Statistics for Machine Learning
Topic 0: Foundations & the Statistics Mindset

Instructor: David Tarkhanyan

Topic 0 Roadmap

- 0.1 From probability to data: what statistics is doing
- 0.2 Parameters, estimators, and loss (why mean/median appear)
- 0.3 What can go wrong: dependence, drift, and selection

Learning Objectives (Topic 0)

By the end of this topic, students should be able to:

- Explain the **population vs sample** distinction and the role of a **data-generating process**.
- Define **estimand**, **estimator**, **loss**, and **risk**.
- Relate **empirical risk minimization (ERM)** to standard ML training objectives.
- Identify common **assumption failures**: dependence, dataset shift, and selection bias.

Why Statistics (in ML)?

Probability:

- Starts with a model/distribution and deduces consequences.
- Example: If $X \sim \mathcal{N}(\mu, \sigma^2)$, compute $\mathbb{P}(X \leq x)$.

Statistics:

- Starts with data and infers unknown quantities (parameters, predictions, uncertainty).
- Example: Given samples $x_1, \ldots, x_n$, estimate μ and quantify uncertainty.

ML connection: training is typically estimating parameters by optimizing a criterion.

Population vs Sample

Population (or data-generating distribution):

$$X \sim P \quad \text{(unknown in practice)}$$

Sample:

$$X_1, \ldots, X_n \stackrel{\text{i.i.d.}}{\sim} P$$

- Population: "all possible observations" under the process of interest.
- Sample: the finite dataset we actually observe.
- **Goal:** learn something about P (or parameters of a model for P) from the sample.

The Data-Generating Process (DGP)

A useful mental model:

1. Nature chooses a distribution P.
2. We observe data $X_1, \ldots, X_n$ as draws from P (sometimes approximately).
3. We compute an estimator $\hat{\theta} = g(X_1, \ldots, X_n)$.
4. We report a value *and* (ideally) uncertainty.

Key point: the dataset is a **realization of random variables**. Estimators are random too.

The i.i.d. Assumption (and Why ML Loves It)

i.i.d. means:

- **Independent:** X_i does not provide information about X_j for $i \neq j$.
- **Identically distributed:** each X_i comes from the same P.

Why it matters:

- Enables Laws of Large Numbers (stability of averages).
- Enables CLT-based approximations (uncertainty / standard errors).
- Justifies train/test splitting under "same distribution" assumption.

Mini-Example: Conversion Rate

Let $X_i \in \{0, 1\}$ indicate whether user i converts.

- Estimand (parameter of interest): $p = \mathbb{P}(X = 1)$.
- Natural estimator: $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Interpretation:

- p is fixed but unknown (population truth).
- $\hat{p}$ is random (depends on which users appear in the sample).

Quick Check: Identify Estimand vs Estimator

For each scenario, identify (i) estimand and (ii) estimator.

1. Average session length from logs.
2. Fraud rate for transactions this month.
3. Mean latency of an API endpoint.
4. Click-through rate for a new UI variant.

Rule of thumb:

$$\text{estimand} = \text{population quantity}, \qquad \text{estimator} = \text{function of the sample}.$$

Estimand, Estimator, and Error

Estimand: θ (a property of P), e.g., mean μ or variance σ^2.
Estimator: $\hat{\theta} = g(X_1, \ldots, X_n)$.
Estimation error: $\hat{\theta} - \theta$ (random).

Why we need a criterion: which estimator is "better" depends on what we penalize.

Loss function: $L(\theta, x)$ measures how bad it is to choose θ when observing x.
Risk (expected loss):

$$R(\theta) = \mathbb{E}_{X \sim P}\big[L(\theta, X)\big].$$

Empirical risk:

$$\hat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(\theta, X_i).$$

ML connection: training often minimizes $\hat{R}_n(\theta)$ (ERM).

Why the Mean Appears: Squared Loss

Consider $L(\theta, x) = (x - \theta)^2$.
Empirical risk:

$$\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta)^2.$$

Minimizer:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (X_i - \theta)^2 = \bar{X}.$$

Interpretation: the sample mean is the best constant predictor under squared loss.

Why the Median Appears: Absolute Loss

Consider $L(\theta, x) = |x - \theta|$.
Empirical risk:

$$\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} |X_i - \theta|.$$

Minimizer:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} |X_i - \theta| = \text{any median of } \{X_i\}.$$

Interpretation: the median is robust to outliers compared to the mean.

Mode and MAP (Brief but Useful)

For discrete X, the **mode** of a distribution maximizes probability mass.
In Bayesian settings:

- Prior: $\pi(\theta)$
- Likelihood: $p(x \mid \theta)$
- Posterior: $p(\theta \mid x) \propto p(x \mid \theta)\, \pi(\theta)$

MAP estimate:

$$\hat{\theta}_{\mathsf{MAP}} = \arg\max_{\theta}\ p(\theta \mid x) = \arg\max_{\theta} \big(\log p(x \mid \theta) + \log \pi(\theta)\big).$$

ML connection: MAP often corresponds to **regularized** optimization.

In-Class Exercise: Mean vs Median Under Outliers

Dataset A: $[10, 11, 9, 10, 10, 11, 9]$
Dataset B: add one outlier: $[10, 11, 9, 10, 10, 11, 9, 100]$

- Compute mean and median for A and B.
- Which estimator changes more? Why?
- What loss function would you choose if outliers reflect measurement errors?

When i.i.d. Breaks: Three Common Failure Modes

1. **Dependence:** observations influence each other or are correlated (time series, grouped users).
2. **Dataset shift:** training and deployment/test distributions differ.
3. **Selection bias:** the sample is not representative of the target population.

Practical consequence: estimates and uncertainty can become systematically wrong.

Dependence (Examples and Implications)

Examples:

- Time dependence: X_t and X_{t+1} correlated (latency, demand).
- Clustered data: many rows per user or per device.
- Network effects: one user's treatment affects others (spillovers).

Implications:

- Effective sample size is smaller than n.
- Naive standard errors / p-values can be overly optimistic.

Mitigation ideas (preview):

- Blocked splits, cluster-robust methods, block bootstrap.

Dataset Shift (High-Level Taxonomy)

Let training distribution be $P_{\text{train}}(X, Y)$ and deployment be $P_{\text{test}}(X, Y)$.
Common cases:

- **Covariate shift:** $P(X)$ changes, $P(Y \mid X)$ stable.
- **Concept drift:** $P(Y \mid X)$ changes over time.
- **Label shift:** $P(Y)$ changes, $P(X \mid Y)$ stable (sometimes plausible).

ML impact: model evaluation and calibration can degrade unexpectedly.

Selection Bias (Sampling Is Part of the DGP)

Selection bias occurs when inclusion in the dataset depends on variables related to the outcome.
Examples:

- Only observing users who remain active (survivorship bias).
- Feedback loops in recommenders: what you show affects what you observe.
- Convenience samples: data from one region or device type only.

Key message: "more data" does not fix biased sampling.

Red-Flags Checklist (Operational)

Ask before trusting an estimate:

- Are observations independent (or clustered/time-correlated)?
- Is the sampling mechanism representative of the target population?
- Is the distribution stable over time (drift/seasonality)?
- Are train and test conditions aligned (same pipeline, same definition of labels)?
- Is there leakage (future information used in features)?

Classification Exercise: What's Wrong Here?

For each scenario, classify the dominant issue:

- Dependence
- Dataset shift
- Selection bias

1. Train on last year's data; deploy during a new marketing campaign.
2. Multiple rows per user; random row-wise train/test split.
3. Evaluate churn model only on users who opened the app last week.
4. Compare two models on the same test set after many rounds of tuning.

Topic 0 Summary

- Statistics starts from data and infers population quantities.
- Estimators are random; we need loss/risk to define "best".
- Mean/median arise as optimal constants under different losses.
- i.i.d. is powerful but fragile: dependence, shift, and selection bias matter.

Next: descriptive summaries, quantiles, ECDF, and sampling variability intuition.