S P Jain School of Global Management
DUBAI • MUMBAI • SINGAPORE • SYDNEY

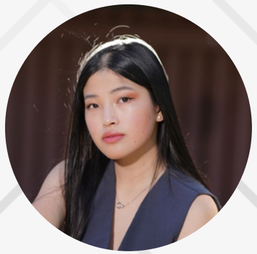
Simulation Modelling of Football First-Goal Scoring Using MOGE Distribution

Group 6

Supervisor: Dr Suchismita Das
Course: Simulation Modelling using Python

Our Team

Ridhi Jain
BS23DMU050



Makhabat Zhyrgalbekova
BS23DSY034

Devanshi Rhea Aucharaz
BJ24DSY005

Naima Dzhunushova
BS23DSY045

Project Abstract

- Study of the MOGE distribution
- EM algorithm for parameter estimation
- Simulation + real data analysis
- Monte Carlo + bootstrapping

Limitations of the Exponential Model

- Assumes constant hazard rate
- Cannot model aging, wear-out, or early failures
- Too restrictive for real lifetime behaviour
- Motivates flexible, multi-parameter alternatives

Need for More Flexibility

- Real systems show complex hazard patterns
- Marshall–Olkin method adds flexibility
- Extra parameter θ modifies hazard shape
- Leads to the more adaptable MOGE model

The Progression of Models

Exponential

Too rigid

Generalized Exponential (GE)

Added shape, more flexible

Marshall–Olkin (MO Extension)

Add shock parameter to adjust failure timing

MOGE

Combines both ideas
→ very flexible
hazard behavior

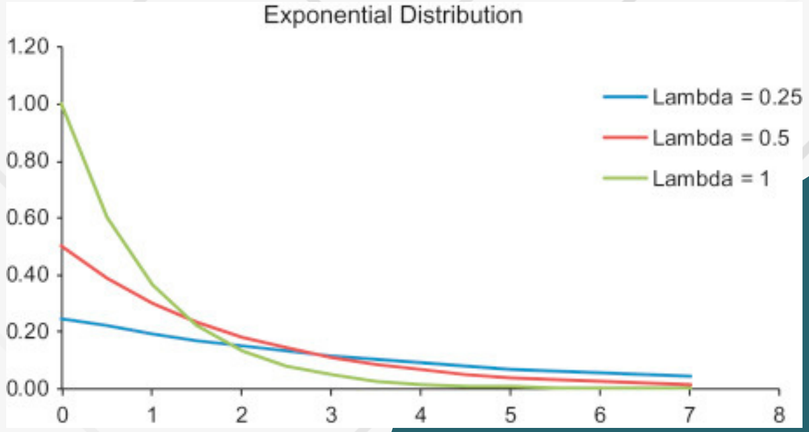
Exponential Distribution

- A basic lifetime distribution used to model “time until an event occurs”
- Assumes events happen randomly at a constant average rate
- PDF:

$$f(x;\lambda) = \lambda e^{-\lambda x}, \qquad x > 0,\ \lambda > 0.$$

- CDF:

$$F(x;\lambda) = 1 - e^{-\lambda x}.$$



- Why it’s popular: Very simple, only one parameter
- Limitations: Constant hazard rate (no increasing or decreasing behaviour, thus not suitable for real-world problems.

Generalized Exponential Distribution (GE)

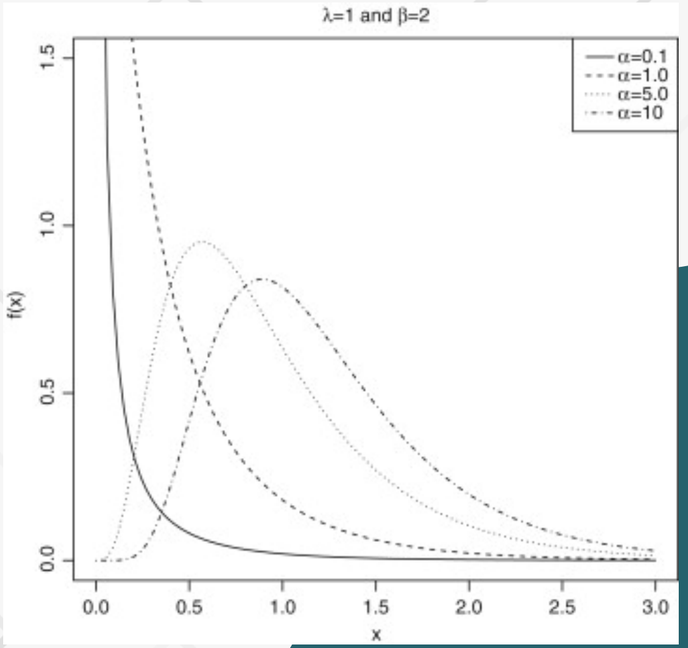
- Extends exponential by adding a shape parameter α
- PDF:

$$F(x;\alpha,\lambda) = \left(1 - e^{-\lambda x}\right)^{\alpha},$$

- CDF:

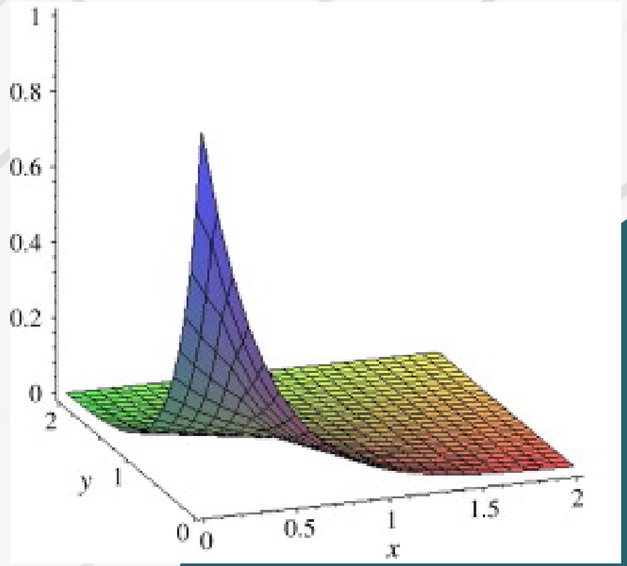
$$f(x;\alpha,\lambda) = \alpha\lambda e^{-\lambda x}\left(1 - e^{-\lambda x}\right)^{\alpha-1}.$$

- Benefits:
  - More flexible than exponential
  - PDF can be increasing or unimodal
  - Better fit for reliability data
- However, still limited: hazard is always monotone.



Marshall–Olkin (MO) Shock Model

- Adds a shock parameter θ to an existing distribution
- Introduces dependence between components
- Models time to failure when external shocks can trigger failure
- MO introduces:
  - Early failure probability increase
  - Extra flexibility in hazard behavior
  - Dependence structure between failure times



The MOGE Distribution

MOGE = MO method + Generalized Exponential distribution

Basic structure

- Extends GE by adding the MO shock parameter θ
- Creates a three-parameter model: α,λ,θ

Key result

- Produces a much more flexible lifetime model

MOGE reduces to:

- GE when $\theta = 1$
- Exponential when $\alpha=1,\theta=1$
- MO-Exponential when $\alpha=1$

Why MOGE Is Powerful

It supports 4 hazard shapes:

- Increasing
- Decreasing
- Unimodal
- Bathtub

Real systems rarely have simple hazard patterns.
MOGE handles:

- Infant mortality
- Wear-out
- Random shocks
- Mixed behaviors

Model Description

- PDF:

$$g(x;\alpha,\lambda,\theta)=\frac{\alpha\lambda\theta e^{-\lambda x}\left(1-e^{-\lambda x}\right)^{\alpha-1}}{\left[\theta+(1-\theta)\left(1-e^{-\lambda x}\right)^{\alpha}\right]^{2}}.$$

- CDF:

$$G(x;\alpha,\lambda,\theta)=\frac{\left(1-e^{-\lambda x}\right)^{\alpha}}{\theta+(1-\theta)\left(1-e^{-\lambda x}\right)^{\alpha}}.$$

- New PDF and CDF allow highly flexible shape behavior
- Includes several special cases
  - Reduces to GE when θ=0
  - Reduces to exponential when α=1, θ=0

Our Goal

Research objectives:

- Understand the distribution
- Derive EM-based parameter estimation
- Evaluate via simulation
- Apply to real data
- Compare with other models

Methodology Overview (EM Algorithm)

Start → E-Step → M-Step → Convergence → Output parameters

- EM is used when likelihood cannot be solved directly
- MOGE has incomplete data → latent variable Z
- EM alternates:
- E-step: compute conditional expectation
- M-step: update parameters

Why EM for MOGE?

MOGE likelihood involves terms like

$$(\theta + (1-\theta)(1-e^{-\lambda x})^{\alpha})^{-2}$$

→ no closed-form MLEs.

Direct optimization of 3D nonlinear system = unstable.

Introducing latent variable Z turns the model into a form that is:

- Easier to differentiate
- Separates parameter updates
- Produces stable EM iterations

EM is not optional but necessary for reliable MOGE estimation.

Parameter Estimation(Log-Likelihood & Score Functions)

Observed-data log-likelihood for MOGE:

$$\ell = n\log(\alpha\lambda\theta) - \lambda\sum x_i + (\alpha-1)\sum\log(1-e^{-\lambda x_i}) - 2\sum\log[\theta+(1-\theta)(1-e^{-\lambda x_i})^{\alpha}].$$

- Score equations for α,λ,θ are nonlinear and cannot be solved analytically.
- This motivates EM and fixed-point formulations.

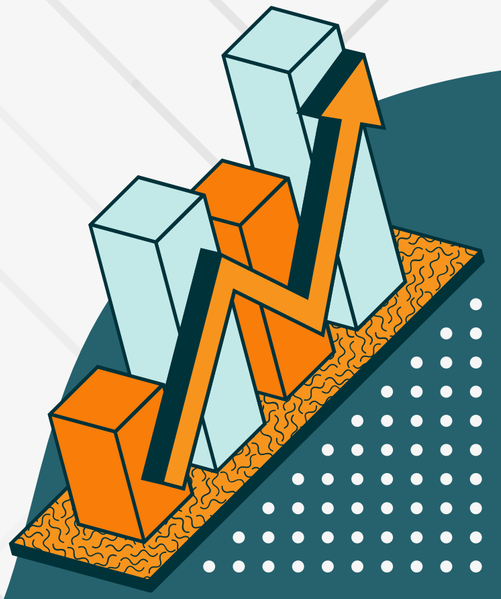
Complete-Data Structure & E-Step

- Introduce latent variable Z → transforms complicated likelihood into a simpler complete-data log-likelihood.
- Conditional expectation:

$$E(Z|X=x)=\frac{2(1-(1-e^{-\lambda x})^{\alpha})}{\theta+(1-\theta)(1-e^{-\lambda x})^{\alpha}}.$$

- E-step: Compute

$$z_i^{(k)}=E(Z|x_i;\alpha^{(k)},\lambda^{(k)},\theta^{(k)}).$$



M-Step & EM Algorithm

- Update for θ:

$$\theta^{(k+1)} = \frac{n}{\sum z_i^{(k)}}.$$

- Compute fixed-point constants c1, c2, from g2.
- Update α,λ using fixed-point equations (20)–(21).
- Iterate E → M until convergence based on log-likelihood change or parameter stability.

Outcome:

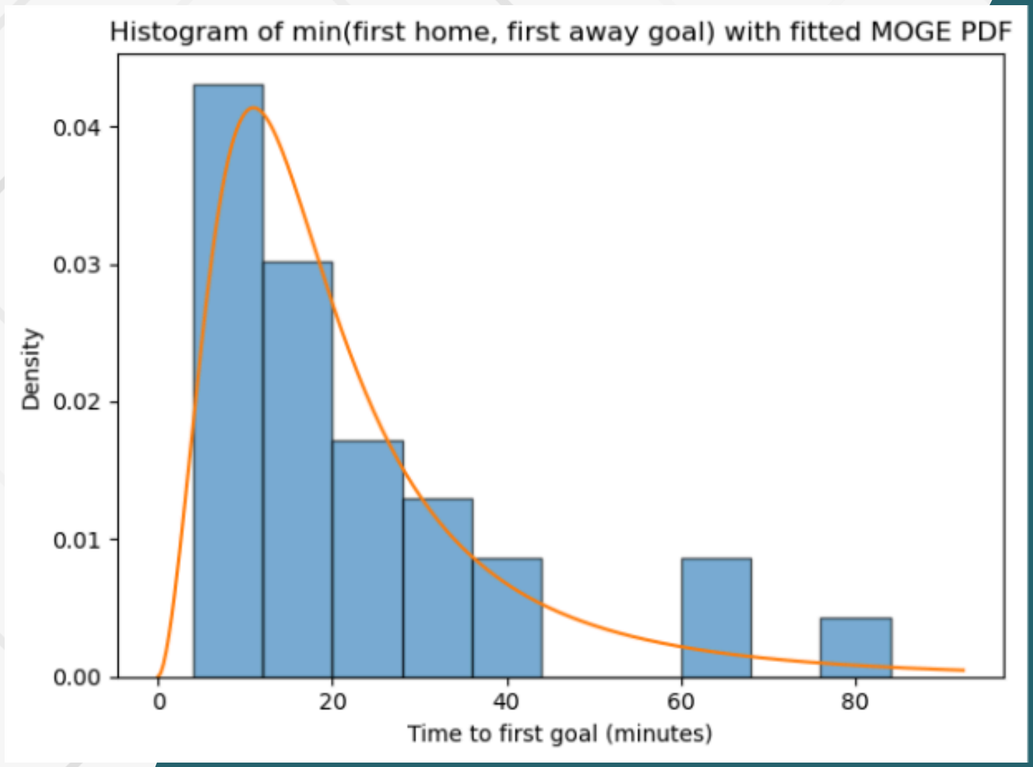
- Efficient EM-based estimation avoids 3D numerical optimization and yields stable parameter estimates for MOGE.

Dataset Used

- Football match data from FIFA Club World Cup (2025) – FootyStats
- Extracted variables:
- X1: First Home goal time
- X2: First Away goal time
- Final cleaned dataset: 29 matches
- Goal times converted to earliest scoring minute

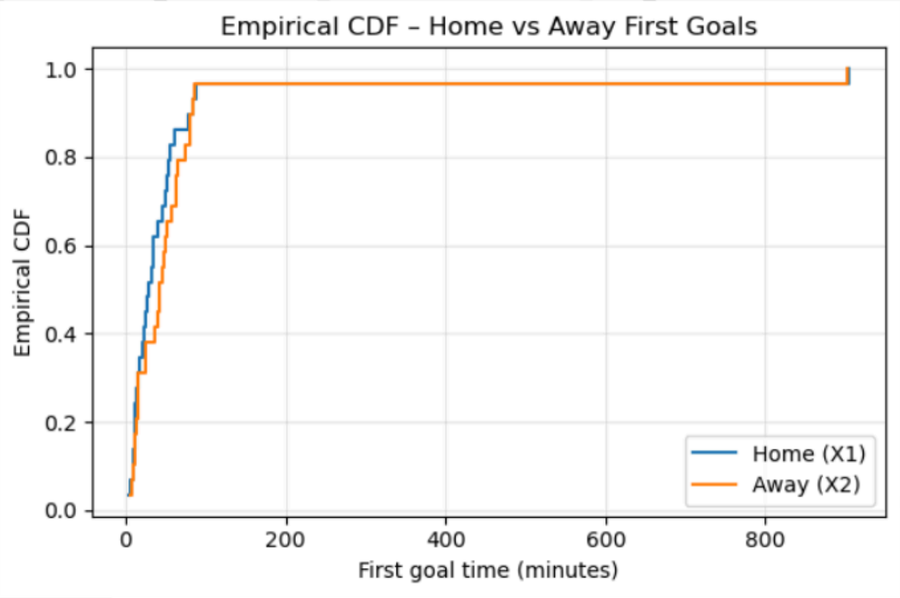
Exploratory Data Analysis

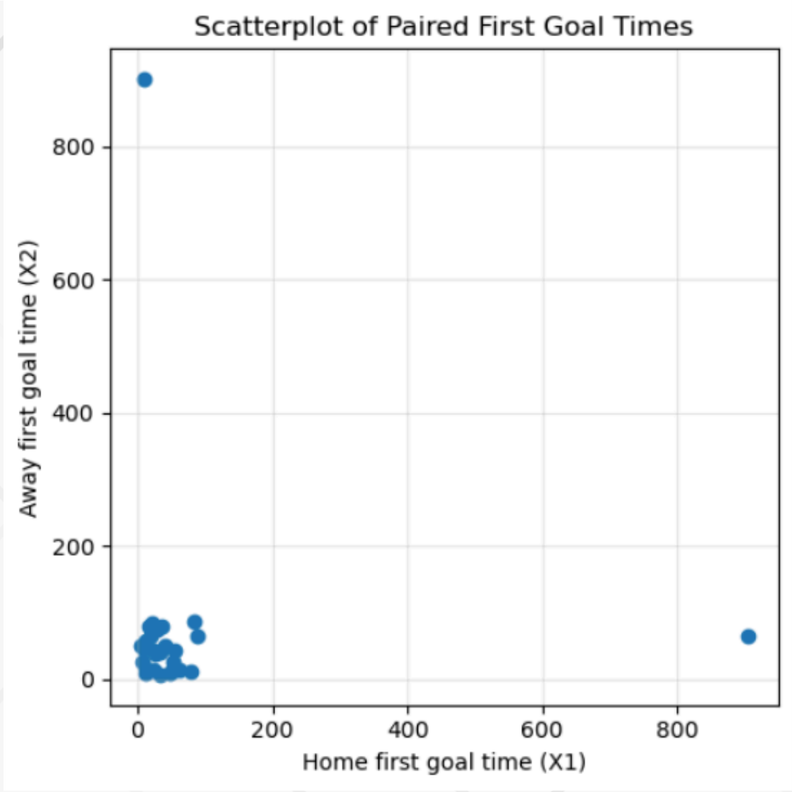
- First-goal times are skewed; rapid early scoring with long-tail cases
- Initial curve comparison suggests heavy-tailed behavior
- Deviations exist in tail → indicates need for model testing
- Later analysis confirms Weibull/Gamma fit better than MOGE



Real Data Analysis

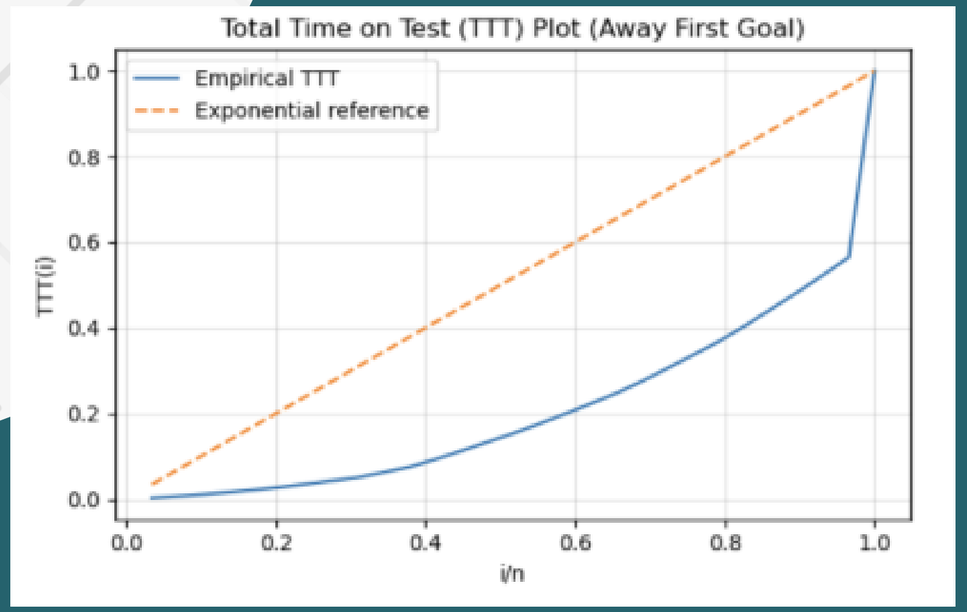
- Both X1 and X2 show strong right skewness → most goals early in the match
- Empirical CDF confirms >50% goals within first 40 minutes
- Weak home-away correlation → timings independent
- Scatter plot: very weak correlation → home & away timings independent





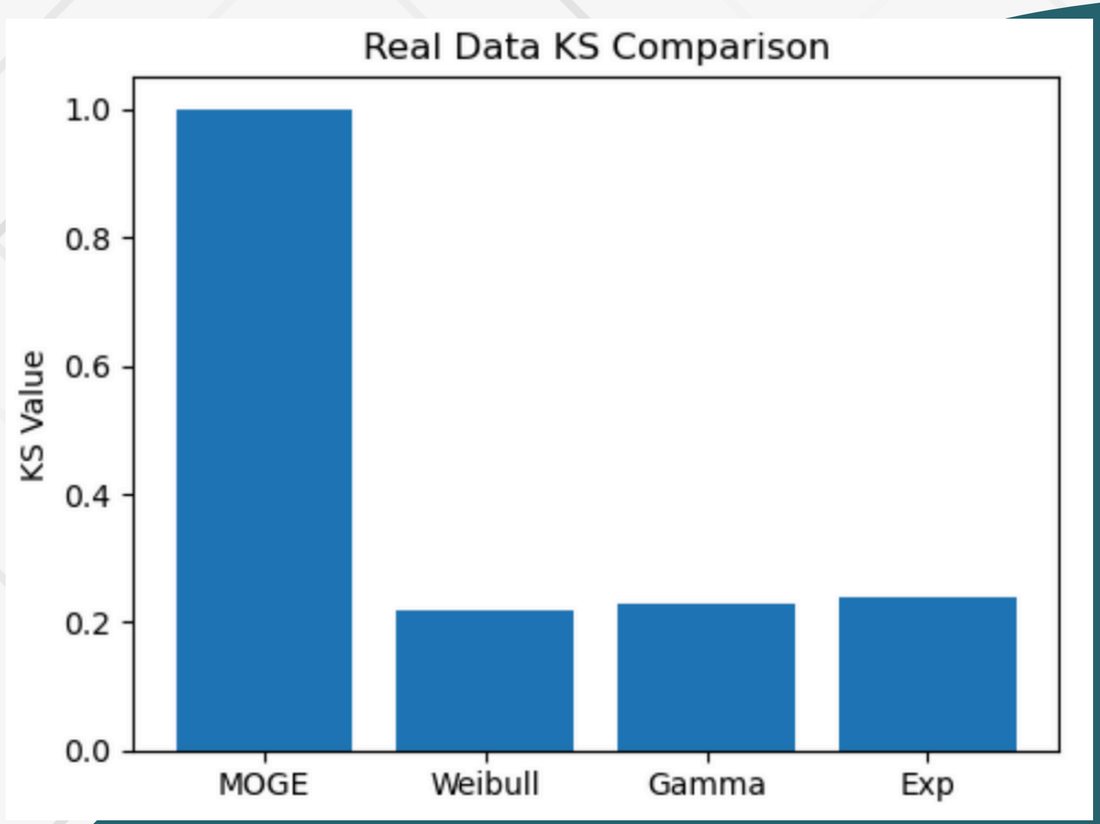
TTT Plot - Hazard Behaviour Analysis

- The empirical TTT curve lies below the 45° exponential reference line
- Early match hazard is low → probability of goal increases later
- Hazard increases → supports Weibull/Gamma suitability
- Confirms the exponential model is unsuitable, motivates Weibull/MOGE modeling



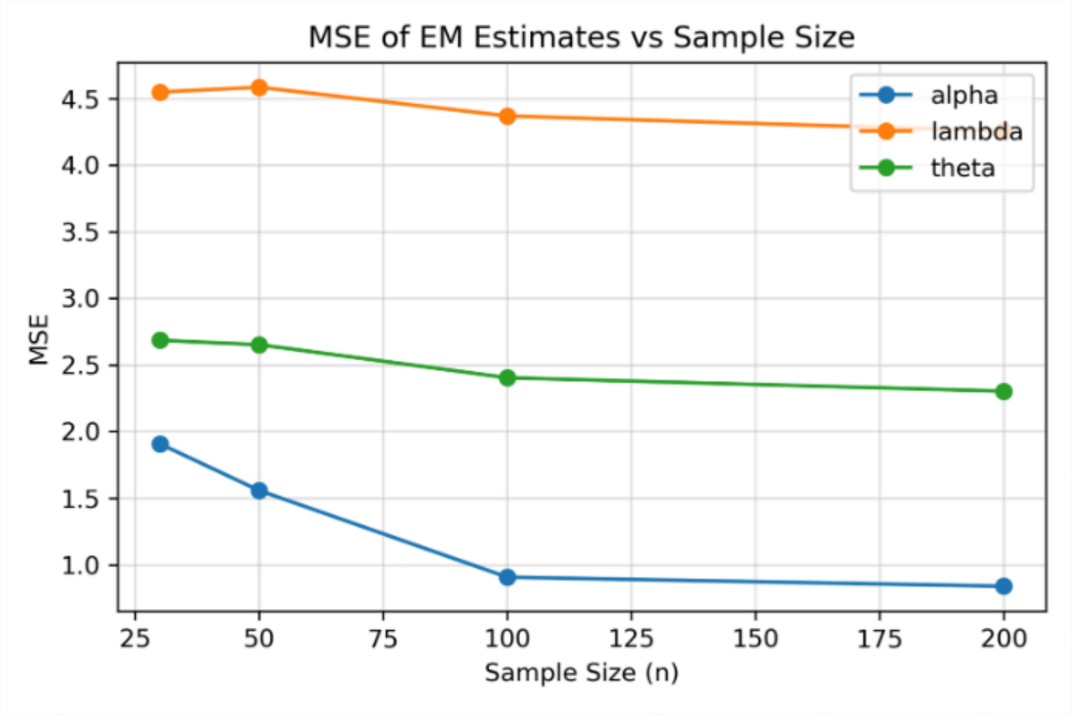
Key Findings

- Real match goal times are irregular and skewed.
- Goal scoring is non-memoryless, supporting flexible hazard models
- Weibull best fits real goal timings, followed by Gamma.
- Exponential borderline. MOGE rejected (KS≈1, p≈0)



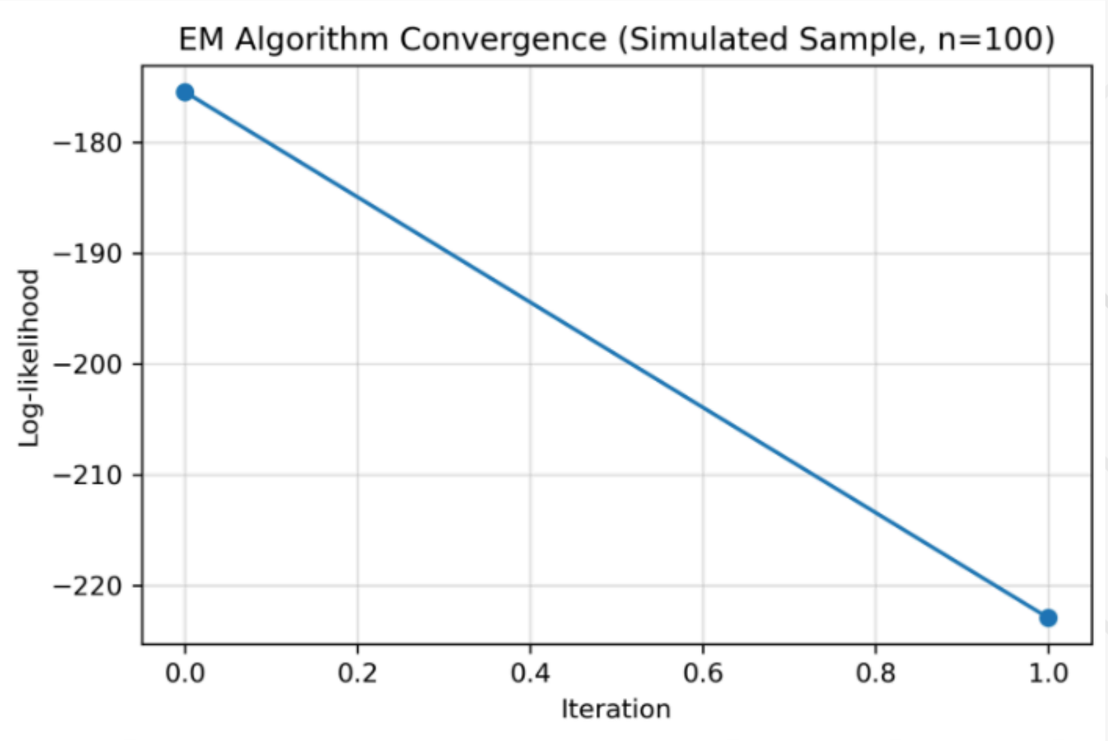
Simulation Study Setup

- Generate synthetic data using MOGE with known parameters ($\alpha=1.5, \lambda=0.8, \theta=1.2$)
- Sample sizes simulated: n = 30, 50, 100, 200
- For each n:
- Fit parameters using EM Algorithm
- Compute Bias & MSE
- Track log-likelihood convergence
- Aim: Check convergence, stability, and accuracy of estimates



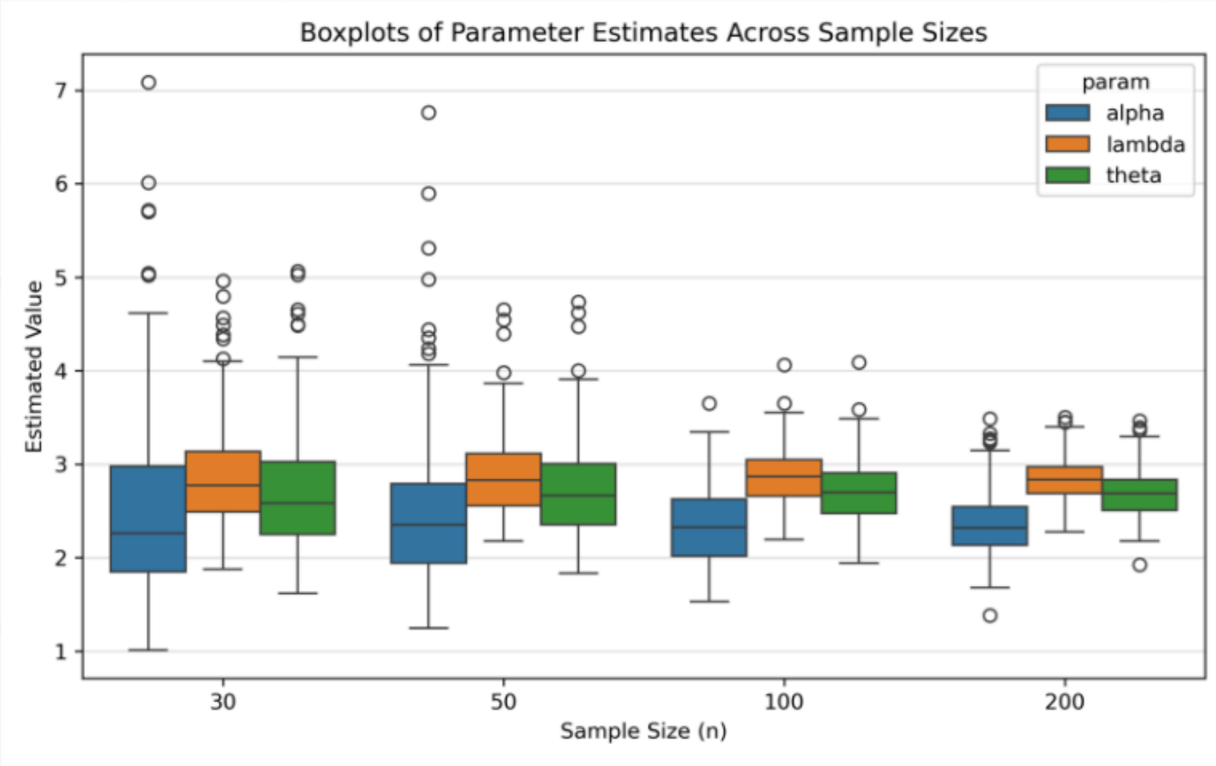
Simulation Study Results

- EM convergence plot shows an increase in log-likelihood across iterations
- Indicates successful movement toward optimal parameter estimates
- MSE decreases as sample size increases → accuracy improves with more data
- Confirms consistency and stability of EM for MOGE estimation



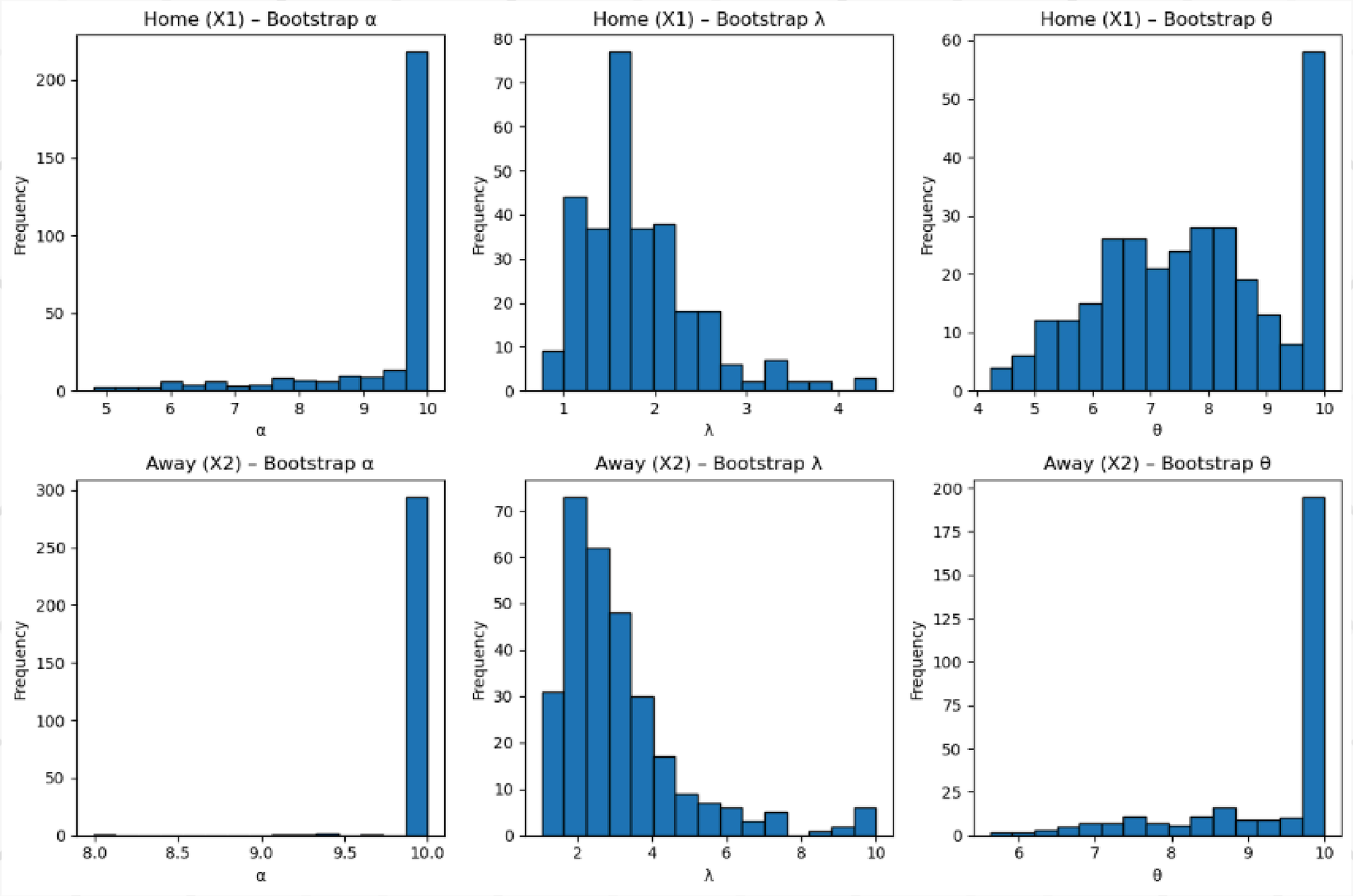
Bias & MSE

- High variance when n = 30 (small sample instability)
- Estimates become more concentrated for n ≥ 100
- MSE decreases steadily → estimator becomes more accurate
- Shows EM is consistent when the data is sufficient
- Small datasets cause noisy estimates the reason real data struggled



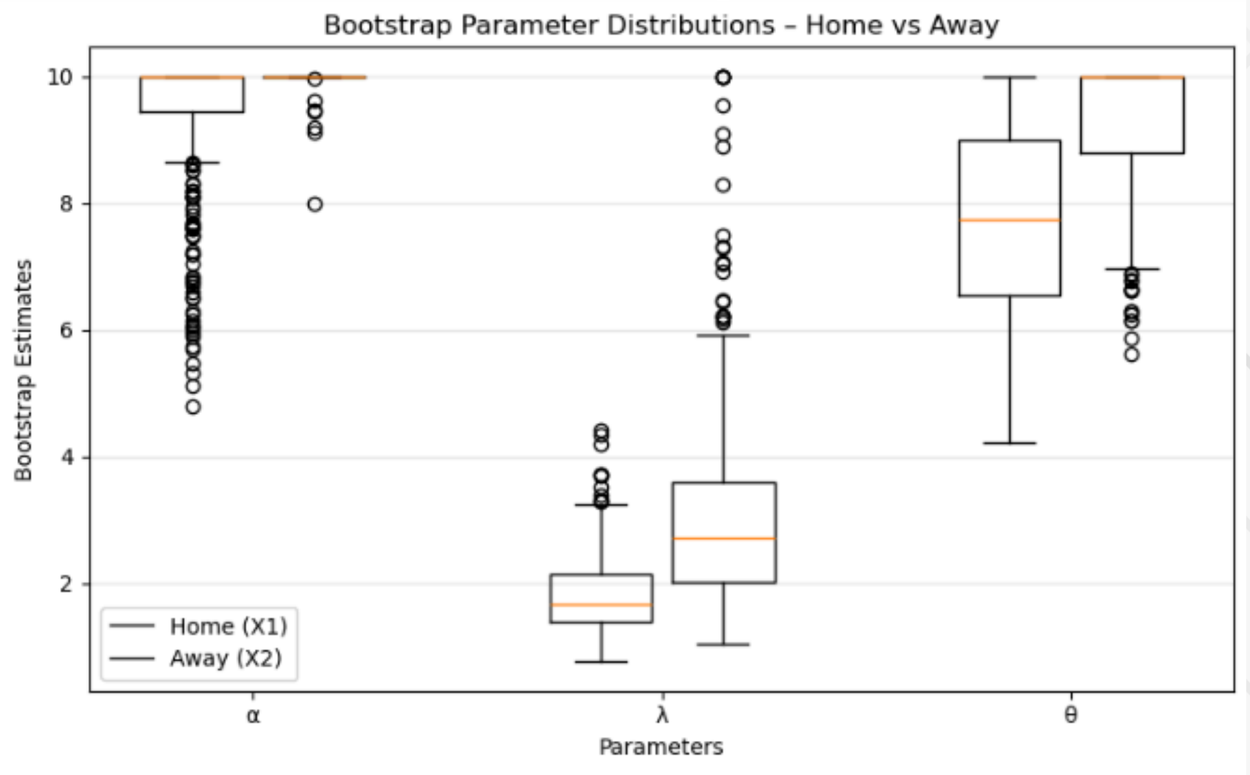
Bootstrap Variability

- Performed bootstrapping with B = 300 resamples
- Parameter-wise histograms plotted for Home (X1) & Away (X2)
- Reveals spread, skewness & uncertainty in estimates
- α more concentrated → shape parameter relatively stable
- λ & θ show higher spread → scoring rate varies more
- Away dataset displays heavier skew & wider tails
- Highlights need for uncertainty quantification with small samples


Home (X1) – Bootstrap α
Frequency
α
Home (X1) – Bootstrap λ
Frequency
λ
Home (X1) – Bootstrap θ
Frequency
θ
Away (X2) – Bootstrap α
Frequency
α
Away (X2) – Bootstrap λ
Frequency
λ
Away (X2) – Bootstrap θ
Frequency
θ

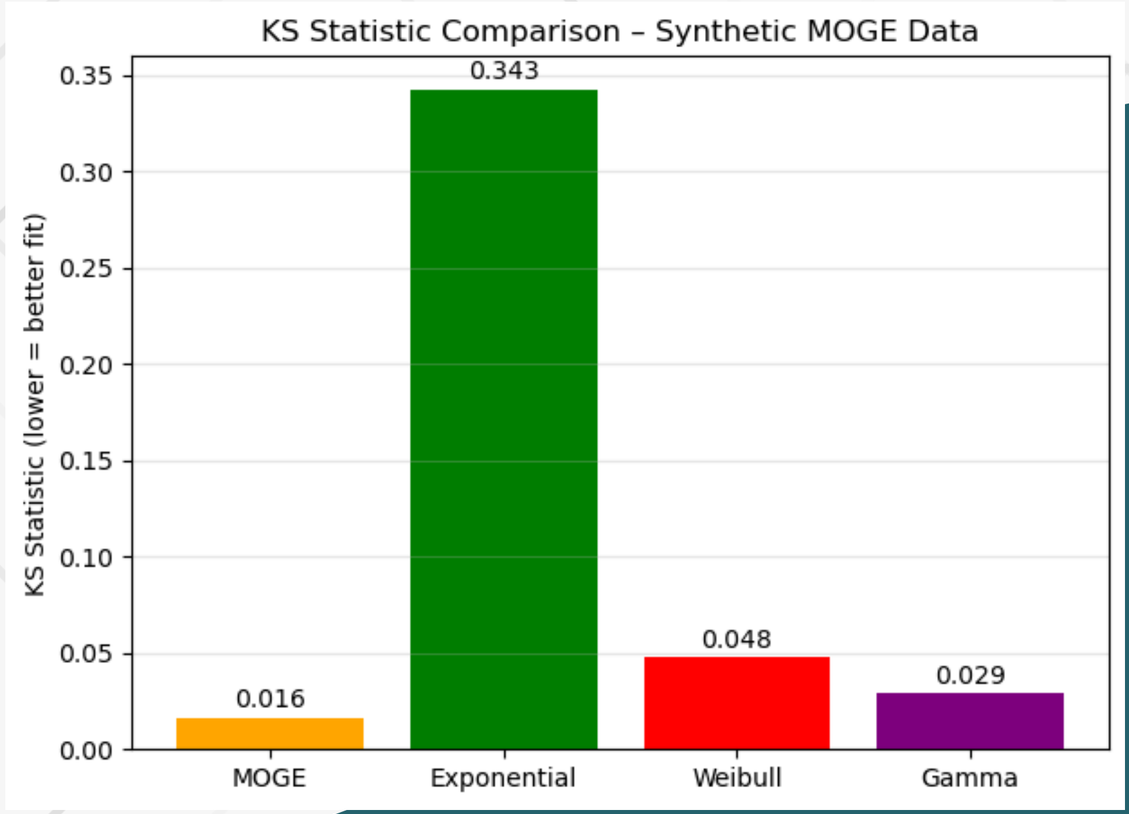
Bootstrap Parameter Uncertainty

- Boxplots visualize distribution of α, λ, θ across bootstraps
- Clear difference between Home (X1) and Away (X2) behaviour
- Useful for constructing confidence intervals
- Shows model uncertainty realistically under n=29
- Reinforces need for larger datasets or pooled seasons



Comparison – Synthetic vs Real Data

- Synthetic data generated using true MOGE parameters
- MOGE fits best → lowest KS statistic
- Weibull & Gamma moderately close, Exponential weakest
- Confirms MOGE works well as a generative & theoretical model
- Real-data mismatch due to small sample size + match randomness, not model failure



Conclusion

- MOGE performs poorly on real match data
- Fits synthetic data extremely well
- EM algorithm stable only with larger samples
- Real-data mismatch caused by small n and irregular scoring patterns

Limitations

- Real dataset too small (n = 29) → unstable estimation
- Bootstrapping showed high variance in λ and θ
- EM algorithm sensitive to initialization under small n
- Real scoring behaviour may not follow MOGE structure
- Alternative models (Weibull/Gamma) fit real data better

Future Work

- Use larger or multi-season football datasets
- Improve EM initialization for small-sample stability
- Explore Bayesian estimation for more robust inference
- Compare against hierarchical or covariate-based models
- Extend study to other sports or different event types

Thank You