Stats

1 Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with

$$\mathbb{E}[X_i] = \mu, \quad \mathrm{Var}(X_i) = \sigma^2 < \infty.$$

Define the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then, as $n \to \infty$,

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

Equivalently,

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{for large } n.$$

The Delta Method

The Delta Method uses a first-order Taylor expansion to transfer asymptotic normality through a smooth function. A way to approximate the distribution of a function of a statistic (e.g., $g(\bar{X})$) using the CLT and a Taylor expansion.
Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with

$$\mathbb{E}[X_i] = \mu, \quad \mathrm{Var}(X_i) = \sigma^2 < \infty.$$

Define the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Let $g(\cdot)$ be differentiable at μ with $g'(\mu) \neq 0$. Then

$$\sqrt{n}\big(g(\bar{X}_n) - g(\mu)\big) \xrightarrow{d} \mathcal{N}\left(0, [g'(\mu)]^2 \sigma^2\right).$$

Equivalently,

$$g(\bar{X}_n) \sim \mathcal{N}\left(g(\mu), \frac{[g'(\mu)]^2 \sigma^2}{n}\right) \quad \text{for large } n.$$

Berry–Esseen Theorem

Berry–Esseen strengthens the CLT by providing a quantitative bound on the accuracy of the normal approximation, with error shrinking at rate $1/\sqrt{n}$. It quantifies the rate of convergence in the CLT, providing a bound on the error of the normal approximation.
Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables such that

$$\mathbb{E}[X_i] = \mu, \quad \mathrm{Var}(X_i) = \sigma^2 > 0, \quad \mathbb{E}[|X_i - \mu|^3] = \rho < \infty.$$

Define

$$S_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu).$$

Let $\Phi(x)$ denote the CDF of the standard normal distribution. Then there exists a universal constant C such that

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(S_n \leq x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}.$$

Context: For symmetric distributions, we can reliably use CLT with $n = 30$; for highly skewed ones (e.g., exponential, income data), we may need $n = 100+$.

Lindeberg–Feller Central Limit Theorem

Lindeberg–Feller CLT extends the classical CLT to independent, non-identically distributed variables by ensuring no single term dominates the variance.
Let $\{X_{n,i}\}_{i=1}^{k_n}$ be a triangular array of independent random variables with

$$\mathbb{E}[X_{n,i}] = 0, \quad \text{Var}(X_{n,i}) = \sigma_{n,i}^2.$$

Define

$$s_n^2 = \sum_{i=1}^{k_n} \sigma_{n,i}^2, \quad \text{and assume } s_n^2 \to 1.$$

Lindeberg Condition

For every $\varepsilon > 0$,

$$\lim_{n\to\infty} \sum_{i=1}^{k_n} \mathbb{E}\left[X_{n,i}^2 \mathbf{1}_{\{|X_{n,i}|>\varepsilon\}}\right] = 0.$$

Conclusion

Then

$$\sum_{i=1}^{k_n} X_{n,i} \xrightarrow{d} \mathcal{N}(0,1).$$

Multivariate Central Limit Theorem

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be i.i.d. random vectors in $\mathbb{R}^d$ such that

$$\mathbb{E}[\mathbf{X}_i] = \boldsymbol{\mu}, \quad \text{Cov}(\mathbf{X}_i) = \boldsymbol{\Sigma}.$$

Define the sample mean vector

$$\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i.$$

Then,

$$\sqrt{n}\big(\bar{\mathbf{X}}_n - \boldsymbol{\mu}\big) \xrightarrow{d} \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma}).$$

What does Multivariate Normal Means?
A d-dimensional random vector $\mathbf{Z}$ is said to follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, denoted by $\mathbf{Z} \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma})$, if and only if for every $\mathbf{a} \in \mathbb{R}^d$,

$$\mathbf{a}^\top \mathbf{Z} \sim \mathcal{N}\big(0, \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}\big)\,.$$

2 Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with

$$\mathbb{E}[X_i] = \mu < \infty.$$

Define the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Weak Law of Large Numbers

$$\bar{X}_n \xrightarrow{P} \mu \quad \text{as } n \to \infty.$$

(convergence in probability)

Strong Law of Large Numbers

If $\mathbb{E}[|X_i|] < \infty$, then

$$\bar{X}_n \xrightarrow{a.s.} \mu \quad \text{as } n \to \infty.$$

(convergence almost surely)

Failure of LLN with Infinite Mean

The LLN relies on the existence of a finite mean. Let $X_1, X_2, \ldots$ be i.i.d. Cauchy$(0, 1)$ random variables. The expected value $\mathbb{E}[X_i]$ does not exist.
Define

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then

$$\bar{X}_n \overset{d}{=} X_1 \quad \text{for all } n,$$

and hence

$$\bar{X}_n \nrightarrow \mu$$

in probability or almost surely.

3 Confidence Interval

A confidence interval is a data-dependent range that captures the true parameter with a specified long-run frequency.
A confidence interval for a parameter θ is an interval $[L(X), U(X)]$ such that

$$\mathbb{P}\big(L(X) \leq \theta \leq U(X)\big) = 1 - \alpha.$$

Assumptions & Conditions

- Random sampling (or at least representative).
- Independence of observations ($n < 10\%$ of population for sampling without replacement).
- Normality condition: (1) Large n ($n \geq 30$) by CLT, OR (2) Original population is normal (for small n).

CI for Mean (Known Variance)

If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

CI for Mean (Unknown Variance)

$$\bar{X}_n \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Asymptotic (CLT-Based) Confidence Interval

For any distribution with finite variance:

$$\bar{X}_n \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}.$$

Outliers: Outliers can violate the normality assumption. Options: 1) Remove if justified as errors, 2) Use a robust method (trimmed mean), 3) Transform data, 4) Use bootstrap CI, 5) Report median with percentile CI instead.

Relation between CIs and hypothesis tests: For a two-sided test at level α, the $(1-\alpha)\%$ CI provides the same information. If μ_0 (null value) is outside the CI, reject H_0. Example: For H_0: $\mu = 0$ with $\alpha = 0.05$, if 95% CI is $(1.2, 3.4)$, reject H_0 since 0 is not in the interval.

A wider CI: A wider CI honestly reflects greater uncertainty (small n or large s); it does not necessarily mean that a wider CI mean the analysis is worse. A narrow CI from biased data is worse than a wide CI from good data. Width tells us about precision, not accuracy.

Difference between confidence intervals and credible intervals

- Frequentist CI: Fixed parameter, random interval. Interpretation: "If I repeated this, 95% of such intervals would contain the true value".
- Bayesian Credible Interval: Random parameter (has distribution), fixed interval. Interpretation: "Given my data and prior, there's 95% probability the parameter is in this interval".

Python code to finding CI

```python
import numpy as np
from scipy import stats
from typing import Iterable, Tuple

def confidence_interval(
    data: Iterable[float],
    confidence: float = 0.95
) -> Tuple[float, float]:
    """
    Compute a confidence interval for the population mean.

    Parameters
    ----------
    data : Iterable[float]
        Sample observations.
    confidence : float, default=0.95
        Confidence level between 0 and 1.

    Returns
    -------
    (lower, upper) : Tuple[float, float]
```

```
    Confidence interval bounds.
"""
data = np.asarray(data)
n = data.size

if n < 2:
    raise ValueError("At least two observations are required")

mean = data.mean()
se = stats.sem(data)

if n < 30:
    return stats.t.interval(confidence, df=n - 1, loc=mean, scale=se)

z = stats.norm.ppf(1 - (1 - confidence) / 2)
margin = z * se
return mean - margin, mean + margin
```

4 Hypothesis Testing

A hypothesis test is a formal statistical procedure used to make a decision about a population parameter based on sample data.

Null Hypothesis (H_0)**:** Default position (no effect/difference)
Alternative Hypothesis (H_1)**:** What we're trying to prove
Significance level α**:** Threshold for rejecting H_0 (commonly 0.05)
Type I error (α): False positive - Reject H_0 when it is true
Type II error (β): False negative - Failing to reject H_0 when it is false
Power ($1 - \beta$): The probability of correctly rejecting H_0 when it is false

p-value

Probability of observing data at least as extreme as what we actually observed, assuming the null hypothesis is true.

$$\text{p-value} = P(\text{data as extreme or more extreme} | H_0)$$

Basically we can divide p-value in three components:

p-value = The probability random chance would result in the observation + The probability of observing somethings else that is equally rare + The probability of observing something rare or more extreme

Example: Hypothesis Test for Coin Bias
A coin tossed 10 times, out of this head occurred only 1, is the coin biased?
Let X denote the number of heads in 10 tosses. Then

$$X \sim \text{Binomial}(10, p).$$

Hypotheses

$$H_0 : p = \tfrac{1}{2} \quad \text{vs} \quad H_1 : p \neq \tfrac{1}{2}.$$

Observed Value

$$x_{\text{obs}} = 1.$$

p-value

The probability random chance would result in the observation, $\mathbb{P}(X = 1) = \binom{10}{1}\left(\frac{1}{2}\right)^{10} = \frac{10}{1024}$
The probability of observing somethings else that is equally rare, $\mathbb{P}(X = 9) = \binom{10}{9}\left(\frac{1}{2}\right)^{10} = \frac{10}{1024}$

The probability of observing something rare or more extreme, $\mathbb{P}(X = 0) + \mathbb{P}(X = 10) = \binom{10}{0}\left(\frac{1}{2}\right)^{10} + \binom{10}{10}\left(\frac{1}{2}\right)^{10} = \frac{2}{1024}$
Under H_0, $X \sim \text{Bin}(10, 0.5)$. For a two-sided test,

$$p\text{-value} = \mathbb{P}(X \leq 1) + \mathbb{P}(X \geq 9).$$

$$\mathbb{P}(X \leq 1) = \binom{10}{0}\left(\tfrac{1}{2}\right)^{10} + \binom{10}{1}\left(\tfrac{1}{2}\right)^{10} = \frac{11}{1024}.$$

Thus,

$$p\text{-value} = \frac{22}{1024} \approx 0.0215.$$

Conclusion
Since p-value < 0.05, we reject H_0. There is evidence that the coin is biased.

Common Statistical Tests and Their Assumptions

Scenario	**Test**	**Assumptions**
Compare 2 means	t-test	Normality, equal variance (check)
Compare >2 means	ANOVA	Normality, equal variance, independence
Compare 2 proportions	z-test for proportions	Binomial, $np \geq 5$, $n(1-p) \geq 5$
Compare >2 proportions	Chi-square test	Expected counts ≥ 5
Test independence	Chi-square test of independence	Expected counts ≥ 5
Test goodness of fit	Chi-square GOF test	Expected counts ≥ 5
Compare medians (non-normal)	Mann–Whitney U (2 groups), Kruskal–Wallis (>2)	Independent samples
Paired data	Paired t-test	Differences normally distributed
Correlation	Pearson's r (linear), Spearman's ρ (monotonic)	Linearity for Pearson

p-Hacking

p-hacking refers to the practice of performing multiple statistical analyses and selectively reporting those that yield statistically significant results ($p < \alpha$), without correcting for multiple testing.

Multiple Testing Problem
If m independent hypothesis tests are conducted at significance level α, then

$$\mathbb{P}(\text{at least one false positive}) = 1 - (1 - \alpha)^m.$$

Example
For $\alpha = 0.05$ and $m = 20$,

$$1 - 0.95^{20} \approx 0.64.$$

Prevention
Common remedies include Bonferroni correction

$$\alpha^* = \frac{\alpha}{m},$$

false discovery rate control, preregistration and replication.

False Discovery Rate (FDR)

FDR controls the expected proportion of false positives among all discoveries, making it ideal for large-scale multiple testing.
The false discovery rate (FDR) is defined as

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{\max(R,1)}\right],$$

where V is the number of false rejections and R is the total number of rejected hypotheses.

Benjamini–Hochberg Procedure

Let $p_1, p_2, \ldots, p_m$ be p-values from m hypothesis tests.

1. Order the p-values: $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$.
2. Find $k = \max\left\{i : p_{(i)} \le \frac{i}{m}q\right\}$.
3. Reject hypotheses $H_{(1)}, \ldots, H_{(k)}$.

Power Analysis

Power analysis determines how large a sample is needed to reliably detect a specified effect while controlling Type-I and Type-II errors.
Power analysis is the study of the probability that a statistical hypothesis test correctly rejects the null hypothesis when the alternative hypothesis is true.

Definition of Power

Let

- α denote the probability of a Type-I error,
- β denote the probability of a Type-II error.

The *power* of a test is defined as

$$\mathrm{Power} = 1 - \beta,$$

which represents the probability of correctly rejecting H_0 when H_1 is true.

Hypothesis Setup

Consider the one-sample mean testing problem:

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

Assume that the observations are independent and identically distributed with

$$X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2),$$

where σ^2 is known.

Test Statistic

The test statistic for the one-sample z-test is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Under the null hypothesis, $Z \sim \mathcal{N}(0,1)$.

Rejection Region

For a two-sided test at significance level α, the rejection region is

$$|Z| > z_{\alpha/2},$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution.

Power Function

Suppose the true mean is $\mu = \mu_1 \neq \mu_0$. Then the distribution of Z becomes

$$Z \sim \mathcal{N}\left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}, 1\right).$$

The power of the test is therefore

$$\begin{aligned}\text{Power}(\mu_1) &= \mathbb{P}_{\mu_1}\left(|Z| > z_{\alpha/2}\right) \\ &= 1 - \Phi\left(z_{\alpha/2} - \frac{|\mu_1 - \mu_0|\sqrt{n}}{\sigma}\right) + \Phi\left(-z_{\alpha/2} - \frac{|\mu_1 - \mu_0|\sqrt{n}}{\sigma}\right),\end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

Effect Size

The standardized effect size (Cohen's d) is defined as

$$d = \frac{\mu_1 - \mu_0}{\sigma}.$$

Typical benchmarks are:

$$d = 0.2 \text{ (small)}, \quad d = 0.5 \text{ (medium)}, \quad d = 0.8 \text{ (large)}.$$

Sample Size Determination

To achieve power $1 - \beta$ at significance level α, the required sample size is

$$n = \left(\frac{z_{\alpha/2} + z_{\beta}}{d}\right)^2.$$

Interpretation

A higher power is achieved by:

- increasing the sample size n,
- increasing the effect size,
- reducing the variance,
- using a larger significance level α.

5 Machine Learning Metrics

Machine learning metrics are quantitative measures used to evaluate the performance of models across different problem types such as classification, regression, ranking, recommendation, and clustering.

Classification Metrics

Confusion Matrix

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Accuracy is suitable for balanced datasets but can be misleading for imbalanced data.

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision measures how many predicted positives are actually correct.

Recall (Sensitivity)

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall measures how many actual positives are correctly identified.

F1 Score

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score balances precision and recall, especially useful for imbalanced datasets.

Specificity

$$\text{Specificity} = \frac{TN}{TN + FP}$$

ROC–AUC

The Receiver Operating Characteristic (ROC) curve plots the true positive rate (on the y-axis) against the false positive rate (on the x-axis), where

$$\text{FPR} = 1 - \text{Specificity} = \frac{FP}{FP + TN}.$$

The Area Under the Curve (AUC) represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.
What is the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative instance?

- AUC $= 1.0 \rightarrow$ Perfect ranking
- AUC $= 0.5 \rightarrow$ 50% chance $\rightarrow$ random
- AUC $< 0.5 \rightarrow$ Worse than random (inverted predictions)

Why ROC Curves Are Sensitive to Class Imbalance

ROC—AUC can be misleading on imbalanced data because FPR is dominated by the large number of negatives, masking poor precision.
The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), defined as

$$\text{TPR} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN}.$$

In highly imbalanced datasets, the number of negative samples (TN) is much larger than the number of positive samples. As a result, even a large number of false positives can yield a small false positive rate.

Illustrative Example
Assume a dataset with 1% positive samples and 99% negative samples:

$$TN = 9900, \quad FP = 100.$$

Then the false positive rate is

$$\text{FPR} = \frac{100}{100 + 9900} = 0.01.$$

Despite having 100 false positives, the ROC curve shows only a small increase in FPR.
However, the precision is

$$\text{Precision} = \frac{TP}{TP + FP}.$$

If $TP = 50$, then

$$\text{Precision} = \frac{50}{150} \approx 0.33,$$

This indicates poor predictive quality. ROC says "great", business says "bad".

Key Observation
ROC curves measure ranking performance and are insensitive to class prevalence. Therefore, ROC–AUC may present an overly optimistic view of model performance on imbalanced datasets.

Comparison with Precision–Recall Curves

Precision–Recall curves explicitly account for false positives and are more informative when dealing with rare positive classes.
Conclusion: For highly imbalanced datasets, Precision–Recall curves are generally preferred over ROC curves.

Precision–Recall AUC

The Precision–Recall curve plots precision against recall and is more informative than ROC–AUC for highly imbalanced datasets.

Regression Metrics

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

MAE is robust to outliers and easy to interpret.

Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

MSE penalizes large errors more heavily.

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\text{MSE}}$$

RMSE has the same units as the target variable.

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

R^2 measures the proportion of variance explained by the model.

Adjusted R^2

The coefficient of determination R^2 measures the proportion of variance in the response variable explained by the regression model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

A key limitation of R^2 is that it never decreases when additional predictors are added, even if those predictors are irrelevant.

Definition of Adjusted R^2

Adjusted R^2 corrects this limitation by penalizing model complexity:

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

where n is the number of observations and p is the number of predictors.

Why Adjusted R^2 Is Preferred?

Adjusted R^2 increases only if a newly added predictor improves the model more than would be expected by chance. Otherwise, it decreases. This makes it suitable for comparing regression models with different numbers of predictors.

Conclusion

Adjusted R^2 provides a more reliable measure of model quality than R^2 by balancing goodness of fit with model complexity.

Ranking and Recommendation Metrics

Precision@K

$$\text{Precision@K} = \frac{\text{Number of relevant items in top } K}{K}$$

Recall@K

$$\text{Recall@K} = \frac{\text{Number of relevant items in top } K}{\text{Total number of relevant items}}$$

Average Precision (AP)

$$\text{AP} = \frac{1}{|\mathcal{R}|} \sum_{k=1}^{N} \text{Precision@k} \cdot \mathbb{I}(\text{item at } k \text{ is relevant})$$

Mean Average Precision (MAP)

$$\text{MAP} = \frac{1}{Q} \sum_{q=1}^{Q} \text{AP}_q$$

Discounted Cumulative Gain (DCG)

$$\text{DCG@K} = \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

Normalized DCG (NDCG)

$$\text{NDCG@K} = \frac{\text{DCG@K}}{\text{IDCG@K}}$$

NDCG accounts for both ranking position and graded relevance.

Clustering Metrics

Silhouette Score

$$s = \frac{b - a}{\max(a, b)}$$

where a is the average intra-cluster distance and b is the average nearest-cluster distance.

Davies–Bouldin Index

The Davies–Bouldin Index evaluates cluster separation, where lower values indicate better clustering.

Adjusted Rand Index (ARI)

ARI measures the similarity between predicted clusters and ground-truth labels.

Probabilistic Metrics

Log Loss (Cross-Entropy)

$$\text{LogLoss} = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

Log loss penalizes confident incorrect predictions.

Brier Score

$$\text{Brier Score} = \frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2$$

Metric Selection Guidelines

- Use Precision, Recall, and F1-score for imbalanced classification.
- Use MAE or RMSE for regression problems.
- Use NDCG, MAP, and Recall@K for recommender systems.
- Use Silhouette Score for clustering evaluation.

Remark: The choice of metric should reflect the business cost of errors rather than relying solely on mathematical convenience.