BUSI 573: Stochastic Models in Operations Management **Spring 2026**

Lecture 2: Stein's Method

Lecturer: Süleyman Kerimov *Date: January 20, 2026*

Disclaimer: *These notes are primarily adapted from expositional texts, including work by Nathan Ross and Remco van der Hofstad. These notes are not meant to be complete or fully rigorous; some proofs are not given, incomplete, or only outlined, as they are discussed in class.*

Stein's method is a powerful tool that helps to prove various central limit theorems and quantify the distance between two probability distributions. Recall the vanilla version of the central limit theorem.

Theorem 2.1 (Central Limit Theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed random variables. Let $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = Var(X_1) < \infty$. Let $S_n = \sum_{i=1}^n X_i$. Then we have*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to_{\mathrm{d}} \mathcal{N}(0,1),$$

where recall that $\mathcal{N}(0,1)$ is the standard normal distribution with density $f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$.

But, under stronger assumptions, one can be more explicit to quantify the error in the approximation, and be more precise regarding the rate of convergence.

Theorem 2.2 (Berry-Esseen Theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed random variables with $\mathbb{E}[X_1] = 0$ and $Var(X_1) = 1$, and assume that $\mathbb{E}[|X_1|^3] < \infty$. Let ϕ be the cumulative distribution function of $\mathcal{N}(0,1)$. Then we have*

$$\left|\mathbb{P}\Big(\frac{S_n}{\sqrt{n}} \leq x\Big) - \phi(x)\right| \leq 7.59\frac{\mathbb{E}[|X_1|^3]}{\sqrt{n}}.$$

This constant 7.59 was later improved by various papers. What central limit theorems suggest is that common events can be approximated by the normal distribution. But, in the context of the rare events, Poisson distribution provides a good approximation as well.

Theorem 2.3 (Poisson's Law of Small Numbers)**.** *Let $X \sim \mathrm{Bin}(n, \lambda/n)$, $\lambda > 0$. Then for any $k \in \mathbb{N}$, we have*

$$\mathbb{P}(X = k) \to e^{-\lambda}\frac{\lambda^k}{k!} = \mathbb{P}(\mathrm{Poi}(\lambda) = k),$$

as $n \to \infty$.

Proof of Theorem 2.3. We have

$$\begin{aligned}\mathbb{P}(X=k) &= \binom{n}{k}(\lambda/n)^k(1-\lambda/n)^{n-k}\\ &= \frac{n(n-1)\cdots(n-k+1)}{k!}(\lambda/n)^k(1-\lambda/n)^{n-k}\\ &= (1-\lambda/n)^n\cdot\frac{\lambda^k}{k!}\cdot\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}\cdot(1-\lambda/n)^{-k}.\end{aligned}$$

For fixed k, as $n\to\infty$,

$$\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}\to 1,\quad (1-\lambda/n)^k\to 1.$$

Now we use the fact that $\exp(-p/(1-p))\le 1-p\le\exp(-p)$ for all $p\in(0,1)$.

$$\exp\left(-\frac{\lambda/n}{1-\lambda/n}\right)\le 1-\frac{\lambda}{n}\le\exp(-\lambda/n),$$

so that

$$\exp\left(-\frac{\lambda}{1-\lambda/n}\right)\le\left(1-\frac{\lambda}{n}\right)^n\le\exp(-\lambda).$$

Therefore, we can conclude

$$\left(1-\frac{\lambda}{n}\right)^n\to\exp(-\lambda)\quad\text{as } n\to\infty.$$

■

Theorem 2.3 implies that $\mathrm{Bin}(n,\lambda/n)\to_{\mathrm{d}}\mathrm{Poi}(\lambda)$. What follows is a discussion around bounding the distance between two probability distributions (e.g., distance between $\mathrm{Bin}(n,\lambda/n)$ and $\mathrm{Poi}(\lambda)$). Before introducing a powerful tool called *coupling*, let us first formalize what we mean by distance.

Definition 2.4. *For two probability measures μ and ν, we define a probability metric as*

$$d_{\mathcal{H}}(\mu,\nu):=\sup_{h\in\mathcal{H}}\left|\int h(x)d\mu(x)-\int h(x)d\nu(x)\right|,$$

where $h(\cdot)$ is called a test function, and $\mathcal{H}$ is the family of test functions. Similarly, for two random variables W and Z, the probability metric has the form

$$d_{\mathcal{H}}(W,Z):=\sup_{h\in\mathcal{H}}\Big|\mathbb{E}[h(W)]-\mathbb{E}[h(Z)]\Big|$$

Here are some examples of probability metrics. Let $X\sim\mu$ and $Y\sim\nu$.

1. If $\mathcal{H}=\{\mathbb{1}_{\{\cdot\le x\}}:x\in\mathbb{R}\}$, then we get the Kolmogorov-Smirnov metric, which is denoted by d_K. Thus, $d_K(\mu,\nu)=\sup_{x\in\mathbb{R}}|F_\mu(x)-F_\nu(x)|=\sup_{x\in\mathbb{R}}|\mathbb{P}(X\le x)-\mathbb{P}(Y\le x)|$, and it can be interpreted as the maximum distance between distribution functions.

2. If $\mathcal{H} = \{\mathbb{1}_{\{\cdot \in A\}} : A \in \mathcal{B}(\mathbb{R})\}$, then we get the total variation metric, which is denoted by d_{TV}. Thus, $d_{TV}(\mu, \nu) = \sup_{A \in \mathcal{B}(\mathbb{R})} |\mu(A) - \nu(A)| = \sup_{A \in \mathcal{B}(\mathbb{R})} |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|$. The total variation metric us the main metric we use for approximation by discrete distributions.

It is immediate that for two random variables X and Y, $d_K(X,Y) \leq d_{TV}(X,Y)$. The following lemma gives a nice characterization of the total variance distance.

Discussion 2.5. *In the homework, you will show that if X and Y are two discrete random variables on Ω, then*

$$d_{TV}(X,Y) = \frac{1}{2} \sum_{w \in \Omega} |\mathbb{P}(X = \omega) - \mathbb{P}(Y = \omega)|.$$

Discussion 2.6. *Let F and G be the distribution functions with continuous densities f and g, respectively, i.e.,*

$$\mu(A) = \int_A f(x)\,dx, \qquad \nu(A) = \int_A g(x)\,dx, \tag{2.1}$$

for all measurable sets $A \subseteq \mathbb{R}$. Then we have

$$d_{\text{TV}}(f,g) = \frac{1}{2} \int_{-\infty}^{\infty} |f(x) - g(x)|\,dx. \tag{2.2}$$

2.1 Coupling

Definition 2.7 (Coupling of random variables)**.** *The random variables $(\hat{X}_1, \ldots, \hat{X}_n)$ are a coupling of the random variables $(X_1, \ldots, X_n)$ when $(\hat{X}_1, \ldots, \hat{X}_n)$ are defined on the same probability space, and are such that the marginal distribution of $\hat{X}_i$ is the same as that of X_i for all $i = 1, \ldots, n$, i.e., for all measurable subsets $\mathcal{E}$ of $\mathbb{R}$,*

$$\mathbb{P}(\hat{X}_i \in \mathcal{E}) = \mathbb{P}(X_i \in \mathcal{E}). \tag{2.3}$$

Note that the following is a trivial coupling: take $(\hat{X}_1, \ldots, \hat{X}_n)$ to be independent, with $\hat{X}_i$ having the same distribution as X_i. The following is another coupling: if $X, Y, U \sim U(0,1)$, then $(U, 1-U)$ is a coupling of (X,Y).

Now let X and Y be two discrete random variables with

$$\mathbb{P}(X = x) = p_x, \qquad \mathbb{P}(Y = y) = q_y, \qquad x \in \mathcal{X},\ y \in \mathcal{Y}.$$

The following result links the total variation distance between two discrete random variables and a coupling of them.

Theorem 2.8 (Maximal coupling)**.** *For any two discrete random variables X and Y, there exists a coupling $(\hat{X}, \hat{Y})$ of X and Y such that*

$$\mathbb{P}(\hat{X} \neq \hat{Y}) = d_{\text{TV}}(p,q), \tag{2.4}$$

while, for any coupling $(\hat{X}, \hat{Y})$ of X and Y,

$$\mathbb{P}(\hat{X} \neq \hat{Y}) \geq d_{\mathrm{TV}}(p, q). \tag{2.5}$$

Moreover, the maximal coupling $(\hat{X}, \hat{Y})$ satisfies the following:

$$\mathbb{P}(\hat{X} = \hat{Y} = x) = \min(p_x, q_x), \tag{2.6}$$

$$\mathbb{P}(\hat{X} = x, \, \hat{Y} = y) = \frac{\max(p_x - q_x, 0) \cdot \max(q_y - p_y, 0)}{d_{TV}(p, q)}, \qquad x \neq y. \tag{2.7}$$

Theorem 2.9 (Poisson limit for binomial random variables)**.** *Let $(I_i)_{i=1}^n$ be independent with $I_i \sim$ Bernoulli(p_i), and let $\lambda = \sum_{i=1}^n p_i$. Let $X = \sum_{i=1}^n I_i$, and let Y be a Poisson random variable with parameter λ. Then, there exists a coupling $(\hat{X}, \hat{Y})$ of X and Y such that*

$$\mathbb{P}(\hat{X} \neq \hat{Y}) \leq \sum_{i=1}^n p_i^2. \tag{2.8}$$

Proof. Let $J_i \sim \mathrm{Poi}(p_i)$ and assume that $(J_i)_{i=1}^n$ are independent. Note that the respective mass functions are

$$p_{i,x} = \mathbb{P}(I_i = x) = p_i^x (1 - p_i)^{1-x}, \qquad q_{i,x} = \mathbb{P}(J_i = x) = e^{-p_i} \frac{p_i^x}{x!} \tag{2.2.24}$$

Let $(\hat{I}_i, \hat{J}_i)$ be a coupling of I_i, J_i, where $(\hat{I}_i, \hat{J}_i)$ are independent for all i. Per Theorem 2.8, for each pair I_i, J_i, the maximal coupling $(\hat{I}_i, \hat{J}_i)$ satisfies

$$\mathbb{P}(\hat{I}_i = \hat{J}_i = x) = \min(p_{i,x}, q_{i,x}) = \begin{cases} 1 - p_i, & x = 0 \\ p_i e^{-p_i}, & x = 1 \\ 0, & x \geq 2 \end{cases} \tag{2.9}$$

since $1 - p_i \leq e^{-p_i}$ for $x = 0$. Since $1 - e^{-p_i} \leq p_i$, we have

$$\mathbb{P}(\hat{I}_i \neq \hat{J}_i) = 1 - \mathbb{P}(\hat{I}_i = \hat{J}_i) = 1 - (1 - p_i) - p_i e^{-p_i} = p_i(1 - e^{-p_i}) \leq p_i^2. \tag{2.10}$$

Next, let $\hat{X} = \sum_{i=1}^n \hat{I}_i$ and $\hat{Y} = \sum_{i=1}^n \hat{J}_i$. Then, $\hat{X}$ has the same distribution as $X = \sum_{i=1}^n I_i$, and $\hat{Y}$ has the same distribution as $Y = \sum_{i=1}^n J_i \sim \mathrm{Poi}(p_1 + \cdots + p_n)$. Per Boole's inequality[1] and (2.10), we have

$$\mathbb{P}(\hat{X} \neq \hat{Y}) \leq \mathbb{P}\left(\bigcup_{i=1}^n \{\hat{I}_i \neq \hat{J}_i\} \right) \leq \sum_{i=1}^n \mathbb{P}(\hat{I}_i \neq \hat{J}_i) \leq \sum_{i=1}^n p_i^2. \tag{2.11}$$

■

[1] Let $(A_i)_{i=1}^\infty$ be a sequence of events. Then, we have $\mathbb{P}(\cup_{i=1}^\infty A_i \leq \sum_{i=1}^\infty \mathbb{P}(A_i)$.

2.2 Stein-Chen Method

Now we discuss the Stein-Chen Method, which upper bounds the total variation metric between W and Z, where W is some random variable and Z is a Poisson random variable. That is, we want to show that

$$d_{\mathrm{TV}}(W, \mathrm{Poi}(\lambda)) := \sup_{A \subset \mathbb{Z}_{\geq 0}} |\mathbb{P}(W \in A) - \mathbb{P}(\mathrm{Poi}(\lambda) \in A)|$$

is small.

Proposition 2.10 (Characterizing operator of Poisson)**.** *For $\lambda > 0$, define the functional operator $\mathcal{A}$ by*

$$\mathcal{A}f(k) = \lambda f(k+1) - kf(k).$$

1. *If the random variable Z has the Poisson distribution with mean λ, then $\mathbb{E}\mathcal{A}f(Z) = 0$ for all bounded f.*
2. *If for some non-negative integer-valued random variable W, $\mathbb{E}\mathcal{A}f(W) = 0$ for all bounded functions f, then W has the Poisson distribution with mean λ.*

Proof of Proposition 2.10. We only prove the first part. Note that

$$\lambda \mathbb{E}[f(Z+1)] = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{k!} f(k+1) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k+1)!} (k+1) f(k+1) = \mathbb{E}[Zf(Z)].$$

■

Having Proposition 2.10, the following two results are very intuitive.

Proposition 2.11. *Let $\mathcal{P}_\lambda(A) := \mathbb{P}(\mathrm{Poi}(\lambda) \in A)$, $A \subseteq \mathbb{Z}_{\geq 0}$. The unique solution f_A of*

$$\lambda f_A(k+1) - kf_A(k) = \mathbf{1}[k \in A] - \mathcal{P}_\lambda(A) \tag{2.12}$$

with $f_A(0) = 0$ is given by

$$f_A(k) = \lambda^{-k} e^{\lambda} (k-1)!\, [\mathcal{P}_\lambda(A \cap U_k) - \mathcal{P}_\lambda(A)\mathcal{P}_\lambda(U_k)],$$

where $U_k = \{0, 1, \ldots, k-1\}$.

Exercise 2.12. *Prove Proposition 2.11.*

Thanks to Proposition 2.11, we have the following immediately.

Proposition 2.13. *If $W \geq 0$ is an integer-valued random variable with mean λ, then*

$$|\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A)| = |\mathbb{E}[\lambda f_A(W+1) - Wf_A(W)]|.$$

One last result is needed to present the main theorem. ■

Proposition 2.14. *If f_A solves* (2.12)*, then*

$$\|f_A\|\leq \min\{1, \lambda^{-1/2}\} \qquad \textit{and} \qquad \Delta f(k) := \|f(k+1)-f(k)\|\leq \frac{1-e^{-\lambda}}{\lambda} \leq \min\{1, \lambda^{-1}\},$$

where $||f||:=\sup_{x\in D}|f(x)|$ and D is the domain of f.

Theorem 2.15 (Poisson Approximation Theorem)**.** *Let $\mathcal{F}$ be the set of functions satisfying the conditions in Proposition 2.14. If $W \geq 0$ is an integer-valued random variable with mean λ and $Z \sim \mathrm{Po}(\lambda)$, then*

$$d_{\mathrm{TV}}(W, Z) \leq \sup_{f\in\mathcal{F}} |\mathbb{E}[\lambda f(W+1) - Wf(W)]|. \tag{4.4}$$

Let's now apply Theorem 2.15 to generalize Theorem 2.3: recall we have already shown that $X_n \sim \mathrm{Bin}(n, \lambda/n)$ and $Z \sim \mathrm{Poi}(\lambda)$ then $d_{\mathrm{TV}}(W_n, Z) \to 0$ as $n \to \infty$,

4.1 Law of small numbers

It is well known that if $X_n \sim \mathrm{Bin}(n, \lambda/n)$ and $Z \sim \mathrm{Poi}(\lambda)$ then $d_{\mathrm{TV}}(W_n, Z) \to 0$ as $n \to \infty$,

Theorem 2.16 (Theorem 4.6)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with $\mathbb{P}(X_i = 1) = p_i$, $W = \sum_{i=1}^n X_i$, and $\lambda = \mathbb{E}[W] = \sum_{i=1}^n p_i$. If $Z \sim \mathrm{Poi}(\lambda)$, then*

$$d_{\mathrm{TV}}(W, Z) \leq \min\{1, \lambda^{-1}\} \sum_{i=1}^n p_i^2$$

Proof. The second inequality is clear and is only included to address the discussion preceding the theorem. For the first inequality, we apply Theorem 4.5. Let f satisfy (4.2) and note that

$$\begin{aligned}\mathbb{E}[Wf(W)] &= \sum_{i=1}^n \mathbb{E}[X_i f(W)]\\ &= \sum_{i=1}^n \mathbb{E}[f(W) \mid X_i = 1]\mathbb{P}(X_i = 1)\\ &= \sum_{i=1}^n p_i \mathbb{E}[f(W_i + 1)], \end{aligned} \tag{4.5}$$

where $W_i = W - X_i$ and (4.5) follows since X_i is independent of W_i. Since $\lambda f(W+1) = \sum_i p_i f(W+1)$, we obtain

$$|\mathbb{E}[\lambda f(W+1) - Wf(W)]| = \left|\sum_{i=1}^n p_i\, \mathbb{E}[f(W+1) - f(W_i+1)]\right| \leq \sum_{i=1}^n p_i \|\Delta f\|\, \mathbb{E}[|W - W_i|].$$

To see why the inequality holds, note that $f(W+1) - f(W_i+1) = \sum_{k=W_i+1}^W f(k+1) - f(k)$ so that by the triangle inequality $|f(W+1) - f(W_i+1)|\leq \sum_{k=W_i+1}^W ||\Delta f||= ||\Delta f|||W - W_i|$, and we just take the expectation.

Since $|W - W_i|= X_i$, we get

$$|\mathbb{E}[\lambda f(W+1) - Wf(W)]| \leq \|\Delta f\| \sum_{i=1}^{n} p_i \mathbb{E}[X_i] = \|\Delta f\| \sum_{i=1}^{n} p_i^2.$$

Using $\|\Delta f\|\leq \min\{1, \lambda^{-1}\}$ from (4.2), we conclude that

$$d_{\mathrm{TV}}(W, Z) \leq \min\{1, \lambda^{-1}\} \sum_{i=1}^{n} p_i^2.$$

By Theorem 2.15, we are done. ■