Probability and Statistics

1 Moment

1.1 Moment and CDF

For random variable X,

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)\,\mathrm{d}t - \int_{-\infty}^0 \mathbb{P}(X \leq t)\,\mathrm{d}t \tag{.1}$$

In general, $X \geq 0$ and a smooth function g with $g(0) = 0$

$$\mathbb{E}[g(X)] = \int_0^\infty g'(t)P(X > t)\,\mathrm{d}t \tag{.2}$$

1.2 k-th moment in the lens of CDF

$$\frac{1}{k}\,\mathbb{E}[X^k] = -\int_{-\infty}^0 x^{k-1}F(x)\,\mathrm{d}x + \int_0^\infty x^{k-1}(1 - F(x))\,\mathrm{d}x \tag{.3}$$

1.3 Clipped random variable

For complementary CDF $\bar{F}$ of random variable $X \geq 0$,

$$\mathbb{E}[\min(X, k)] = \int_0^k xf(x)\,\mathrm{d}x + k\bar{F}(k), \tag{.4}$$

2 Conditional Distribution

Tip: ML/Statistics 분야에서 흔히 쓰는 notation $p(x|y), p(x,y)$ 같은 것은 X, Y 들의 pdf/pmf 라고 생각하자. 거의 measure는 안나옴.

Conditioning on event. Let $f_{X,Y}$ be the joint density of X and Y, and $f_X(x)$ is the marginal density of X.

1. Single point conditioning $X = 1$

$$f_{Y|X=1}(y) = \frac{f_{X,Y}(1,y)}{f_X(1)}$$

2. Set conditioning $X \in S$

$$f_{Y|X\in S}(y) = \frac{\int_S f_{X,Y}(x,y)\,\mathrm{d}x}{\int_S f_X(x)\,\mathrm{d}x} = \frac{\int_S f_{Y|X=x}(y)f_X(x)\,\mathrm{d}x}{\int_S f_X(x)\,\mathrm{d}x}$$

3. Event conditioning

$$f_{Y|A}(y) = \frac{f_Y(y)1[y \in A]}{\int_A f_Y(y)\,\mathrm{d}y}$$

3 Weak convergence

3.1 Delta method

If there is a sequence of random variables X_n satisfying

$$\sqrt{n}[X_n - \theta] \xrightarrow{w} \mathcal{N}(0, \sigma^2) \tag{.5}$$

then

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{w} \mathcal{N}(0, \sigma^2 \cdot [g'(\theta)]^2) \tag{.6}$$

given that $g'(\theta)$ exists and is non-zero value.

3.2 Slutsky's theorem

If X_n converges in distribution to a random element X and Y_n converges in probability to a constant c, then

1. $X_n + Y_n \xrightarrow{w} X + c$
2. $X_n Y_n \xrightarrow{w} Xc$
3. $X_n / Y_n \xrightarrow{w} X/c$

where $\xrightarrow{w}$ denotes convergence in distribution.

4 Central limit theorem

4.1 Salem-Zygmund

Theorem. Let U be a uniform random variable with support $(0, 2\pi)$, and let $X_k = r_k \cos(n_k U + a_k)(0 \leq a_k < 2\pi)$, where

1. n_k satisfy the *lacunarity condition*: there exists $q > 1$ such that $n_{k+1} \geq q n_k$ for all k
2. $\sum_{i=1}^{\infty} r_i^2 = \infty$ and $\frac{r_k^2}{r_1^2 + \cdots + r_k^2} \to 0$

Then,

$$\frac{X_1 + \cdots + X_k}{\sqrt{r_1^2 + \cdots + r_k^2}} \tag{.7}$$

converges in distribution to $\mathcal{N}(0, 1/2)$.

5 Conditional independence

Theorem. Let p_{XYZ} be the joint PDF/PMF of X, Y and Z. Then the following are equivalent with up to almost-everywhere equivalence:

1. $X \perp Y \mid Z$
2. $p_{XYZ}(x, y, |z) = p_{X|Z}(x|z) p_{Y|Z}(y|z)$
3. $p_{X|YZ}(x|y, z) = p_{X|Z}(x|z)$
4. $p_{XYZ}(x, y, z) = \frac{p_{XZ}(x,z) p_{YZ}(y,z)}{p_Z(z)}$
5. $p_{XYZ}(x, y, z) = g(x, z) h(y, z)$ for some measurable functions g and h
6. $p_{X|YZ}(x|y, z) = w(x, z)$ for some measurable function w

Properties. Let X, Y, Z, W be RVs

1. (symmetry) $X \perp Y \mid Z \iff Y \perp X \mid Z$
2. (decomposition) $X \perp Y \mid Z \Rightarrow h(X) \perp Y \mid Z$ for any measurable function h
3. (weak union) $X \perp Y \mid Z \Rightarrow X \perp Y \mid Z, h(X)$ for any measurable function h
4. (contraction)
$$X \perp Y \mid Z \text{ and } X \perp W \mid (Y, Z) \iff X \perp (W, Y) \mid Z \tag{.8}$$
5. If the joint PDF $P_{XYZW}(x, y, z, w)$ satisfies $f_{YZW}(y, z, w) > 0$ almost everywhere. Then
$$X \perp Y \mid (W, Z) \text{ and } X \perp W \mid (Y, Z) \iff X \perp (W, Y) \mid Z \tag{.9}$$

5.1 Bayes' Theorem

Assume that X is a random variable on $(\Omega, \mathcal{F}, P)$,and let Q be another probability measure on $(\Omega, \mathcal{F})$ with Radon-Nikodym derivative
$$L = \frac{\mathrm{d}Q}{\mathrm{d}P} \text{ on } \mathcal{F} \tag{.10}$$
Assume that $X \in L^1(\Omega, \mathcal{F}, Q)$ and that $\mathcal{G}$ is a sigma-algebra with $\mathcal{G} \subseteq \mathcal{F}$. Then
$$\overset{Q}{\mathbb{E}}[X \mid \mathcal{G}] = \frac{\mathbb{E}^P[L \cdot X \mid \mathcal{G}]}{\mathbb{E}^P[L \mid \mathcal{G}]}, \quad Q - a.s. \tag{.11}$$

5.2 Conditional expectation under independence

Proposition. Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space, $(\mathbb{X}, \mathcal{M}), (\mathbb{Y}, \mathcal{N})$ be measurable spaces, $X : \Omega \to X$ and $Y : \Omega \to \mathbb{Y}$ be measurable functions. If X and Y are *independent* and $f \in (\mathcal{M} \otimes \mathcal{N})_b$ then
$$\mathbb{E}[f(X, Y) \mid X] = \mathbb{E}[f(x, Y)]|_{x=X} \text{ a.s.} \tag{.12}$$

6 Regular conditional distribution

Theorem. If X is a real random variable defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ then for every σ-algebra $\mathcal{G} \subset \mathcal{F}$ there is a regular conditional distribution for X given $\mathcal{G}$.

Regular conditional distributions are useful in part because they allow one to reduce many problems concerning conditional expectations to problems concerning only ordinary expectations. For such applications the following disintegration formula for conditional expectations is essential.

6.1 Disintegration formula

Theorem. Let $\mu_w(\,\mathrm{d}x)$ be a regular conditional distribution for X given $\mathcal{G}$, let Y be $\mathcal{G}$-measurable, and let $f(x, y)$ be a jointly measurable real-valued function such that $\mathbb{E}[|f(X, Y)|] < \infty$. Then,
$$\mathbb{E}[f(X, Y) \mid \mathcal{G}] = \int f(x, Y(w))\mu_w(\,\mathrm{d}x) \quad \text{ a.s.} \tag{.13}$$

Theorem2. Let Y and X be two Radon spaces. Let $\mu \in P(Y)$, let $\pi : Y \to X$ be a

Borel-measurable function, and let $\nu \in P(X)$ be the pushforward measure from Y to X by π. Then there exists a ν-almost everywhere uniquely determined family of probability measures $\{\mu_x\}_{x \in X} \subseteq P(Y)$ such that

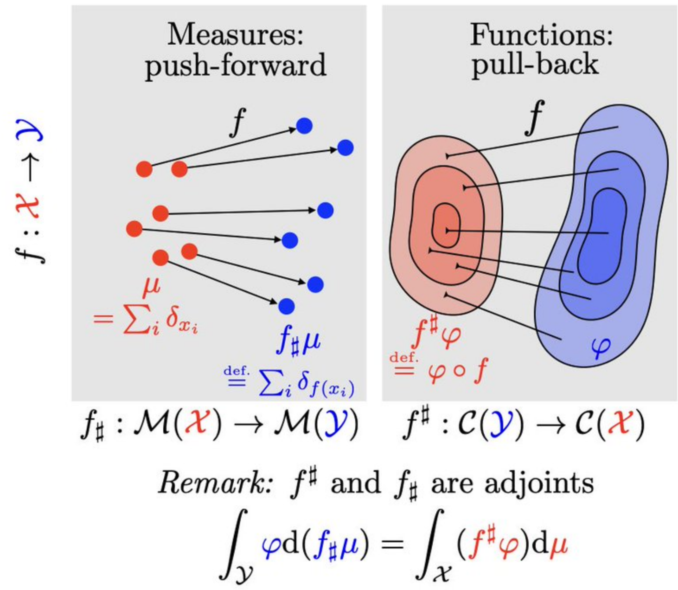

Figure 1: Pullback of functions and pushforward of measures are dual one with each other!

1. the function $x \mapsto \mu_x$ is Borel measurable
2. μ_x lives on the fiber $\pi^{-1}(x)$
3. for every Borel-measurable function $f : Y \to [0, +\infty]$,

$$\int_Y f(y)\, \mathrm{d}\mu(y) = \int_X \int_{\pi^{-1}(x)} f(y)\, \mathrm{d}\mu_x(y)\, \mathrm{d}\nu(x) \tag{.14}$$

7 Markov kernel

A Markov kernel (also called transition kernel, stochastic kernel, or probability kernel) is a mathematical formalization of a "function with random outcomes".

8 Mutual Information

8.1 Concavity of Mutual information

Let α be the law of X and π be the conditions law of $Y|X$. Let I_1 be $I(X;Y)$ where $(X,Y) \sim (\alpha_1, \pi)$, let I_2 be $I(X;Y)$ where $(X,Y) \sim (\alpha_2, \pi)$, let I be $I(X;Y)$ where $(X,Y) \sim (\lambda\alpha_1 + (1-\lambda)\alpha_2, \pi)$, for some $0 \leq \lambda \leq 1$, then

$$I \geq \lambda I_1 + (1-\lambda) I_2.$$

9 Fisher Information

Given the score function $\log p(\theta; X)$, the Fisher Information is defined as

$$I(\theta) := \mathop{\mathbb{E}}_{\theta}\left[-\frac{\partial^2}{\partial\theta^2} \log p(\theta; X)\right] \tag{.15}$$

It gives you uncertainty about the estimation since

$$\underbrace{\mathrm{Var}\left[\frac{\partial \ell(\theta; X)}{\partial \theta}\right]}_{\text{variance of score}} = -\,\mathbb{E}\left[\frac{\partial^2 \ell(\theta; X)}{\partial \theta^2}\right] \tag{.16}$$

holds.

9.1 Cramér–Rao bound

Unbiased means there

$$\mathbb{E}[\hat{\theta}(X) - \theta \mid \theta] = \int \left(\hat{\theta}(x) - \theta\right) f(x;\theta)\,\mathrm{d}x = 0 \text{ regardless of the value of } \theta \tag{.17}$$

Then, the following holds

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)} \tag{.18}$$

The precision to which we can estimate θ is fundamentally limited by the Fisher information of the likelihood function.

10 MLE estimation

10.1 Asymptotic normality

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathcal{N}(0, I^{-1}(\theta)) \tag{.19}$$

The mean square error (MSE) of $\hat{\theta}_n$ is

$$\mathrm{MSE}(\hat{\theta}_n, \theta_0) = \mathrm{bias}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta}_n) \approx \frac{1}{nI(\theta_0)} \tag{.20}$$

Moreover, if we know about $I(\theta_0)$, we can construct a $1-\alpha$ confidence interval using

$$\left[\hat{\theta}_n - \frac{z_{1-\alpha/2}}{\sqrt{n\hat{I}(\theta_0)}}, \hat{\theta}_n + \frac{z_{1-\alpha/2}}{\sqrt{n\hat{I}(\theta_0)}}\right] \tag{.21}$$

[1]

11 Famous Family

11.1 Poission

Binomial of Poisson trials is Poisson. Let $\lambda \geq 0, p \in [0,1]$. Suppose $(X_i)_{i=1}^{\infty}$ are i.i.d Bernoulli random variables with parameter p, and N is a Poisson(λ) random variable independent of the X_i's. Then $\sum_{i=1}^{N} X_i \sim \mathrm{Poi}(\lambda p)$.

Tail Distribution. Let $X \sim \mathrm{Poi}(\lambda)$, for some parameter $\lambda > 0$. Then for any $\boldsymbol{x} > 0$, we have

$$\mathbb{P}[X \geq \lambda + \boldsymbol{x}] \leq e^{-\frac{\boldsymbol{x}^2}{2\lambda} h\left(\frac{\boldsymbol{x}}{\lambda}\right)} \tag{.22}$$

, and, for any $0 < \boldsymbol{x} < \lambda$,

$$\mathbb{P}[X \leq \lambda - \boldsymbol{x}] \leq e^{-\frac{\boldsymbol{x}^2}{2\lambda} h\left(-\frac{\boldsymbol{x}}{\lambda}\right)}. \tag{.23}$$

where $h(u) := 2\frac{(1+u)\log(1+u)-u}{u^2}$. In particular, this implies that for every $\boldsymbol{x} > 0$,

$$\mathbb{P}[|X - \lambda| \geq \boldsymbol{x}] \leq 2e^{-\frac{\boldsymbol{x}^2}{2(\lambda+\boldsymbol{x})}}. \tag{.24}$$

[1] CI: estimator ± z-value * (SD of estimator)

11.2 Binomial

Fact. If X is Binomial(n, p), then $\mathbb{E}[1/(X+1)] \leq 1/((n+1) \cdot p)$

11.3 Negative Binomial

Negative binomial as a Poisson + logarithmic.

1. Draw T from a Poisson distribution and draw $K_1, K_2, \ldots, K_T$ independently from a logarithmic distribution.
2. Then $K = \sum_{t=1}^{T} K_t$ follows a negative binomial distribution.

12 M-estimator

12.1 Consistency

Theorem. Suppose that

1. $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \to 0$ in probability (ULLN)
2. For all $\varepsilon > 0$, $\sup \{M(\theta) : d(\theta, \theta_0) \geq \varepsilon\} < M(\theta_0)$ (identifiability)
3. $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$

Then $\hat{\theta}_n \to \theta_0$ in probability.