Probabilistic Modelling

Jingyuan Sun
ChatGPT Claude

Dec, 2024

1 Probability Theory and Statistics Recap

1.1 Probability Space

A probability space is defined as a triple $(\Omega, \mathcal{F}, P)$, where:

- Ω is the **sample space**, the set of all possible outcomes.
- $\mathcal{F}$ is the **event space**, a σ-algebra of subsets of Ω, containing all events of interest.
- P is the **probability measure**, a function $P : \mathcal{F} \to [0, 1]$ that assigns a probability to each event, satisfying:
  1. Non-negativity: $P(A) \geq 0$ for all $A \in \mathcal{F}$,
  2. Normalization: $P(\Omega) = 1$,
  3. Countable additivity: If $\{A_i\}_{i=1}^{\infty}$ are disjoint, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

Example: In the experiment of rolling a die:

- $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- $\mathcal{F} = 2^{\Omega}$ (the power set of Ω).
- P assigns equal probability $P(\{i\}) = \frac{1}{6}$ to each outcome i.

1.2 Probability Distributions

1.2.1 Discrete Distributions

Binomial Distribution: The binomial distribution models the number of successes in n independent Bernoulli trials, each with success probability p. Its probability mass function is:

$$X \sim \text{Binomial}(n, p), \quad P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Example: Suppose a coin with a probability $p = 0.6$ of landing heads is flipped $n = 10$ times. What is the probability of exactly 6 heads?

$$P(X = 6) = \binom{10}{6} (0.6)^6 (0.4)^4 = \frac{10!}{6! \cdot 4!} \cdot (0.6)^6 \cdot (0.4)^4 \approx 0.2508.$$

Bernoulli Distribution: The Bernoulli distribution is a special case of the binomial distribution with $n = 1$. Its probability mass function is:

$$X \sim \text{Bernoulli}(p), \quad P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

Example: A light bulb has a 70% chance of functioning. What is the probability that it works (1) or fails (0)?

$$P(X = 1) = 0.7, \quad P(X = 0) = 1 - 0.7 = 0.3.$$

Poisson Distribution: The Poisson distribution models the number of events occurring in a fixed interval of time or space, given the mean rate λ. Its probability mass function is:

$$X \sim \text{Poisson}(\lambda), \quad P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \geq 0.$$

Example: On average, 5 customers arrive at a store per hour. What is the probability of exactly 3 customers arriving in an hour?

$$P(X = 3) = \frac{5^3 e^{-5}}{3!} = \frac{125 \cdot e^{-5}}{6} \approx 0.1404.$$

Hypergeometric Distribution: The hypergeometric distribution models the number of successes X in k draws without replacement from a population of m successes and n failures:

$$X \sim \text{Hypergeometric}(m, n, k), \quad P(X = x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}, \quad x = 0, 1, \ldots, \min(k, m).$$

Example: A box contains 7 defective and 13 functional items. If 5 are selected at random, what is the probability of exactly 2 defective items?

$$P(X = 2) = \frac{\binom{7}{2}\binom{13}{3}}{\binom{20}{5}} = \frac{21 \cdot 286}{15504} \approx 0.386.$$

1.2.2 Continuous Distributions

Normal Distribution: The normal distribution, or Gaussian distribution, is defined by its mean μ and variance σ^2. Its probability density function (PDF) is:

$$X \sim \mathcal{N}(\mu, \sigma^2), \quad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Example: If $X \sim \mathcal{N}(0, 1)$, find the probability $P(-1 \leq X \leq 1)$.

$$P(-1 \leq X \leq 1) = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx \approx 0.6826.$$

(The result comes from standard normal tables or numerical integration.)

Beta Distribution: The beta distribution models probabilities and is parameterized by $\alpha > 0$ and $\beta > 0$. Its PDF is:

$$X \sim \text{Beta}(\alpha, \beta), \quad f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad x \in [0, 1],$$

where $B(\alpha, \beta)$ is the beta function.

Beta Function: The beta function, denoted as $B(\alpha, \beta)$, is a special function defined for $\alpha > 0$ and $\beta > 0$. It is expressed as:

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt,$$

where $t \in [0, 1]$.

The beta function can also be expressed in terms of the gamma function:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)},$$

where $\Gamma(x)$ is the gamma function defined by:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt.$$

Explanation: The beta function arises in the computation of the Beta distribution's normalization constant and is frequently used in probability and statistics. It represents the area under a curve defined by $t^{\alpha-1}(1-t)^{\beta-1}$ over the interval $[0, 1]$.

Example: Compute $B(2, 3)$:

$$B(2, 3) = \int_0^1 t^{2-1}(1-t)^{3-1}\, dt = \int_0^1 t(1-t)^2\, dt.$$

To solve:

$$\int_0^1 t(1-t)^2\, dt = \int_0^1 (t - 2t^2 + t^3)\, dt = \left[\frac{t^2}{2} - \frac{2t^3}{3} + \frac{t^4}{4}\right]_0^1 = \frac{1}{2} - \frac{2}{3} + \frac{1}{4}.$$

Simplifying:

$$B(2, 3) = \frac{6}{24} - \frac{16}{24} + \frac{6}{24} = \frac{6}{24} = \frac{1}{4}.$$

Thus, $B(2, 3) = \frac{1}{4}$.

Gamma Distribution: The gamma distribution is parameterized by shape $\alpha > 0$ and rate $\beta > 0$. Its PDF is:

$$X \sim \text{Gamma}(\alpha, \beta), \quad f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, \quad x > 0.$$

Example: If $X \sim \text{Gamma}(3, 2)$, find $P(1 \leq X \leq 2)$:

$$P(1 \leq X \leq 2) = \int_1^2 \frac{2^3 x^2 e^{-2x}}{\Gamma(3)} dx.$$

(This integral can be evaluated numerically or using software.)

Student's t-Distribution: The t-distribution is used in estimating population means. Its PDF is:

$$X \sim t_k, \quad f(x) \propto \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}},$$

where k is the degrees of freedom.

Example: If $X \sim t_5$, find $P(-2 \leq X \leq 2)$. Use the cumulative distribution function (CDF) or numerical methods to find:

$$P(-2 \leq X \leq 2) \approx 0.919.$$

Chi-Squared Distribution: The chi-squared distribution is a special case of the gamma distribution, used in hypothesis testing. Its PDF is:

$$X \sim \chi^2_k, \quad f(x) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}, \quad x \geq 0.$$

Example: If $X \sim \chi^2_3$, find $P(X \geq 4)$:

$$P(X \geq 4) = \int_4^\infty \frac{x^{1/2}e^{-x/2}}{2^{3/2}\Gamma(3/2)} dx.$$

(This integral can be evaluated numerically or using chi-squared tables.)

1.3 Expectation

The expectation of a random variable X is:

$$\mathbb{E}[X] = \begin{cases} \sum_x xP(X = x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} xf(x)\,dx, & \text{if } X \text{ is continuous.} \end{cases}$$

Example: The expected value of a die roll is:

$$\mathbb{E}[X] = \sum_{i=1}^{6} i \cdot \frac{1}{6} = 3.5.$$

1.4 Summary Statistics

Mean: The mean is the average value:

$$\mu = \mathbb{E}[X].$$

Standard Deviation: The standard deviation measures spread:

$$\sigma = \sqrt{\mathbb{E}[(X - \mu)^2]}.$$

Skewness: Skewness quantifies asymmetry:

$$\text{Skewness} = \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^3\right].$$

Median: The median is the value that splits the data into two equal parts.

Full Width Half Maximum (FWHM): For a unimodal distribution, FWHM is the width at half the maximum PDF value.

Covariance: The covariance between X and Y is:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

Correlation: The correlation coefficient is:

$$r(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

2 Sampling and Estimation

2.1 Sampling

Sampling is the process of selecting a subset (a sample) of individuals or observations from a larger population to estimate population parameters. Sampling allows us to analyze data when the population is too large to study in its entirety.

Simple Random Sampling

A **simple random sample** is a sample where each subset of a fixed size n has an equal probability of being selected. Sampling can occur **with replacement** (selected individuals can appear multiple times) or **without replacement** (each individual appears only once).

Example: Consider a population of 100 individuals. Selecting 10 individuals randomly ensures each possible group of 10 has an equal chance of selection. In practice, this ensures unbiased estimates of population parameters.

2.2 Estimators

An **estimator** is a statistic calculated from a sample used to infer the value of a population parameter. For a parameter θ, the estimator $\hat{\theta}$ is a random variable whose distribution depends on the sampling procedure.

Properties of Estimators

- **Unbiasedness:** An estimator $\hat{\theta}$ is unbiased if $\mathbb{E}[\hat{\theta}] = \theta$, where θ is the true population parameter.

- **Variance:** The variance of an estimator $\hat{\theta}$ measures its spread and is defined as $\mathrm{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$.

- **Mean Squared Error (MSE):** $\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$, which decomposes into bias and variance terms: $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (\mathrm{Bias}(\hat{\theta}))^2$.

Example: Sample Mean as an Estimator The sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an unbiased estimator of the population mean μ. Its variance is $\frac{\sigma^2}{n}$, where σ^2 is the population variance.

2.3 Sampling Strategies

Cluster Sampling

Cluster sampling divides the population into clusters and randomly selects clusters to study. Sampling can involve:

- **One-stage cluster sampling:** Entire clusters are sampled.
- **Two-stage cluster sampling:** A sample is drawn within each selected cluster.

Example: In a city with 100 schools, we randomly select 10 schools and then survey every student in the chosen schools (one-stage) or a subset of students from each chosen school (two-stage).

Stratified Sampling

Stratified sampling divides the population into distinct subgroups, called strata, based on specific characteristics (e.g., age, income). A random sample is then taken from each stratum. This ensures representation from each subgroup and improves the precision of the estimates compared to simple random sampling.

Definition: If the population is divided into k strata, with the i-th stratum containing N_i individuals, and the total population size is $N = \sum_{i=1}^{k} N_i$, then:

- Proportional allocation: Sample size from the i-th stratum is $n_i = n \cdot \frac{N_i}{N}$, where n is the total sample size.
- Optimal allocation (for minimizing variance): Sample size from the i-th stratum is proportional to $\frac{N_i \sigma_i}{\sum_{j=1}^{k} N_j \sigma_j}$, where σ_i is the standard deviation within the i-th stratum.

Example: Consider a population divided into three age groups: children, adults, and seniors. If children make up 30% of the population, adults 50%, and seniors 20%, a stratified sample of size $n = 100$ would include 30 children, 50 adults, and 20 seniors.

Sampling a Diverse Population

When sampling from a diverse population, it is essential to ensure that all subgroups are represented to avoid bias. This is particularly relevant in studies where the population is heterogeneous in terms of demographics, income levels, or geographic location.

Key Considerations:

- Ensure adequate representation of smaller subgroups to capture their variability.
- Use stratified sampling or oversampling techniques if certain groups are underrepresented.

Example: In a survey of healthcare preferences across urban and rural areas, a simple random sample may underrepresent rural participants. Stratified sampling ensures proportional representation of urban and rural residents, allowing for more accurate insights into population-wide preferences.

2.4 Resampling

Resampling involves repeatedly drawing samples from the original data to estimate variability or adjust for bias.

The Jackknife Estimator

For a dataset of n observations, the jackknife estimator recalculates the statistic of interest, excluding one observation at a time:

$$\hat{\theta}_{(i)} \quad \text{(statistic recalculated without the i-th observation).}$$

The jackknife estimate is then:

$$\hat{\theta}_J = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}.$$

Bootstrap Method

The bootstrap method involves sampling with replacement from the data to generate multiple datasets (replicates), calculating the statistic for each replicate, and using the distribution of these statistics to estimate variability or confidence intervals.

Example: Given a sample of size n, we create B bootstrap samples (each of size n), compute the mean for each sample, and use these means to estimate the confidence interval for the population mean.

3 Probabilistic Models

3.1 Linear Regression

3.1.1 Model Definition

The linear regression model assumes the relationship between a dependent variable y and independent variables $x_1, x_2, \ldots, x_p$:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

or equivalently in matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where:

- $\mathbf{y} \in \mathbb{R}^{n\times 1}$: Vector of observed outcomes.

- $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$: Design matrix with the first column as 1 (intercept) and remaining columns as predictors.
- $\boldsymbol{\beta} \in \mathbb{R}^{(p+1)\times 1}$: Coefficient vector.
- $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times 1}$: Error vector, assumed to satisfy $\mathbb{E}[\boldsymbol{\epsilon}] = 0$, $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

3.1.2 Objective Function: Least Squares

The coefficients $\boldsymbol{\beta}$ are estimated by minimizing the sum of squared residuals:

$$e(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Expanding the quadratic form:

$$e(\boldsymbol{\beta}) = \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}.$$

3.1.3 Optimization: First-Order Condition

To minimize $e(\boldsymbol{\beta})$, we compute the gradient with respect to $\boldsymbol{\beta}$:

$$\frac{\partial e(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}.$$

Setting the gradient to zero:

$$-2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = 0.$$

Simplify:

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y}.$$

3.1.4 Solution: OLS Estimator

Assuming $\mathbf{X}^\top\mathbf{X}$ is invertible, the solution for $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

3.1.5 Predicted Values and Residuals

Using the estimated coefficients $\hat{\boldsymbol{\beta}}$, the predicted values $\hat{\mathbf{y}}$ and residuals $\mathbf{r}$ are:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y},$$

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}.$$

3.1.6 Variance of the Error Term

The variance of the error term σ^2 is estimated using the residual sum of squares:

$$\hat{\sigma}^2 = \frac{\mathbf{r}^\top\mathbf{r}}{n-p-1} = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-p-1}.$$

3.1.7 Statistical Properties of OLS

- If $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, then $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$.
- The variance of $\hat{\boldsymbol{\beta}}$ is given by:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}.$$

3.1.8 Example: Simple Linear Regression

Suppose we want to predict the Gross National Product (GNP) y based on Employment x:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Given the data:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

we calculate:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

The predictions are:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \ldots, n.$$

3.2 Logistic Regression

For binary classification ($y \in \{0, 1\}$), the **logistic regression model** uses:

$$z_i = \sum_{j=1}^{p} x_{ij}\beta_j, \quad P(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + e^{-z_i}},$$

where z_i represents the *log-odds*. The parameters $\boldsymbol{\beta}$ are estimated via **maximum likelihood estimation (MLE)**.

Example: Predicting whether a customer will purchase a product ($y = 1$) based on age and income.

3.3 Time Series Models

3.4 Introduction to Time Series

A time series is a sequence of observations indexed by time t. It is denoted as:

$$\{X_t\}_{t=1}^{T},$$

where $t = 1, 2, \ldots, T$ represents time.

3.4.1 Stationarity

A time series is weakly stationary if it satisfies:

- Constant mean: $\mu = \mathbb{E}[X_t]$,
- Constant variance: $\sigma^2 = \mathrm{Var}(X_t)$,
- Autocovariance depends only on lag: $\gamma_k = \mathrm{Cov}(X_t, X_{t-k})$.

3.4.2 Autoregressive (AR) Model

The AR model of order p is defined as:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t,$$

where ϕ_i are the AR coefficients, $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.

3.4.3 Moving Average (MA) Model

The MA model of order q is defined as:

$$X_t = c + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t,$$

where θ_j are the MA coefficients.

3.4.4 Autoregressive Moving Average (ARMA) Model

Combining the AR and MA models, the ARMA model of order (p, q) is:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t.$$

3.4.5 Autoregressive Integrated Moving Average (ARIMA) Model

For non-stationary time series, the ARIMA model includes differencing:

$$\nabla^d X_t = (1 - B)^d X_t,$$

where $\nabla^d X_t$ is the d-th differenced series, B is the backward shift operator.

The ARIMA model of order (p, d, q) is expressed as:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t,$$

where p, d, q represent the orders of autoregression, differencing, and moving average.

3.4.6 Example: ARIMA(1, 1, 1)

Suppose we have a time series with:

- $p = 1$: one autoregressive term,
- $d = 1$: first-order differencing,
- $q = 1$: one moving average term.

The model is:

$$\nabla X_t = \phi_1 X_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t,$$

where $\nabla X_t = X_t - X_{t-1}$.

Given $\phi_1 = 0.8, \theta_1 = 0.5$, and $\epsilon_t \sim \mathcal{N}(0, 1)$, we can forecast future values using the past observations.

3.4.7 Seasonality and Decomposition

A time series can also exhibit:

- Trend: long-term movement,
- Seasonality: periodic fluctuations,
- Noise: random variation.

The Seasonal-Trend Decomposition (STL) method splits a series into components:

$$X_t = \text{Trend}_t + \text{Seasonality}_t + \text{Residual}_t.$$

3.5 Polynomial Regression

A **polynomial regression** model includes higher-order terms of the predictors:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i.$$

Example: Modeling the trajectory of a projectile with quadratic terms.

3.6 Underfitting and Overfitting

- **Underfitting**: Model is too simple, failing to capture data complexity.
- **Overfitting**: Model is too complex, fitting noise instead of patterns.

Example: Predicting exam scores with too few or too many predictors.

3.7 Regularisation

Regularisation reduces model complexity by adding penalties to the loss function.

Ridge Regression:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

This introduces a penalty term λ, controlling the magnitude of coefficients.

Example: Predicting house prices using correlated predictors, where regularisation avoids overfitting.