Probability and Statistics

Module 5: Applied Probability and Statistical Tools

Complete classroom and examination notes

B.Sc. Computer Science • Course Code DASH250011

Based on the NIELIT Deemed to be University 2025 curriculum

Contents

1 Module map 1
2 1. Law of Large Numbers 1
2.1 Weak Law of Large Numbers 1
2.2 Strong Law of Large Numbers 1
2.3 Intuition and example 1
2.4 Chebyshev's inequality connection 1
2.5 Applications 2
3 2. Central Limit Theorem 2
3.1 CLT for sums 2
3.2 Standardizing a sample mean 2
3.3 CLT and Bernoulli proportions 3
3.4 Law of Large Numbers versus CLT 3
4 3. Simulation and Monte Carlo methods 3
4.1 Monte Carlo workflow 3
5 4. Markov chains 4
5.1 Transition matrix 4
5.2 Multi-step transition probabilities 4
5.3 State classification 4
6 5. Stationary distributions and long-run behavior 5
6.1 Worked stationary-distribution example 5
6.2 Absorbing chains 5
7 6. Markov-chain applications 5
8 7. Introduction to statistical machine learning 6
8.1 Learning settings 6
8.2 Features, targets, and datasets 6
8.3 Parameters and loss functions 6
9 8. Overfitting, underfitting, and regularization 6
10 9. Common statistical learning models 7
10.1 Linear regression 7
10.2 Logistic regression 7
10.3 Naive Bayes 7
10.4 Clustering 7
11 10. Model evaluation and uncertainty 7
12 11. Ethical and practical considerations 8
13 12. Problem-solving checklist 8
14 13. Quick formula sheet 8
15 14. Practice questions with answers 9

1 Module map

The syllabus module covers the Law of Large Numbers; Central Limit Theorem; Markov chains and their applications; and an introduction to statistical machine-learning concepts.

> **Learning outcomes.** You should be able to explain why averages stabilize, use the Central Limit Theorem to approximate sampling distributions, represent a finite-state Markov chain with a transition matrix, calculate multi-step and long-run probabilities, and connect probability and statistics to supervised and unsupervised machine-learning workflows.

2 1. Law of Large Numbers

The **Law of Large Numbers** explains why averages of repeated observations tend to approach the expected value as the number of observations increases.

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean μ, and define the sample average

$$|X|_n = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

2.1 Weak Law of Large Numbers

For every $\varepsilon > 0$,

$$\lim_{n\to\infty} P(||X|_n - \mu| > \varepsilon) = 0.$$

Equivalently, $|X|_n$ converges in probability to μ. For any desired tolerance, the probability that the average differs from μ by more than that tolerance becomes small as n grows.

2.2 Strong Law of Large Numbers

Under standard conditions,

$$P\left(\lim_{n\to\infty} |X|_n = \mu\right) = 1.$$

This is almost-sure convergence. It is a stronger statement than convergence in probability, although both express long-run stabilization.

2.3 Intuition and example

For fair coin tosses, let $X_i = 1$ for heads and 0 for tails. Then $E[X_i] = 0.5$. The average $|X|_n$ is the observed proportion of heads. It may fluctuate substantially for small n, but tends to approach 0.5 as the number of tosses increases.

> **What the law does not say.** It does not say that every short run must look balanced, that the average equals the expectation for a finite sample, or that dependent observations automatically satisfy the same result.

2.4 Chebyshev's inequality connection

If $X_1, \ldots, X_n$ are independent with common variance σ^2, then

$$V(|X|_n) = \frac{\sigma^2}{n}.$$

Chebyshev's inequality gives

$$P(||X|_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2},$$

which tends to 0 as n increases. This provides one route to the Weak Law.

2.5 Applications

- Estimating average latency from repeated requests.
- Estimating a conversion rate from many independent visitors.
- Estimating sensor bias through repeated measurements.
- Monte Carlo estimation of an integral or expected loss.
- Measuring the long-run failure rate of a service.

3 2. Central Limit Theorem

The **Central Limit Theorem** describes the distribution of a properly standardized sum or sample mean. It is one of the main reasons the Normal distribution appears throughout statistics.

If $X_1, ..., X_n$ are independent and identically distributed with mean μ and finite variance σ^2, then for large n,

$$Z = \frac{|X|_n - \mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0, 1).$$

Equivalently,

$$|X|_n \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

The approximation becomes more accurate as n increases, although the required sample size depends on the shape of the population distribution. Strong skewness, heavy tails, dependence, and outliers may require larger samples or different methods.

3.1 CLT for sums

For $S_n = X_1 + ... + X_n$:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \approx N(0, 1),$$

so $S_n \approx N(n\mu, n\sigma^2)$ for large n.

3.2 Standardizing a sample mean

To calculate $P(a \leq |X|_n \leq b)$ approximately:

1. Find the mean μ and standard error $\frac{\sigma}{\sqrt{n}}$.
2. Convert both endpoints to z values.
3. Use the standard Normal CDF Φ.

$$P(a \leq |X|_n \leq b) \approx \Phi\left(\frac{b - \mu}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{a - \mu}{\frac{\sigma}{\sqrt{n}}}\right).$$

Worked example: mean of measurements

Suppose individual measurements have $\mu = 50$ and $\sigma = 12$. For $n = 36$, $|X|$ has approximate mean 50 and standard error $\frac{12}{6} = 2$. Thus $P(46 \le |X| \le 54) \approx P(-2 \le Z \le 2) \approx 0.9545$.

3.3 CLT and Bernoulli proportions

If X_i are Bernoulli(p), then $|X| = \hat{p}$ and

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right).$$

This supports approximate confidence intervals and tests for large samples. The approximation is more reliable when both np and $n(1-p)$ are sufficiently large.

3.4 Law of Large Numbers versus CLT

Feature	Law of Large Numbers	Central Limit Theorem
Main question	Does the average approach μ?	How is the average distributed around μ?
Conclusion	$\|X\|_n$ becomes close to μ.	Standardized average is approximately Normal.
Typical use	Consistency and long-run behavior.	Approximate probabilities, intervals, and tests.

4 3. Simulation and Monte Carlo methods

A **Monte Carlo method** uses repeated random sampling to estimate a probability, expectation, integral, or risk quantity.

For a quantity $\theta = E[g(X)]$, simulate independent values $X_1, \ldots, X_n$ and estimate

$$\hat{\theta} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} g(X_i).$$

The Law of Large Numbers supports convergence of $\hat{\theta}$ to θ. The CLT can often quantify its sampling uncertainty:

$$\mathrm{SE}\left(\hat{\theta}\right) \approx \frac{s_g}{\sqrt{n}},$$

where s_g is the sample standard deviation of the simulated values $g(X_i)$.

4.1 Monte Carlo workflow

1. Define the target quantity precisely.
2. Generate random samples from the model.
3. Compute the quantity of interest for each sample.
4. Average the results.
5. Estimate uncertainty and repeat with a larger sample if needed.
6. Check random-number seeds, convergence, and sensitivity to model assumptions.

Applications include estimating rare-event probabilities, queueing performance, portfolio risk, reliability, Bayesian posterior quantities, and multidimensional integrals.

5 4. Markov chains

A **Markov chain** is a stochastic process in which the distribution of the next state depends on the current state, not on the full history.

For states $X_0, X_1, \ldots$,

$$P(X_{t+1} = j \mid X_t = i, X_{t-1}, \ldots, X_0) = P(X_{t+1} = j \mid X_t = i).$$

This is the **Markov property**.

A finite-state Markov chain has states such as "healthy," "degraded," and "failed," or pages in a web-navigation model.

5.1 Transition matrix

For m states, define the transition matrix P by

$$P_{ij} = P(X_{t+1} = j \mid X_t = i).$$

Using the row-vector convention, each row sums to 1:

$$P_{ij} \geq 0, \quad \sum_j P_{ij} = 1.$$

If the current distribution is the row vector π_t, then

$$\pi_{t+1} = \pi_t P, \quad \pi_t = \pi_0 P^t.$$

A matrix is **stochastic** when its entries are nonnegative and every row sums to 1 under this convention.

> **Worked example: two-state reliability**
>
> Let state G mean "good" and F mean "fault." Suppose $P(G \to G) = 0.9$, $P(G \to F) = 0.1$, $P(F \to G) = 0.4$, and $P(F \to F) = 0.6$. With state order (G, F), $P = ((0.9, 0.1), (0.4, 0.6))$. If the system starts good, $\pi_0 = (1, 0)$ and $\pi_1 = (0.9, 0.1)$.

5.2 Multi-step transition probabilities

The entry $(P^n)_{ij}$ is the probability of being in state j after n steps when starting in state i.

For two steps,

$$(P^2)_{ij} = \sum_k P_{ik} P_{kj}.$$

This sums the probabilities over every possible intermediate state.

5.3 State classification

- **Accessible**: state j is accessible from i if some $n \geq 0$ has $(P^n)_{ij} > 0$.
- **Communicating**: i and j communicate if each is accessible from the other.
- **Closed class**: once entered, the chain cannot leave the class.
- **Absorbing state**: $P_{ii} = 1$; once entered, it is never left.
- **Irreducible**: every state communicates with every other state.
- **Periodic**: returns to a state only at multiples of a common period greater than 1.
- **Aperiodic**: the period is 1.

These properties affect long-run behavior and whether a unique stationary distribution exists.

6 5. Stationary distributions and long-run behavior

A probability vector π is **stationary** if

$$\pi P = \pi, \quad \sum_i \pi_i = 1, \quad \pi_i \geq 0.$$

To find it, solve the linear equations $\pi P = \pi$ together with the normalization equation.
For a finite irreducible, aperiodic Markov chain, the distribution often converges to a unique stationary distribution regardless of the initial distribution:

$$\lim_{n\to\infty} \pi_0 P^n = \pi.$$

6.1 Worked stationary-distribution example

For the two-state matrix

$$P = ((0.9, 0.1), (0.4, 0.6)),$$

let $\pi = (a, 1-a)$. The first stationary equation is

$$a = 0.9a + 0.4(1-a).$$

Thus $0.5a = 0.4$, so $a = 0.8$ and $\pi = (0.8, 0.2)$. In the long run, the system is good about 80% of the time and faulty about 20% of the time under this model.

6.2 Absorbing chains

An absorbing chain has at least one absorbing state. It can model completion, failure, exit, or adoption. Useful questions include:

- What is the probability of eventual absorption in each absorbing state?
- What is the expected number of steps before absorption?
- Which transient states are visited frequently?

Reorder states into transient and absorbing groups. The transient block Q is used to form the fundamental matrix

$$N = (I - Q)^{-1}.$$

The entry N_{ij} gives the expected number of visits to transient state j starting from transient state i, including the initial visit. This topic is often developed further in advanced stochastic-process courses.

7 6. Markov-chain applications

- **Web navigation**: states are pages and transitions are click probabilities; stationary probabilities motivate ranking models.
- **Weather and environment**: today's state predicts tomorrow's state probabilistically.
- **Computer systems**: healthy, overloaded, degraded, and failed states support availability analysis.
- **Queueing**: state is the number of jobs or customers in a system.
- **Text generation**: states can represent words or contexts; transitions define an n-gram model.
- **Customer behaviour**: states may represent lifecycle stages, with transitions for retention and churn.
- **Bioinformatics**: hidden Markov models represent hidden biological states generating observed sequences.

A Markov model is useful only when its state definition captures enough information for the Markov property to be a reasonable approximation.

8 7. Introduction to statistical machine learning

Statistical machine learning builds models that learn patterns from data and quantify uncertainty or prediction error. Probability supplies the language for uncertainty; statistics supplies estimation and inference; optimization fits model parameters.

8.1 Learning settings

Setting	Goal	Examples
Supervised learning	Learn a mapping from features X to known target Y.	Regression, classification.
Unsupervised learning	Find structure without labelled targets.	Clustering, dimensionality reduction.
Semi-supervised	Use a small labelled set with many unlabelled examples.	Document or image classification.
Reinforcement learning	Learn actions from rewards and state transitions.	Control, games, resource allocation.

8.2 Features, targets, and datasets

A **feature** is an input variable; a **target** or label is the outcome to predict. A dataset is commonly split into training, validation, and test sets.

- **Training set**: fit model parameters.
- **Validation set**: choose hyperparameters or compare model choices.
- **Test set**: final unbiased performance estimate, used sparingly.

Never allow test-set information to influence model selection. Otherwise, reported test performance may be optimistically biased.

8.3 Parameters and loss functions

A model $f_{\theta(x)}$ has parameters θ. Training chooses parameters to minimize a loss over the training data:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f_{\theta(x_i)}\big).$$

Common losses:

- Squared error: $L(y, \hat{y}) = (y - \hat{y})^2$ for regression.
- Absolute error: $|y - \hat{y}|$, more resistant to large errors.
- Log loss / cross-entropy for probabilistic classification.
- Hinge loss in some margin-based classifiers.

The empirical risk is average training loss. Generalization performance is expected loss on new data from the target population.

9 8. Overfitting, underfitting, and regularization

- **Underfitting**: model is too simple to capture systematic structure; training and test errors are both high.
- **Overfitting**: model learns noise or peculiarities of training data; training error is low but test error is high.
- **Generalization**: ability to perform well on unseen data.

The bias–variance perspective separates prediction error into approximation bias, estimation variance, and irreducible noise. Model complexity often lowers training error but can increase variance.
Regularization adds a penalty to discourage overly complex parameters:

$$\hat{\theta} = \arg\min_{\theta}[\text{empirical loss} + \lambda\ \text{penalty}(\theta)].$$

Examples include ridge-style squared-parameter penalties and lasso-style absolute-parameter penalties. The parameter λ controls the strength of regularization and should be selected without using the final test set.

> **Data leakage warning.** Scaling, feature selection, imputation, or target-derived variables must be fitted using training data only within each validation split. Leakage can produce impressive but invalid test results.

10 9. Common statistical learning models

10.1 Linear regression

For a numerical target:

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_p x_p.$$

Coefficients describe associations conditional on the included features, not automatically causal effects. Evaluate with mean squared error, root mean squared error, mean absolute error, and residual diagnostics.

10.2 Logistic regression

For a binary target, logistic regression models a probability:

$$P(Y = 1 \texttt{ mid } x) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + \ldots + b_p x_p)}}.$$

The log-odds are linear in the features. Classification requires a threshold or decision rule, which should reflect the application's costs of false positives and false negatives.

10.3 Naive Bayes

Using Bayes' theorem and a conditional-independence assumption:

$$P(C \texttt{ mid } x_1, \ldots, x_p) \propto P(C) P(x_1 \texttt{ mid } C) \ldots P(x_p \texttt{ mid } C).$$

The assumption is often unrealistic but can work well, especially with text features and small datasets.

10.4 Clustering

Clustering groups observations by similarity without target labels. In k-means, choose k centroids to minimize within-cluster squared distance. Results depend on scaling, initialization, distance choice, and the selected k.

11 10. Model evaluation and uncertainty

A model evaluation must match the task and decision:

- Regression: MAE, RMSE, R^2, calibration of prediction intervals.
- Classification: accuracy, precision, recall, specificity, F1 score, ROC-AUC, and calibration.

- Imbalanced classification: accuracy can be misleading; inspect class-specific measures and the confusion matrix.
- Probabilistic predictions: evaluate calibration, log loss, or Brier score.
- Time-dependent data: use temporal splits rather than random shuffling when future prediction is the goal.

Use cross-validation to estimate performance and compare models while preserving the structure of the data. Report uncertainty through repeated folds, bootstrap intervals, or other suitable methods rather than presenting one score as exact.

12 11. Ethical and practical considerations

Statistical learning systems can amplify sampling bias, encode historical discrimination, expose personal data, and fail under distribution shift. Good practice includes:

- define the intended population and use;
- measure subgroup performance where relevant;
- document data provenance, missingness, and labelling;
- protect privacy and minimize unnecessary data;
- monitor drift after deployment;
- provide human oversight for high-impact decisions;
- distinguish correlation from causation;
- communicate uncertainty and limitations.

13 12. Problem-solving checklist

1. For averages, ask whether the Law of Large Numbers or CLT applies.
2. Check independence, identical distribution, finite variance, and sample size.
3. Standardize sums and means carefully; use the correct standard error.
4. For a Markov chain, define states, transition direction, and matrix convention.
5. Verify each transition row sums to 1.
6. Use matrix powers for multi-step probabilities and solve $\pi P = \pi$ for stationarity.
7. For machine learning, define target, features, loss, split strategy, and evaluation metric.
8. Prevent leakage and report performance on data not used for fitting.
9. Interpret results in the context of assumptions, uncertainty, and consequences.

14 13. Quick formula sheet

Concept	Formula
Sample average	$\lvert X\rvert_n = \left(\frac{1}{n}\right)\sum_i X_i$
Variance of average	$V(\lvert X\rvert_n) = \frac{\sigma^2}{n}$
CLT statistic	$Z = \frac{\lvert X\rvert_n - \mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0,1)$
Monte Carlo estimate	$\hat{\theta} = \left(\frac{1}{n}\right)\sum_i g(X_i)$
Markov update	$\pi_{t+1} = \pi_t P$
Multi-step distribution	$\pi_t = \pi_0 P^t$

Stationary distribution	$\pi P = \pi, \sum_i \pi_i = 1$
Fundamental matrix	$N = (I - Q)^{-1}$
Regularized risk	$\text{loss} + \lambda \text{ penalty}$

15 14. Practice questions with answers

1. State the main difference between the Law of Large Numbers and the Central Limit Theorem. **Answer:** LLN says the sample average approaches the population mean; CLT describes the approximate Normal distribution of the standardized average.
2. If $\sigma = 20$ and $n = 100$, find the standard error of the sample mean. **Answer:** $\frac{20}{\sqrt{100}} = 2$.
3. For a two-state chain, $P = ((0.8, 0.2), (0.3, 0.7))$. Verify that it is a valid transition matrix. **Answer:** All entries are nonnegative and each row sums to 1.
4. What equation defines a stationary distribution? **Answer:** $\pi P = \pi$, together with nonnegative entries summing to 1.
5. Why does a larger simulation sample usually reduce Monte Carlo error? **Answer:** The Law of Large Numbers stabilizes the average and the standard error typically decreases at rate $\frac{1}{\sqrt{n}}$.
6. Give one example of data leakage. **Answer:** Scaling the complete dataset before cross-validation, so validation folds influence the scaling parameters.
7. Why can accuracy be misleading for an imbalanced classification problem? **Answer:** A classifier can achieve high accuracy by predicting the majority class while failing to detect the rare but important class.
8. What is the role of a test set? **Answer:** To provide a final performance estimate on data not used for fitting or model selection.

> **Exam reminder.** For a Markov-chain question, clearly state whether you use row vectors or column vectors. For a CLT question, write the mean and standard error before standardizing. For a machine-learning question, identify the data split and evaluation metric before discussing performance.