Lecture Summary: Illustrations with Data

Source: lec7.3.pdf

Key Points

- **Purpose:**
  - Explore real-world datasets to compute sample statistics and infer insights.
  - Assess the applicability of iid sampling models for such datasets.
- **Iris Dataset:**
  - Data:
    - Three classes (0, 1, 2) with 50 samples each.
    - Features: Sepal length, Sepal width, Petal length, Petal width.
  - Analysis:
    - Compute sample mean, variance, and proportions for features like sepal length.
    - Example:

$$S(\text{sepal length} > 5) = \frac{22}{50}, \quad S(4.8 \leq \text{sepal length} \leq 5.2) = \frac{20}{50}.$$

  - Observations:
    - Model the data as iid samples from an unknown distribution.
    - While the iid model seems reasonable for this dataset, it's a first-order approximation.
- **Taj Mahal Air Quality Dataset:**
  - Data:
    - 11 observations of pollution levels (SO2, NO2, PM2.5, PM10) in April 2021.
    - Maximum allowable limits: 80 (SO2, NO2), 60 (PM2.5), 100 (PM10).
  - Statistics:
    - Sample means and variances computed for all pollutants.
    - Proportions:

$$P(\text{exceeds max}) = \begin{cases} 0, & \text{SO2, NO2,} \\ \frac{7}{11}, & \text{PM2.5,} \\ \frac{11}{11}, & \text{PM10.} \end{cases}$$

  - Observations:
    - The iid sampling model may not be appropriate due to temporal correlations.
    - External factors (e.g., fires, seasonal effects) could influence the data.
  - Limitations:
    - Small dataset with 11 observations is insufficient for strong statistical conclusions.
- **IPL Dataset:**

  - Data:
    - Runs scored on first three deliveries of IPL matches (1598 innings).
  - Statistics:
    - Sample means and variances:

$$\bar{X}_{0.1} = 0.73, \quad \bar{X}_{0.2} = 0.87, \quad \bar{X}_{0.3} = 0.95.$$

    - Proportions:

$$P(\text{dot ball}) = \begin{cases} 0.5989, & 0.1, \\ 0.55, & 0.2, \\ 0.53, & 0.3. \end{cases}$$

$$P(\text{boundary}) = \begin{cases} 0.1, & 0.1, \\ 0.1145, & 0.2, \\ 0.13, & 0.3. \end{cases}$$

  - Observations:
    - Clear trends: Runs and boundaries increase from 0.1 to 0.3.
    - The iid model is reasonable but may require further checks for dependencies (e.g., psychological effects on bowlers after a boundary).

- **Lessons Learned:**
  - Statistical models and conclusions depend on the dataset size, quality, and context.
  - Large datasets (like IPL) allow for more reliable inferences compared to smaller datasets (like Taj Mahal).

Simplified Explanation

Key Datasets Analyzed: 1. **Iris Dataset:** Modeled as iid samples; computed basic statistics. 2. **Taj Mahal Air Quality:** Insufficient data for strong conclusions; temporal effects likely. 3. **IPL Data:** Large dataset revealing trends; iid model reasonable with caveats.

Insights: - Larger datasets provide stronger confidence in statistical stories. - Sample statistics offer valuable summaries but must be interpreted in context.

Conclusion

In this lecture, we:

- Explored sample statistics through three datasets.
- Discussed the suitability of iid models for different contexts.
- Highlighted the role of data size and context in statistical conclusions.

Real-world datasets illustrate the importance of understanding the assumptions and limitations behind statistical models.