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$$

satisfies $(n-1)\mathbf{S} \sim W_d(\boldsymbol{\Sigma}, n-1)$, the Wishart distribution. The induced density of the sample correlation matrix $\mathbf{R}_S = \mathbf{D}_S^{-1/2}\mathbf{S}\mathbf{D}_S^{-1/2}$ is (Muirhead, 1982):

$$f(\mathbf{R}_S) \propto |\mathbf{R}_S|^{(n-d-2)/2}\left|\mathbf{I}_d - \mathbf{R}_S^2\right|^{-(n-1)/2}, \quad \mathbf{R}_S \in \mathcal{R}_d.$$

This is the analytical distribution of the correlation matrix.

(a) Use the given Matlab module bootstrap the sample correlation matrix to estimate the empirical distribution $\hat{\mathbf{R}}^{*(b)} = \mathrm{corr}(\mathbf{X}^{*(b)}), \qquad b = 1, \ldots, B.$ and store these in vectorize unique correlations $\mathbf{r}^{*(b)} = \mathrm{vech}(\hat{\mathbf{R}}^{*(b)})$.

(b) Which probability distribution closely approximates the distribution of the correlations
$\mathbf{r}^{*(b)} = \mathrm{vech}(\hat{\mathbf{R}}^{*(b)})$ as in your Project#1.

(c) Apply **slice sampling** to approximate marginal densities for selected r_{jk}.

(d) Compare bootstrap-based confidence intervals with analytical Fisher-z intervals:

$$z_{jk} = \frac{1}{2}\log\frac{1+r_{jk}}{1-r_{jk}}, \quad \mathrm{Var}(z_{jk}) \approx \frac{1}{n-3}.$$

Discuss the geometry of uncertainty in $\mathbf{R} \in R^{d\times d}$ as revealed by the slice-sampled distribution.

Q5 Interpret your results: Conclusion and Discussion

Grading Rubric (Total 100 points)

Parts	**Description**	**Points**
Q1: Visualization and Interpretation	Plots, interpretation of μ, s, and comparison to Normal	20
Q2: Finding MLEs	Coding and showing results	25
Q3: Wald CI and Bootstrap (mean)	CI accuracy, and discussion	15
Q4a: Correct simulation	Bootstrap simulation	5
Q4b: Approx. pdf of correlation	Fitting different distributions	10
Q4c: Slice sampling (corr matrix)	Correct use, intervals estimates	5
Q4d: Fisher-z intervals	Interval estimates	5
Q5: Conclusion and Discussion	Interpretation and clarity of conclusions	10
Formatting	Labeled plots, clear comments and explanation	5
Total		**100**