Statistics for Machine Learning

"Statistics is the grammar of science." — Karl Pearson

What is Statistics?

Statistics is the discipline that *collects, organises, summarises, analyses, and draws conclusions* from data. In machine–learning (ML) it underpins model building, uncertainty quantification, and validation.

- **Population**: the full set we care about.
- **Sample**: a subset used to infer about the population.
- **Parameter** (Greek, e.g. μ, σ): a population measure (unknown).
- **Statistic** (Latin, e.g. $\bar{x}, s$): computed from a sample.

Types of Data

Category	**Subtype**	**Description**	**Examples**	**ML Encoding**
Qualitative	Nominal	Unordered labels	Browser {Chrome, Safari}	One–hot
	Ordinal	Ordered labels	Likert {Poor→Excellent}	Ordinal / Target
Quantitative	Discrete	Integer counts	Clicks per session	(Scaled) integer
	Continuous	Real-valued	Temperature °C	Normalisation

Why it matters

Choosing the wrong encoding can break distance-based models (e.g. k-NN treats encoded categories as numeric).

Common Encoding Tricks

- **Label Encoding** – ordinal only.
- **One-Hot / Dummy** – nominal default.
- **Frequency Encoding** – map category $\to$ empirical probability.
- **Target Encoding** – replace with mean target (beware leakage).

Two Main Branches of Statistics

Branch	**Goal**	**Typical ML Question**
Descriptive	Condense & visualise existing data	"What is the average click-through-rate?"
Inferential	Generalise and quantify uncertainty	"Will the new UI raise CTR across *all* users?"

Descriptive Statistics

1. Measures of Central Tendency (MCT)

Symbol	**Name**	**Formula**	**Derivation Idea**
$\bar{x}$	Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Minimise $\sum (x_i - c)^2$ w.r.t. c
$\tilde{x}$	Median	Middle value (or average of two middles)	Minimise $\sum \lvert x_i - c \rvert$
–	Mode	Most frequent value	Useful for categorical data

2. Measures of Dispersion (MD)

Symbol	Name	Formula	Interpretation
s^2	Sample Variance	$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$	Avg. squared deviation (Bessel correction)
s	Std. Dev.	$s = \sqrt{s^2}$	Back to original units
IQR	Inter-Quartile Range	$Q_3 - Q_1$	Robust spread (box□plot)

ML tie-in

- Z-score scaling uses mean & std-dev.
- Robust scaling uses median & IQR.

3. Distribution Shape

Skewness $\gamma_1 = \frac{\mu_3}{\sigma^3}$, Kurtosis $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$.

4. Visual Tools

Histogram, box-plot, pair-plot, correlation heat-map.

Inferential Statistics

1. Estimation

Type	Output	Formula / Method	ML Context
Point	$\hat{\theta}$	MLE: maximise $L(\theta) = \prod f(x_i \vert \theta)$	Fit model weights by MLE
CI	$[\hat{\theta} \pm z_{\alpha/2}\, SE]$	$SE = \frac{s}{\sqrt{n}}$ (for mean)	Report $\pm 1.96\, SE$ around accuracy

2. Hypothesis Testing

- State H_0 vs H_1.
- Compute test-statistic (t, z, χ^2, F).
- p-value = Prob. of statistic $\geq$ observed if H_0 true.
- Reject if $p < \alpha$.

Common tests: t-test, z-test, χ^2, ANOVA.

3. Resampling & CLT

Bootstrap for CIs; k-fold cross-validation for generalisation error.

4. Bias–Variance Trade-off

$$\mathbb{E}\big[(y - \hat{f}(x))^2\big] = \underbrace{\big(\text{Bias}[\hat{f}(x)]\big)^2}_{\text{under-fit}} + \underbrace{\text{Var}[\hat{f}(x)]}_{\text{over-fit}} + \sigma^2_{\text{irreducible}}$$

High bias $\Rightarrow$ under-fit; High variance $\Rightarrow$ over-fit.

Real-World ML Use-Cases

- **EDA**: detect skewness ⇒ log-transform.
- **A/B Testing**: hypothesis test on conversion rates.
- **Early stopping**: monitor CV error to balance variance.
- **Ensembles**: bagging (Random Forest) lowers variance.

Sampling Techniques

Why Sampling?

- **Cost & Time**: measuring an entire population (N units) is often impractical.
- **Feasibility**: some units may be inaccessible or destroyed by measurement (e.g. crash-testing).
- **Precision**: a well–designed sample can achieve *lower variance* than a poorly executed census.

Key Terminology

Population (N)	Entire set of interest (all transactions, voters, etc.)
Sample (n)	Subset drawn from the population, $n \ll N$.
Sampling Frame	List or mechanism that identifies every population unit.
Parameter	Fixed but unknown value (e.g. μ, σ).
Estimator	Rule that maps the sample to an estimate $\hat{\theta}$.
Sampling Error	$\theta - \hat{\theta}$ (random, has variance).

Notation & Basic Formulas

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad SE(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} \times \sqrt{\frac{N-n}{N-1}} \text{ (fpc)}$$

where the last factor is the *finite-population correction*.

Probability Sampling (each unit has known $p > 0$)

1. **Simple Random Sampling (SRS)** Choose n units with equal probability $1/\binom{N}{n}$.

$$SE_{\text{SRS}}(\bar{X}) = \sqrt{\frac{S^2}{n}}$$

2. **Systematic Sampling** Pick every k-th unit after a random start; $k = N/n$.
3. **Stratified Sampling** Partition into H strata, sample n_h per stratum.

$$\bar{X}_{\text{str}} = \sum_{h=1}^{H} W_h \, \bar{X}_h, \quad SE^2 = \sum_{h=1}^{H} W_h^2 \, \frac{S_h^2}{n_h}$$

   where $W_h = N_h/N$.
4. **Cluster Sampling** Sample g entire clusters (e.g. schools) out of G. Use cluster means to estimate population mean; design effect DEFF $= 1 + \rho(m-1)$ where ρ is intra-cluster correlation.
5. **Multi-stage Sampling** Combine stages (e.g. clusters → households → individuals).

Non-Probability Sampling (selection p unknown)

Convenience
Grab units that are easy to reach (e.g. street survey).

Judgment / Purposive
Expert chooses 'typical' cases.

Quota
Ensure sample proportions match some traits (age, gender).

Snowball
Existing subjects recruit future ones (common in hidden populations).

These methods can introduce selection bias; inferential statistics (e.g. confidence intervals) are generally invalid without strong assumptions.

Design Effect & Effective Sample Size

$$\text{DEFF} = \frac{\text{Var}_{\text{actual}}(\hat{\theta})}{\text{Var}_{\text{srs}}(\hat{\theta})}, \qquad n_{\text{eff}} = \frac{n}{\text{DEFF}}$$

A complex design with $\text{DEFF} = 2$ cuts precision in half relative to an SRS.

ML Use-Cases of Sampling

- **Mini-batch SGD**: treats each mini-batch as an SRS from data.
- **Stratified train/test split**: maintain class balance to stabilise metrics.
- **Negative sampling**: randomly sample negative pairs in word2vec.

Probability Basics

Definition

For an experiment with sample space Ω, a *probability measure* $P\colon \mathcal{F} \to [0,1]$ assigns numbers to events (σ-algebra $\mathcal{F}$) such that $P(\Omega) = 1$ and countable additivity holds.

Types of Events

- **Mutually Exclusive**: $A \cap B = \varnothing$.
- **Exhaustive**: $A_1 \cup \cdots \cup A_k = \Omega$.
- **Independent**: $P(A \cap B) = P(A)P(B)$.
- **Complementary**: $A^c = \Omega \setminus A$, $P(A^c) = 1 - P(A)$.

Addition & Complement Rules

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \qquad P(A^c) = 1 - P(A).$$

Cumulative Probability

For a real r.v. X the *cumulative distribution function (cdf)* is

$$F_X(x) = P(X \le x) \quad (\text{non-decreasing},\ 0 \le F_X \le 1).$$

Conditional Probability and Bayes' Theorem

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad P(B \mid A) = \frac{P(A \mid B)P(B)}{\sum_i P(A \mid B_i)P(B_i)}.$$

Probability Distributions

Discrete Distributions

Dist.	**When to Use**	**pmf** $p(k)$	$E[X]$	$\mathrm{Var}(X)$
Bernoulli(p)	Single 0/1 trial	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Binomial(n,p)	# successes in n iid trials	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Geometric(p)	Trials until 1st success	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	Rare events per interval	$\dfrac{e^{-\lambda}\lambda^k}{k!}$	λ	λ

Continuous Distributions

Dist.	**When to Use**	**pdf** $f(x)$	$E[X]$	$\mathrm{Var}(X)$
Uniform(a,b)	Equal likelihood in $[a,b]$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal(μ,σ^2)	Sum of many effects	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Exponential(λ)	Waiting time between Poisson events	$\lambda e^{-\lambda x}\ (x \geq 0)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma(k,θ)	Sum of k Expo. vars	$\dfrac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k}$	$k\theta$	$k\theta^2$

The Bell Curve and the Central Limit Theorem

$$\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \qquad (n\to\infty),$$

i.e. the standardized sample mean of iid variables converges in distribution to the Normal—explaining why the "bell curve" appears so often in practice.

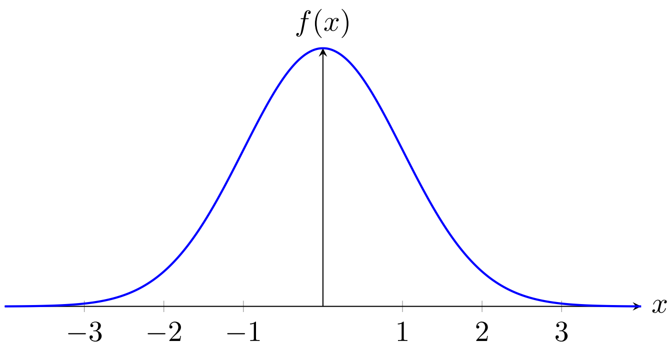

Figure 1: Standard Normal ("bell curve") density

Skewness: Definition, Concept, and Effects

Population Definition. The *skewness* of a random variable X with mean μ and standard deviation σ is the **third standardised moment**

$$\gamma_1 \;=\; \frac{E[(X-\mu)^3]}{\sigma^3} \quad \Big(\text{dimensionless}\Big).$$

Sample Estimator. For observations $x_1,\dots,x_n$ with sample mean $\bar{x}$ and sample standard deviation s:

$$g_1 \;=\; \frac{\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2\right)^{3/2}} \qquad (\text{bias□corrected: } G_1 = \tfrac{n}{(n-1)(n-2)}g_1).$$

Conceptual Meaning.

– $\gamma_1 > 0$: **right□skew**. Mass piles to the left, long tail to the right; typically mean $>$ median $>$ mode.
– $\gamma_1 < 0$: **left□skew**. Long tail to the left; relation reverses.
– $|\gamma_1| \lesssim 0.5$: practically symmetric for many analyses.

Why it matters in practice.

– Many parametric tests (e.g. t-test) assume near□normality; heavy skew may inflate Type-I error.
– In regression, skewed residuals violate homoscedasticity and influence coefficients.
– Transformations (log, Box–Cox) often aim to reduce skewness before modelling.

Quick sanity check (rule of thumb). If mean$-$median $>$ 0 the distribution is likely right-skewed, and vice-versa; the magnitude divided by the standard deviation approximates γ_1 for moderate skew.

Skewness and Kurtosis

1. Skewness

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{E[(X-\mu)^3]}{\left(E[(X-\mu)^2]\right)^{3/2}}$$

– $\gamma_1 > 0$: *right□skewed* (long tail to the right).
– $\gamma_1 < 0$: *left□skewed*.
– $\gamma_1 = 0$: symmetric (e.g. Normal).

2. Kurtosis

$$\gamma_2 = \frac{\mu_4}{\sigma^4} \implies \text{excess kurtosis} = \gamma_2 - 3$$

Type	Shape	Excess Kurtosis
Leptokurtic	Sharp peak, heavy tails	> 0
Mesokurtic	Normal reference	$= 0$
Platykurtic	Flat peak, light tails	< 0

Interpretation. Higher kurtosis signals more probability in the tails *and* the centre, implying outlier□prone data; lower kurtosis indicates a flatter, “short□tailed” distribution.

3. Illustrative Shapes

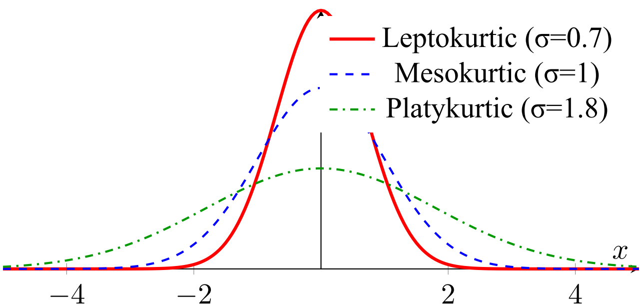

Figure 2: Relative shapes of leptokurtic, mesokurtic, and platykurtic curves (all areas = 1).

© 2025 Divyanshu Kumar