### Introduction

- **Ultra-Marathons**
  - An race longer than a marathon (26.2 miles or 42.195 km)
  - Races: 50km, 100km, 250km, 6 Hour, 12 Hour.
- **International Association of Ultrarunners (IAU) World Championships**
  - 24 Hour, 2012; Katowice, Poland; Sept. 8-9; 260 runners, 34 countries

- **Running Strategies and Terminology**
  - Running at consistent pace (avg. min/mile) throughout race
  - Starting fast, relative to runner’s ability, and dropping out of the race
  - Starting at a consistent pace, slowing down, & picking it up at end

### Data

- The runners ran on a 1554 meter loop (slightly less than a mile)
- Cumulative lap counts after each hour were recorded
- Can view each runner as a trajectory of their lap counts for each hour
- Characteristics of these trajectories may indicate a running strategy

### Variables

- **Continuous**
  - Avg. Pace over 24 hours
  - Avg. Day Pace (12pm-6pm, 6am-12pm), Avg. Night Pace (6pm-6am) Hourly Running “Runs” Pace (in/4 lap per hour)
- **Ordinal**
  - Hour of Largest Lap Count Decrease
  - Hour of Largest Lap Count Increase
  - Both categorized as: 1-6am, 7am-12am, 1-6pm, 7pm-12am
- **Nominal**
  - Most Laps Dropped from Hour to Hour (1-2, 3-4, 5+)
  - Most Laps Gained from Hour to Hour (1-2, 3-4, 5+)

- **GOAL:** Determine the number and type of different running strategies

### Methodology

- **Gaussian Mixture Models for Continuous Data**
  - Assumes that the population density \( f(x) \) represents a weighted combination of \( G \) Gaussian densities, \( f_j(x) \), one per group [Fraley & Raftery]
  \[
  f(x) = \sum_{j=1}^{G} \pi_j f_j(x; \theta_j)
  \]
  - \( \theta_j \) mean, \( \pi_j \) variance matrix \( \Sigma_j \) (shape, volume, orientation of cluster)
  - Fit with Expectation-Maximization (EM) algorithm [McLachlan & Krishnan]

- **Latent Class Analysis for Categorical Variables**
  - For category \( k \), probability estimated by the difference of two cumulative distribution functions:
  \[
  P(Y_k = j) = \Phi\left(\frac{x_{kj} - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{x_{kj} - \mu_{j-1}}{\sigma_j}\right)
  \]
  - Calculate the cumulative sum of the table of category counts; use corresponding quantiles from normal distribution

- **Mixture Models with Mixed Data Type**
  - Continuous Variables: Multivariate Gaussian mixture model
  - Nominal Variables: Multivariate Categorical distribution

### Complexities with Estimation for Nominal Variables

- **Identifiability issues in the presence of nominal variables**
  - Solved by employing constraints on volume parameters in covariance [McLachlan & Krishnan]
  - These constraints still don’t mean we can fit all proposed models

- **Calculating probabilities in E-step intractable for nominal variables**
  - Solved by Monte Carlo approximation
  - Computes probability of response category using a large simulation from a normal distribution with mean \( p_k \) and covariance \( \Sigma_k \)
  - \( p_k \): Sample mean of all Monte Carlo samples that equal \( k \)
  - \( \Sigma_k \): Mean of inner products of simulated vectors from Monte Carlo samples that equal \( k \)

### Future Work/Issues

- **Future Work/Issues**
  - Approximate mixture model with mixed data for higher than \( G = 4 \)
  - Investigate problems in methodology of defining nominal variables
  - Fit mixture of Poisson processes, cluster runners by count trajectories
  - Further analysis on mixed membership models
  - (e.g., allowing runners to move between strategies)

---

**Carnegie Mellon University, Department of Statistics**

**References:**

- Dean, N. & Raftery, A. (2010)
- Heatmap: As the color gets lighter = larger BIC = better model

---

### Simulation: Distance Runner Strategies

- **Three Simulated Strategies**: 1000 runners each
  - **Strategy 1: Elite, Consistent, Low Variability**
  - **Strategy 2: Second, Somewhat consistent, High Variability**
  - **Strategy 3: Slow, Inconsistent, High Variability**

- 260 runners, approach chose a four strategy model
- Our approach extended the model estimation capability; two models still could not be fit due to small sample size
- Trajectory plots show different strategies across the clusters
  - Red = Elite. Seem to be most consistent and fastest runners
  - Green = Slowest. Most drop outs and slowest average pace
  - Black = Middle 1. More variability in middle of race
  - Blue = Middle 2. More variability at end of race

---

**Carnegie Mellon University, Department of Statistics**

**Future Work/Issues**

- Approximate mixture model with mixed data for higher than \( G = 4 \)
- Investigate problems in methodology of defining nominal variables
- Fit mixture of Poisson processes, cluster runners by count trajectories
- Further analysis on mixed membership models
- (e.g., allowing runners to move between strategies)