

Lecture 1: Mathematical basics for statistics

*Lecturer: Yiqiao Zhong**Last update: 21:09, 2024-09-09*

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

Welcome to STAT 709 “Mathematical Statistics”! Statistics, and more broadly, data science is an ever-evolving discipline. I will try to introduce basic ideas and theories of statistics from a modern perspective whenever possible. Caution: Typos and incorrect statements should be not unexpected.

The first two lectures provide a recap of some basic mathematics you need to know. If you are not aware of the results, please reconsider taking this course.

1 Set theory

Throughout this course, we will mostly consider Euclidean spaces, unless specified otherwise. A d -dimensional Euclidean space is denoted by $\mathbb{R}^d$. The Euclidean distance is denoted by $\|\cdot\|$ or $\|\cdot\|_2$.

- **Open set:** A set $U \subset \mathbb{R}^d$ is open if $\forall x \in U$, there exists a positive number ϵ such that $y \in U$, $\forall \|y - x\| < \epsilon$.
- **Closed set:** A set $K \subset \mathbb{R}^d$ is closed if the complement K^c is open.
- **Union of open sets:** suppose that we have a collection of open sets $(U_i)_{i \in \mathcal{I}}$ indexed by i , then the union $\cup_{i \in \mathcal{I}}$ is also open. Note that $\mathcal{I}$ can be an infinite index set.
- **Countable set:** a set is countable if it is finite or it can be made in one-to-one correspondence with the set of natural numbers. For example, the set of rational numbers $\mathbb{Q}$ is countable. The product of two countable sets is countable.
- **Closure:** The closure of a set $S \subset \mathbb{R}^d$ is the set of all limit points of S . For example, the closure of $\mathbb{Q}$ is $\mathbb{R}$, and the closure of $\mathbb{Q} \times \cdots \times \mathbb{Q}$ is $\mathbb{R}^d$.

2 Limit

One of the prerequisites is basic proof-based calculus. For example, you are expected to know how to prove things like “the subsequence of a converging sequence converges to the same limit”.

- **Small-o notation.** For sequences $(a_n)_{n \geq 1}, (b_n)_{n \geq 1}$, we write $a_n = o(b_n)$ if $\lim_{n \rightarrow \infty} a_n/b_n = 0$. The $o(1)$ is interpreted as a sequence that converges to zero. Under this notation, we may write $a_n = (1 + o(1))b_n$ if $\lim_{n \rightarrow \infty} a_n/b_n = 1$.
- **Big-O notation.** For sequences $(a_n)_{n \geq 1}, (b_n)_{n \geq 1}$, we write $a_n = O(b_n)$ if there exists constants $C, N > 0$ such that $|a_n| \leq Cb_n$ holds for all $n > N$. If a sequence depends on many indices or parameters, the “constant” C means that it does not depend on n , but may depend on other indices or parameters.

3 Probability

- **Gaussian density.** The probability density function of the standard normal variable is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

A basic property is that the density integrates to one, namely $\int_{-\infty}^{\infty} f(x) = 1$. This is based on the following fact

$$\int_{-\infty}^{\infty} e^{-x^2/2} = \sqrt{2\pi}.$$

- **Gaussian tail probability (Mills ratio).** Suppose that G is a standard normal random variable $N(0, 1)$. Then for $t > 0$,

$$\frac{t}{\sqrt{2\pi}(t^2 + 1)} e^{-t^2/2} \leq P(G > t) \leq \frac{1}{\sqrt{2\pi}t} e^{-t^2/2}.$$

We can prove this inequality by using $f'(x) + xf(x) = 0$ where f is the Gaussian density and integration by parts.

- **Linear transform of normal variables.** Suppose $G = (G_1, \dots, G_n)^\top$ where $G_1, \dots, G_n \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$, namely $G_1, G_2, \dots, G_n$ be independent and identically distributed (i.i.d.) standard normal random variables, and $A \in \mathbb{R}^{m \times n}$. Then $AG \sim N(0, AA^\top)$ is still a normal random vector. In particular, if $AA^\top = I_m$, then AG is a standard normal random vector.

4 Combinatorics

Recall that the notation for combination is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

where $n! = 1 \cdot 2 \cdot \dots \cdot n$ is the factorial.

- Stirling's formula: $\lim_{n \rightarrow \infty} n! / [\sqrt{2\pi n} (n/e)^n] = 1$.

Example 1. Suppose that $X_1, \dots, X_n \in \{0, 1\}$ are independent Bernoulli random variables. Assume n is an even number and k is a fixed integer. Then

$$P(X_1 + \dots + X_n = \frac{n}{2} + k) = \frac{1}{2^n} \binom{n}{n/2 + k} = \frac{1}{2^n} \frac{n!}{(n/2 + k)!(n/2 - k)!} = (1 + o(1)) \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{n}} \exp(-2k^2/n).$$

More generally, we define the entropy function as

$$h(\gamma) = \gamma \log \frac{1}{\gamma} + (1 - \gamma) \log \frac{1}{1 - \gamma}$$

for $\gamma \in (0, 1)$.

Example 2. Suppose that $\gamma \in (0, 1)$ is fixed, and m is an integer of the form $m = (\gamma + o(1))n$. Then (exercise)

$$\log \binom{n}{m} = (h(\gamma) + o(1))n, \quad \text{as } n \rightarrow \infty.$$

5 Linear algebra

Let $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{m \times k}$, $C \in \mathbb{R}^{k \times n}$ be matrices.

- **Trace.** The trace of a square matrix M is defined by $\text{Tr}(M) = \sum_i M_{ii}$. It satisfies

$$\text{Tr}(ABC) = \text{Tr}(BCA).$$

Example 3. Suppose that $X \in \mathbb{R}^d$ has zero mean and covariance matrix Σ , and $M \in \mathbb{R}^{d \times d}$ is a matrix. Then

$$\mathbb{E}[X^\top M X] = \text{Tr}(M \Sigma).$$

Let $A \in \mathbb{R}^{n \times n}$ be a real symmetric matrix. Then A has n real eigenvalues $\lambda_1 \geq \dots \geq \lambda_n$. We may write $\lambda_{\max}(A)$ to denote λ_1 .

- **Variational formulation.** The largest eigenvalue satisfies

$$\lambda_{\max}(A) = \max_{x: \|x\|_2=1} x^\top A x.$$

- **Min-max formula.** A more general formula for the eigenvalues is

$$\begin{aligned} \lambda_i(A) &= \max_{V: \dim(V)=i} \min_{x \in V: \|x\|_2=1} x^\top A x \\ &= \max_{V: \dim(V)=n-i+1} \min_{x \in V: \|x\|_2=1} x^\top A x \end{aligned}$$

for all $1 \leq i \leq n$, where V ranges from all subspaces of the indicated dimension.

- **Weyl's inequality.** A consequence of the min-max formula is Weyl's inequality. Let $A, B \in \mathbb{R}^{n \times n}$ be real symmetric matrices. Then for all $i \leq n$,

$$|\lambda_i(A+B) - \lambda_i(A)| \leq \|B\|_{\text{op}}, \quad \text{where } \|B\|_{\text{op}} = \max\{|\lambda_1(B)|, |\lambda_n(B)|\}.$$

Similar results hold for Hermitian matrices, which are matrices in the complex domain that satisfy $\bar{A}^\top A$ where $\bar{A}$ is the conjugate.

For general matrix $A \in \mathbb{R}^{n \times m}$, the singular value decomposition (SVD) $A = U S V^\top$ where $U \in \mathbb{R}^{n \times r}$, $V \in \mathbb{R}^{m \times r}$ are orthonormal matrices, and $S = \text{diag}\{s_1, \dots, s_r\}$ are singular values. Here $r = \text{rank}(A) \leq \min(n, m)$. The convention is $s_1 \geq s_2 \geq \dots \geq s_r > 0$, and we may denote $s_{\max}(A) = s_1$.

- **Connection between SVD and eigendecomposition.** The nonzero eigenvalues of $AA^\top$ are given by $[s_1(A)]^2, \dots, [s_r(A)]^2$. The same result holds for $A^\top A$. The corresponding eigenvectors of $AA^\top$ are given by columns of U , and corresponding eigenvectors of $A^\top A$ are given by columns of V .
- **Variational formulation.** The largest singular value of A is

$$s_1(A) = \max_{\substack{x \in \mathbb{R}^n: \|x\|_2=1 \\ y \in \mathbb{R}^m: \|y\|_2=1}} x^\top A y = \max_{y \in \mathbb{R}^m: \|y\|_2=1} \|A y\|_2$$

The top singular value of a matrix is also known as the **operator norm**, denoted by $\|A\|_{\text{op}}$. By convention, we denote by $O(n)$ the space of all orthogonal matrices, and by $O(n, r)$ the space of orthonormal matrices of size $n \times r$. Under this notation, $U \in O(n, r)$ and $V \in O(m, r)$.

Example 4. Let $X \in \mathbb{R}^{n \times d}$ be matrix data of sample size n , each containing d features (aka attributes, covariates). Suppose $n > d$. We apply SVD¹ to X , yielding $X = U S V^\top$. The orthogonal $V = [v_1, \dots, v_d] \in O(d)$ provides a new coordinate system for the data.

¹For simplicity, we assume that it is already pre-processed including centering.

1. **Principal components (PCs):** Xv_1 contains the 1st PCs, Xv_2 contains the 2nd PCs, etc. Note that $US = XV$.

2. **Principal directions:** v_1, v_2 are the 1st, 2nd principal directions, etc.

It is well known that PCA is a **dimension reduction** method because it gives the best low-rank approximation:

$$\|X - U_r S_r V_r^\top\|_{\text{op}} = \min_{B: \text{rank}(B) \leq r} \|X - B\|_{\text{op}}$$

where $U_r = [u_1, \dots, u_r]$, $S_r = \text{diag}(S_{11}, \dots, S_{rr})$, $V_r = [v_1, \dots, v_r]$.

6 Vector norms, inequalities

Let $x, y \in \mathbb{R}^n$ be vectors and $\gamma \in \mathbb{R}$ be a scalar. The norm $\|\cdot\|$ of a vector satisfies (i) $\|x\| = 0$ iff $x = 0$ (ii) $\|\gamma x\| = |\gamma| \cdot \|x\|$, (iii) $\|x + y\| \leq \|x\| + \|y\|$. Vector norms generalize the Euclidean distance. Some important examples.

- **ℓ_1 -norm.** $\|x\|_1 = \sum_{i=1}^n |x_i|$.
- **ℓ_∞ -norm.** $\|x\|_\infty = \max_{i=1}^n |x_i|$.
- **ℓ_q -norm.** More generally, given $q \geq 1$, we define $\|x\|_q = (\sum_{i=1}^n |x_i|^q)^{1/q}$. Note that $\|x\|_\infty = \lim_{q \rightarrow \infty} \|x\|_q$.

Define the inner product

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i.$$

- **Cauchy-Schwarz inequality.**

$$\left| \sum_{i=1}^n x_i y_i \right| \leq \left(\sum_{i=1}^n x_i^2 \right)^{1/2} \cdot \left(\sum_{i=1}^n y_i^2 \right)^{1/2}$$

Or equivalently, $|\langle x, y \rangle| \leq \|x\| \|y\|$. We can define $\cos(\theta) = \langle x, y \rangle / \|x\|_2 \cdot \|y\|_2$. In two dimensions, θ is exactly the angle between the two vectors x, y .

Perhaps one of the most widely used inequalities, it can be used to verify the triangle inequality of the Euclidean norm.

$$\|x + y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 + 2\langle x, y \rangle \leq \|x\|_2^2 + \|y\|_2^2 + 2\|x\|_2 \|y\|_2 = (\|x\|_2 + \|y\|_2)^2.$$

- **Hölder's inequality.** It is a generalization of the Cauchy-Schwarz inequality. Let $q, r \geq 1$ be such that $(1/q) + (1/r) = 1$. Then

$$\left| \sum_{i=1}^n x_i y_i \right| \leq \left(\sum_{i=1}^n |x_i|^q \right)^{1/q} \cdot \left(\sum_{i=1}^n |y_i|^r \right)^{1/r}$$

Hölder's inequality can be used to verify the triangle inequality of ℓ_q -norm.

- **Norm conversion.**

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2 \leq n \|x\|_\infty.$$

7 Functional analysis

We may generalize the notation of vectors to infinite-dimensional space. Suppose that we have a function $f : \mathbb{R}^r \rightarrow \mathbb{R}$. We can choose a grid $t_1, \dots, t_M \in \mathbb{R}^r$ and set $x_1 = f(t_1), \dots, x_M = f(t_M)$. By choosing a very dense grid and letting $M \rightarrow \infty$, the vector norm can be generalized to functions. Of course, there are technical details when we take the limit; nevertheless, we may view vectors as a discrete version of functions.

Let $I \subset \mathbb{R}$ be a set and $h(x) > 0$ be a function² on I satisfying $\int_I h(x) = 1$. Define

$$\|f\|_q = \left(\int_I |f(x)|^q h(x) dx \right)^{1/q}$$

and

$$\langle f, g \rangle = \int_I f(x)g(x)h(x) dx. \quad (1)$$

The results about vector norms generalize to functions. For example,

$$\left| \int_I f(x)g(x)h(x) dx \right| \leq \left(\int_I |f(x)|^2 h(x) dx \right)^{1/2} \cdot \left(\int_I |g(x)|^2 h(x) dx \right)^{1/2}.$$

Orthonormal basis. We may also generalize linear algebra from finite-dimensional spaces to functions. Let $\mathcal{H}$ be the space of square-integrable functions³, i.e., $f \in \mathcal{H}$ if and only if

$$\|f\|_2^2 = \int_I [f(x)]^2 h(x) dx < \infty.$$

For example, all bounded functions satisfy this inequality. A nice property is that $\mathcal{H}$ with the inner product (??) is a Hilbert space. We can find a sequence of functions $\beta_1, \beta_2, \dots \in \mathcal{H}$ such that $\|\beta_k\|_2 = 1$ and $\langle \beta_j, \beta_k \rangle = 0$. These functions $(\beta_j)_{j=1}^\infty$ generalize the notation of orthonormal vectors to the function space, and they satisfy the good properties such as below

- **Basis expansion.** For $f \in \mathcal{H}$, we have⁴

$$f = \sum_{j=1}^{\infty} \langle f, \beta_j \rangle \beta_j$$

- **Parseval's identity.** For $f \in \mathcal{H}$,

$$\|f\|_2^2 = \sum_{j=1}^{\infty} \langle f, \beta_j \rangle^2$$

Fourier series form an orthonormal basis for square-integrable periodic functions. Let $I = [-\pi, \pi]$ and $h(x) \equiv 1/(2\pi)$. Then $1, (\sqrt{2} \cos(nx))_{n \geq 1}, (\sqrt{2} \sin(nx))_{n \geq 1}$ give the orthonormal basis.

Hermite polynomials. They are polynomials of increasing degrees, for instance,

$$H_0(x) = 1, \quad H_1(x) = x, \quad H_2(x) = \frac{1}{\sqrt{2}}(x^2 - 1), \quad H_3(x) = \frac{1}{\sqrt{6}}(x^3 - 3x).$$

²For now, we assume that h is a piecewise continuous function. The results hold for measurable functions.

³Functions refer to Lebesgue/Borel measurable function, which we will introduce later.

⁴Infinite series means $\lim_{M \rightarrow \infty} \|f - \sum_{j=1}^M \langle f, \beta_j \rangle \beta_j\|_2 = 0$.

If $I = \mathbb{R}$ and $h(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ (normal density), then $(H_j(x))_{j \geq 0}$ forms an orthonormal basis. If we have a standard normal variable $G \sim N(0, 1)$, then we have

$$f(G) = \sum_{j=0}^{\infty} \langle f, H_j \rangle H_j(G)$$

This identity is useful because it decomposes f into a sum of components of increasing complexity (low-degree has low complexity).

8 Convexity

A set $S \subset \mathbb{R}^d$ is convex if $\lambda x + (1 - \lambda)y \in S$ for any $x, y \in S$ and $\lambda \in (0, 1)$. A function f is convex if

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y), \quad \text{for any } \lambda \in (0, 1) \text{ and for any } x, y$$

We say a function is strictly convex if $\leq$ is replaced by $<$. There are many equivalent definitions. For example, if f is twice differentiable, then

$$f \text{ is convex} \iff \nabla^2 f \text{ is positive semi-definite}$$

Connection to optimization. Convexity is important in optimization theory. A strictly convex function in a bounded closed set has a unique minimum. A convex problem (convex loss, convex domain) generally has polynomial-time algorithms.

Example 5. *The least-square loss*

$$\ell(\beta) = \|y - X\beta\|_2^2$$

is convex in β . Moreover, it is strictly convex if and only if $X^\top X$ is of full rank. The regularized loss

$$\ell_{\lambda, q}(\beta) = \|y - X\beta\|_2^2 + \lambda \|\beta\|_q^q$$

is strictly convex for $\lambda > 0$ and $q > 1$.

Some further results of interest.

- **KKT condition.** A fundamental result for convex optimization states that the first-order condition is strongly connected to minimum of a convex problem.
- **Subgradient.** If a convex function is not smooth, we can generalize the notation of gradients to subgradients. There are similar results to the smooth case, e.g., KKT condition.

Example 6. *The subgradient (subdifferential) of $f(x) = |x|$ is*

$$\partial f = \begin{cases} 1 & x > 0 \\ [-1, 1] & x = 0 \\ -1, & x < 0 \end{cases}$$

The subgradient of the regularized squared loss $\ell_{\lambda, 1}(\beta)$ is

$$\partial \ell_{\lambda, 1}(\beta) = -2X^\top(y - X\beta) + \lambda \partial \|\beta\|_1$$

and at the minimum point β , we have $0 \in \partial \ell_{\lambda, 1}(\beta)$.

9 Further readings

For a review of basic mathematical analysis as well as elementary functional analysis, a good reference is [?]. Terrance Tao's book on random matrix theory [?] provides a good review of probability theory and linear algebra. A standard reference to convex optimization is [?].

Most materials can be also found in various sources. Wikipedia is a good starting point for learning detailed proofs and generalizations.