WSABI Summer Research Lab **Summer 2025**

Lab 8: Central Limit Theorem and Binomial Confidence Intervals

Instructor: Jonathan Pipping *Author: Ryan Brill*

8.1 NBA Free Throws

8.1.1 Data

NBA free throw data from the 2023-2024 season is stored in `08_nba-free-throws.csv`. Each row is a player-team combination, and each contains the following variables:

- `Player-Team`: the player's name and team
- `G`: the number of games the player played
- `FT-G`: the number of free throws made per game
- `FTA-G`: the number of free throws attempted per game

8.1.2 Your Task

1. Modify the above dataset to include the following variables:
    - `Player`: the player's name
    - `FT`: total number of free throws made
    - `FTA`: total number of free throws attempted
    - `FT%`: free throw percentage across a season
2. Filter the dataset to include only players who shot at least 25 free throws.
3. For each player, plot $\widehat{p} = \text{FT\%}$ (x axis) vs player name (y axis), overlaying the Wald and Agresti-Coull confidence intervals. What do you observe?

8.2 Simulation Study

We want to understand how Wald and Agresti-Coull confidence intervals perform across a range of values of p. To do this, we will simulate data from a binomial distribution with varying values of p and compute the confidence intervals for each.

8.2.1 Your Task

1. Discretize the interval $[0, 1]$ into 1000 equally spaced points.
2. For each value of p and each n in $\{10, 50, 100, 250, 500, 1000\}$, generate n free throws using the binomial distribution with parameter p.
3. Compute the 95% Wald and Agresti-Coull confidence intervals for each value of p and n.
4. Repeat steps 2 and 3 $M = 100$ times each.
5. For each value of p and n, plot the coverage probability of the Wald and Agresti-Coull confidence intervals.
6. For each (n, p) pair, note in what percentage of the M simulations the Wald and Agresti-Coull confidence intervals contain the true parameter p. Plot this **coverage** vs p for each n and interval method (faceted).

8.3 Math Homework

8.3.1 Problem 1

Prove that for

$$S_n = \sum_{i=1}^{n} X_i$$

where $X_i \sim \text{Bernoulli}(p)$, the estimate $\widehat{p} = \frac{S_n}{n}$ is the maximum likelihood estimator for p. This means that it maximizes the probability of observing the data given the parameter, or equivalently,

$$\widehat{p}_{\text{MLE}} = \underset{p \in [0,1]}{\text{argmax}} \mathbb{P}(X_1, \ldots, X_n \mid p)$$

8.3.2 Problem 2

Assume that you place n bets on -110 odds (meaning you bet \$110 to win \$100 each time). Answer the following questions:

1. What is the break-even success percentage for each bet? (i.e. the probability of winning that you need to achieve to make \$0 profit). Call this p_{BE}.
2. Discretize the interval $[p_{BE}, 1]$ into 100 equally spaced points. At each point, find how many bets you'd need to make to be confident that you're actually profitable (ex: that the confidence interval for your true success rate exceeds p_{BE}). Plot this as a function of p.