Theory of Statistical Practice

Derek Li

Contents

1 Probability Review

1.1 Basic Definition

Definition 1.1. ***Random experiment*** is a mechanism producing an outcome (result) perceived as random or uncertain.

Definition 1.2. ***Sample space*** is a set of all possible outcomes of the experiment:

$$\mathcal{S} = \{\omega_1, \omega_2, \cdots\}.$$

Example 1.1. Waiting time until the next bus arrives: $\mathcal{S} = \{t : t \geqslant 0\}$.

1.2 Probability Function/Measure

Definition 1.3. Given a sample space $\mathcal{S}$, define $\mathcal{A}$ to be a collection of subsets (events) of $\mathcal{S}$ satisfying the following conditions:

1. $\mathcal{S} \in \mathcal{A}$;
2. $A \in \mathcal{A} \Rightarrow A^C \in \mathcal{A}$;
3. $A_1, A_2, \cdots \in \mathcal{A} \Rightarrow A_1 \cup A_2 \cup \cdots \in \mathcal{A}$.

If $\mathcal{S}$ is finite or countably infinite, then $\mathcal{A}$ could consist of all subsets of $\mathcal{S}$ including $\varnothing$.

Definition 1.4. The ***probability function*** (***measure***) P on $\mathcal{A}$ satisfies the following conditions:

1. $P(A) \geqslant 0, \forall A \in \mathcal{A}$;
2. $P(\varnothing) = 0$ and $P(\mathcal{S}) = 1$;
3. If $A_1, A_2, \cdots$ are disjoint (mutually exclusive) events, i.e., $A_i \cap A_j = \varnothing$ for $i \neq j$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Property 1.1. $P(A^C) = 1 - P(A)$.

Proof. $1 = P(\mathcal{S}) = P(A \cup A^C) = P(A) + P(A^C)$. □

Property 1.2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof. $P(A) = P(A \cap B) + P(A \cap B^C)$ and $P(A \cup B) = P(B) + P(A \cap B^C)$. □

Property 1.3. $P(A \cup B) \leqslant P(A) + P(B)$.

Property 1.4. In general,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots - (-1)^n P(A_1 \cap \cdots \cap A_n).$$

Property 1.5 (Bonferroni's Inequality)**.** In general,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} P(A_i).$$

1.3 Conditional Probability

Definition 1.5. The probability of A ***conditional*** on B is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

if $P(B) > 0$. Note that if $P(B) = 0$, we can still define $P(A|B)$ but we need to be more careful mathematically.

Theorem 1.1 (Bayes Theorem)**.** If $B_1, \cdots, B_k$ are disjoint events with $B_1 \cup \cdots \cup B_k = \mathcal{S}$, then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}.$$

1.4 Independence

Definition 1.6. Two events A and B are ***independent*** if

$$P(A \cap B) = P(A)P(B).$$

When $P(A), P(B) > 0$, we can also say

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B).$$

Events $A_1, \cdots, A_k$ are independent if

$$P\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} P(A_i).$$

1.5 Interpretation of Probability

- Long-Run frequencies: If we repeat the experiment many times, then $P(A)$ is the proportion of times the event A occurs.
- Degrees of belief (subjective probability): If $P(A) > P(B)$, then we believe that A is more likely to occur than B.
- Frequentist versus Bayesian statistical methods:
  - ∗ Frequentists: Pretend that an experiment is at least conceptually repeatable.
  - ∗ Bayesians: Use subjective probability to describe uncertainty in parameters and data.

1.6 Random Variable

Definition 1.7. ***Random variable*** is a real-valued function defined on a sample space $\mathcal{S}, X : \mathcal{S} \to \mathbb{R}$. In other words, for each outcome $\omega \in \mathcal{S}, X(\omega)$ is a real number.

Definition 1.8. The ***probability distribution*** of X depends on the probabilities assigned to the outcomes in $\mathcal{S}$.

Definition 1.9. The ***cumulative distribution function*** (CDF) of X is

$$F(x) = P(X \leqslant x) = P(\omega \in \mathcal{S} : X(\omega) \leqslant x).$$

We denote it $X \sim F$.

Property 1.6. CDF satisfies:
1. If $x_1 \leqslant x_2$, then $F(x_1) \leqslant F(x_2)$;
2. $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$;
3. F is right-continuous with left-hand limits:

$$\lim_{y \to x^+} F(y) = F(x), \lim_{y \to x^-} F(y) = F(x-) = P(X < x);$$

4. $P(X = x) = F(x) - F(x-)$.

Definition 1.10. If $X \sim F$ where F is a continuous function, then X is a ***continuous r.v.***, and we can typically find a non-negative ***probability density function*** (PDF) f s.t.

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t.$$

Definition 1.11. If X takes only a finite or countably infinite number of possible values, then X is a ***discrete r.v.***, and F is a step function. We can define its ***probability mass function*** (PMF) by

$$f(x) = F(x) - F(x-) = P(X = x).$$

1.7 Expected Value

Definition 1.12. Suppose X with PDF $f(x)$ and Y with PMF $f(y)$. We can define the ***expected value*** of $h(X)$ and $h(Y)$ by

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)\mathrm{d}x \text{ and } \mathbb{E}[h(Y)] = \sum_{y} h(y)f(y).$$

We can also write $h(x) = h^+(x) - h^-(x)$ where $h^+(x) = \max\{h(x), 0\}$ and $h^-(x) = \max\{-h(x), 0\}$, then $\mathbb{E}[h(X)] = \mathbb{E}[h^+(X)] - \mathbb{E}[h^-(X)]$:
1. If $\mathbb{E}[h^+(X)]$ and $\mathbb{E}[h^-(X)]$ are finite, then $\mathbb{E}[h(X)]$ is well defined.
2. If $\mathbb{E}[h^+(X)] = \infty$ and $\mathbb{E}[h^-(X)]$ is finite, then $\mathbb{E}[h(X)] = \infty$.
3. If $\mathbb{E}[h^+(X)]$ is finite and $\mathbb{E}[h^-(X)] = \infty$, then $\mathbb{E}[h(X)] = -\infty$.
4. If $\mathbb{E}[h^+(X)]$ and $\mathbb{E}[h^-(X)]$ are infinite, then $\mathbb{E}[h(X)]$ does not exist.

Example 1.2 (Expected Values of Cauchy Distribution)**.** X is a continuous r.v. with

$$f(x) = \frac{1}{\pi(1 + x^2)}, -\infty < x < \infty.$$

We have

$$\mathbb{E}[X^+] = \mathbb{E}[X^-] = \int_0^{\infty} \frac{x}{\pi(1 + x^2)}\mathrm{d}x = \lim_{x \to \infty} \frac{1}{2\pi} \ln(1 + x^2) = +\infty.$$

Thus, $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$ does not exist.

1.8 Independent Random Variable

Definition 1.13. R.v.s. $X_1, X_2, \cdots$ are ***independent*** if the events $[X_1 \in A_1], [X_2 \in A_2], \cdots$ are independent events for any $A_1, A_2, \cdots$.

If $X_1, \cdots, X_n$ are independent r.v.s. with PDF or PMF $f_1, \cdots, f_n$, then the joint PDF or PMF of $(X_1, \cdots, X_n)$ is

$$f(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_i(x_i).$$

Suppose $X_1, \cdots, X_n$ are independent r.v.s. with mean $\mu_1, \cdots, \mu_n$ and variance $\sigma_1^2, \cdots, \sigma_n^2$. Define $S = X_1 + \cdots + X_n$, then $\mathbb{E}[S] = \mu_1 + \cdots + \mu_n$ (which is true even if $X_1, \cdots, X_n$ are not independent) and $\mathrm{Var}[S] = \sigma_1^2 + \cdots + \sigma_n^2$.

1.9 Convergence of Random Variable

Theorem 1.2 (Markov's Inequality)**.** Suppose Y is a random variable with $\mathbb{E}[|Y|^r] < \infty$ for some $r > 0$, then

$$P(|Y| > \varepsilon) \leqslant \frac{\mathbb{E}[|Y|^r]}{\varepsilon^r}.$$

Proof. For any $\varepsilon > 0$,

$$\mathbb{E}[|Y|^r] = \mathbb{E}[|Y|^r I(|Y| \leqslant \varepsilon)] + \mathbb{E}[|Y|^r I(|Y| > \varepsilon)] \geqslant 0 + \varepsilon^r P(|Y| > \varepsilon),$$

then $P(|Y| > \varepsilon) \leqslant \frac{\mathbb{E}[|Y|^r]}{\varepsilon^r}$. □

Theorem 1.3 (Chebyshev's Inequality)**.**

$$P(|X - \mathbb{E}[X]| > \varepsilon) \leqslant \frac{\mathrm{Var}[X]}{\varepsilon^2}.$$

Proof. Take $r = 2, Y = X - \mathbb{E}[X]$ in Markov's Inequality. □

1.9.1 Convergence in Probability

Definition 1.14. A sequence of r.v.s. $\{Y_n\}$ ***converges in probability*** to a r.v. Y (denoted $Y_n \xrightarrow{p} Y$) if for each $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|Y_n - Y| > \varepsilon) = 0.$$

Typically, the limiting r.v. Y is a constant.

Theorem 1.4 (Weak Law of Large Numbers)**.** If $X_1, X_2, \cdots$ are independent r.v.s. with finite mean μ, then

$$\overline{X} = \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$

1.9.2 Convergence in Distribution

Definition 1.15. A sequence of r.v.s. $\{X_n\}$ ***converges in distribution*** to a r.v. X if

$$\lim_{n \to \infty} F_n(x) = F(x),$$

for every $x \in \mathbb{R}$ at which F is continuous. F_n and F are CDF of X_n and X, respectively.

Let $S_n = \sqrt{n}(\overline{X}_n - \mu)$, then we have $\mathbb{E}[S_n] = 0$ and $\mathrm{Var}[S_n] = \sigma^2$. $\{S_n\}$ is bounded in probability since

$$P(|S_n| > M) \leqslant \frac{\mathbb{E}[S_n^2]}{\varepsilon^2} = \frac{\sigma^2}{M^2} \to 0 \text{ as } M \to \infty.$$

Theorem 1.5 (Basic Central Limit Theorem)**.** If $X_1, X_2, \cdots$ are independent r.v.s. with common CDF F with finite mean and variance μ and σ^2, then

$$\lim_{n\to\infty} P(S_n \leqslant x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2\sigma^2}} \mathrm{d}t,$$

denoted as $S_n \xrightarrow{d} S \sim \mathcal{N}(0, \sigma^2)$.

As a consequence, the distribution of $\overline{X}_n$ is approximately $\mathcal{N}(\mu, \frac{\sigma^2}{n})$, when n is sufficiently large, denoted as $\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

We can approximate $g(\overline{X}_n)$ by Taylor's Formula. We have

$$g(\overline{X}_n) = g(\mu) + g'(\mu)(\overline{X}_n - \mu) + o(\overline{X}_n - \mu)$$

and thus

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) = g'(\mu)\sqrt{n}(\overline{X}_n - \mu) + \sqrt{n}o(\overline{X}_n - \mu)$$

with $o(\overline{X}_n - \mu) \to 0$, suggesting that

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

Theorem 1.6 (General Central Limit Theorem)**.** Suppose $X_1, X_2, \cdots$ are independent with $\mathbb{E}[X_i] = \mu_i$ and $\mathrm{Var}[X_i] = \sigma_i^2$ and let

$$S_n = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i\right),$$

then

$$S_n \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2\right)$$

provided that $\sum_{i=1}^{n} \sigma_i^2$ is not dominated by a small number of terms and the tails of the distributions of $\{X_i\}$ are not too dissimilar.

1.9.3 Quality of Normal Approximation

Definition 1.16. Define ***skewness*** of X_i as

$$\mathrm{Skew}(X_i) = \frac{\mathbb{E}[(X_i - \mu)^3]}{\sigma^3}.$$

Definition 1.17. Define ***kurtosis*** of X_i as

$$\mathrm{Kurt}(X_i) = \frac{\mathbb{E}[(X_i - \mu)^4]}{\sigma^4}.$$

Let $S_n = \sqrt{n}(\overline{X}_n - \mu)$, where $\overline{X}_n$ is the sample mean of independent $X_1, \cdots, X_n$ with CDF F with mean μ and variance σ^2. The normal approximation works better for fixed n if $\mathrm{Skew}(X_i)$ and $\mathrm{Kurt}(X_i)$ are close to the values for normal distribution (0 and 3, respectively).

1.9.4 Distribution Approximation

Theorem 1.7 (Slutsky's Theorem)**.** Suppose $X_n \xrightarrow{d} X \sim G$ and $Y_n \xrightarrow{p} \theta$, where θ is a constant, then as $n \to \infty, \psi(X_n, Y_n) \xrightarrow{d} \psi(X, \theta)$, where ψ is continuous.

Theorem 1.8 (Delta Method)**.** Suppose $a_n(X_n - \theta) \xrightarrow{d} Z$ where $a_n \uparrow \infty$. If $g(x)$ is differentiable at $x = \theta$, then

$$a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z.$$

Proof. By Taylor's Formula,

$$g(X_n) = g(\theta) + g'(\theta)(X_n - \theta) + \Delta(X_n)(X_n - \theta),$$

where $\Delta(X_n) \xrightarrow{p} 0$. Therefore,

$$a_n(g(X_n) - g(\theta)) = g'(\theta)a_n(X_n - \theta) + \Delta(X_n)a_n(X_n - \theta) \xrightarrow{d} g'(\theta)Z + 0 = g'(\theta)Z.$$

□

We can use the Delta Method to estimate standard errors of parameter estimators, and extend the Delta Method to functions of several sample means: $g(\overline{X}_n, \overline{Y}_n, \overline{Z}_n, \cdots)$.

Another application with the Delta Method is the variance stabilizing transformations for some distributions with $\mathrm{Var}[X_i] = \phi(\mu)$.

Example 1.3. For the Poisson distribution, $\mathrm{Var}[X_i] = \phi(\mu) = \mu$, then by CLT,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \mu).$$

We find g s.t.

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, 1),$$

i.e.,

$$[g'(\mu)]^2\mu = 1 \Rightarrow g'(\mu) = \frac{1}{\sqrt{\mu}}.$$

Thus, $g(x) = 2\sqrt{x} + C$ and

$$2\sqrt{\overline{X}_n} \dot{\sim} \mathcal{N}\left(2\sqrt{\mu}, \frac{1}{n}\right).$$

2 Statistical Models

2.1 Probability versus Statistics

Suppose $X_1, \cdots, X_n$ are independent r.v.s., with some CDF F. For probability, F is known and we can calculate probabilities involving the r.v.s. $X_1, \cdots X_n$. Knowledge of the population F gives information about the nature of samples from the population. For statistics, F is unknown and we observe outcomes of $X_1, \cdots X_n : x_1, \cdots, x_n$ (data).

Definition 2.1. ***Statistical inference*** uses the information in the data to estimate of infer properties of the unknown F.

Definition 2.2. Assume that the data $x_1, \cdots, x_n$ are outcomes of r.v.s. $X_1, \cdots, X_n$ whose joint distribution is F (which is assumed to be unknown to some degree). A ***statistical model*** is a family $\mathcal{F}$ of probability distributions of $(X_1, \cdots, X_n)$.

We assume that true distribution $F \in \mathcal{F}$ but in practice, $\mathcal{F}$ typically represents only an approximation to the truth, i.e., $F \notin \mathcal{F}$ but F is close to some $F_0 \in \mathcal{F}$.

Definition 2.3. $\mathcal{F}$ is called a ***parametric model*** if

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}^p,$$

where θ is the parameter and Θ is the parameter space. We can write $\theta = (\theta_1, \cdots, \theta_p)$.

Example 2.1. $X_1, \cdots, X_n$ are independent Poisson r.v.s. with unknown mean $\lambda > 0$.

Example 2.2. $X_1, \cdots, X_n$ are independent normal r.v.s. with unknown mean μ and variance σ^2.

Example 2.3. Observe $(x_1, Y_1), \cdots, (x_n, Y_n)$ with $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. This is a parametric model with parameters $(\beta_0, \beta_1, \sigma^2)$.

Definition 2.4. The model is said to be ***non-parametric*** if the parameter space Θ is not finite dimensional.

Example 2.4. $X_1, \cdots, X_n$ are independent continuous r.v.s. with unknown PDF $f(x)$.

Example 2.5. $Y_i = g(x_i) + \varepsilon_i$ for $i = 1, \cdots, n$ where g is unknown.

In practice, we often approximate the infinite dimensional parameter by a finite dimensional parameter. For example, we assume $g(x) \approx \sum_{k=1}^{p} \beta_k \phi_k(x)$ for some functions ϕ_k's and unknown parameters β_k's.

Definition 2.5. The model is said to be ***semi-parametric*** if non-parametric model has a finite dimensional parametric component.

Example 2.6. $X_1, \cdots, X_n$ are independent continuous r.v.s. with unknown PDF $f(x)$ on the interval $[0, \theta]$ where $\theta > 0$ is unknown.

Example 2.7. $Y_i = g(x_i) + \varepsilon_i$ for $i = 1, \cdots, n$ with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ where g and σ^2 are unknown.

Example 2.8. Observe $(x_1, Y_1), \cdots, (x_n, Y_n)$ with $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where we make no assumptions about the distribution of ε_i apart from 0 mean and finite variance. This is a semi-parametric model.

2.2 Bayesian Models

Assume we have a parametric model with parameter space $\Theta \subset \mathbb{R}^p$. For each $\theta \in \Theta$, the joint CDF F_θ is the conditional distribution of $(X_1, \cdots, X_n)$ given θ. ***Bayesian inference*** is the process that we put a probability distribution on Θ - ***prior distribution***, and then after observing $X_1 = x_1, \cdots, X_n = x_n$, we can use Bayes Theorem to obtain a ***posterior distribution*** of θ.

Note that we can take Bayesian inference for non-parametric models.