Charlas Sobre el Método de Stein
Larry Goldstein, USC

El método de Stein es una manera de probar cotas para aproximaciones por distribuciones, como el Teorema del límite central. Además, podemos aplicar algunas técnicas desarrollado el método para mostrar concentración de medida de variables aleatorias.

Estas charlas están organizadas asi:

1. Introduciíon, las ideas bsicas, y la distribución Gaussiana
2. La distribución Poisson, y aplicaciones.
3. Desigualdades de concentración de medida
4. Otras distribuciones

Antes de empezar, quiero decir que hay mucho más de lo que voy a presentar. Además, no hay tal cosa como 'el' método de Stein, seria mejor decir 'los métodos de Stein' porque toma muchas formas. A primera vista, tal vez pareciera que 'el método' solo consiste en una bolsa de trucos vagamente interrelacionados. Hay mucho de cierto en este pensamiento, y yo también no se si es un método, pero sin embargo espero de convencerlos que se trata de una idea muy útil. Las referencias generales son [10], [21] y [6], y además los artículos originales de Stein, [22] y [23].

1 Introducción, el Caso Gaussiana, y Las Ideas Basicas

Empezamos con la Teorema de limite central. El ámbito para el teorema del límite central es el siguiente. Dadas variables aleatorias $X_1, \ldots, X_n$, independientes y con la mismas distibuciones, con media cero y varianza uno,

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \to_d \mathcal{N}(0,1),$$

donde la distribución $Z \sim \mathcal{N}(0,1)$ tiene la densidad

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2),$$

y decimos $Y \sim \mathcal{N}(\mu, \sigma^2)$ cuando la distribución de Y es igual a $\sigma Z + \mu$.

Clasicamente, el TLC es probado usando funciones características, con la forma $\phi(t) = E[e^{itX}]$, poniendonos en la tierra de los números imaginarios, y consiguientemente, perdiendo parte de nuestra intuición.

El método de Stein nos ofrece una alternativa, usando en lugar de $\phi(t)$, una ecuación que caracteriza la distribución. En esta caso, lo que se necesita es provisto por el Lema de Stein, que dice

$$X \sim \mathcal{N}(0, \sigma^2) \quad \text{si y solo si} \quad E[Xf(X)] = \sigma^2 E[f'(X)]$$

para todas las f's tal que éstas expresiones existan.

A primera vista, esta identidad parece inocente y tal vez muy poquito interesante, quizás solo una curiosidad. La idea enorme, es usarla como el lado derecho en una ecuación diferencial. Pero primero, explico la forma de algunas métricas, o distancias.

Sea $\mathcal{H}$ una colección de funciónes $\mathbb{R} \to \mathbb{R}$, dadas X y Y variables aleatorias, podemos formar

$$d_{\mathcal{H}}(X, Y) = \sup_{h \in \mathcal{H}} |Ef(X) - Ef(Y)|.$$

Muchas distancias bien conocidas son de esta forma. Per ejemplo, eligiendo

$$\mathcal{H} = \{h : h(x) = \mathbf{1}(x \leq a), a \in \mathbb{R}\}$$

produce la distancia 'de Kolmogorov' o la distancia L^∞, que tambien podemos escribir como

$$d_\infty(X, Y) = \sup_{a \in \mathbb{R}} |P(X \leq a) - P(Y \leq a)|.$$

O podemos elegir la clase

$$\mathcal{H} = \mathrm{Lip}_1 \quad \text{donde} \quad \mathrm{Lip}_1 = \{h : |h(x) - h(y)| \leq |x - y|\},$$

que resulta en la distancia 'Wasserstein' o L^1, que también satisface

$$d_1(X, Y) = \int_{-\infty}^{\infty} |P(X \leq t) - P(Y \leq t)| dt \quad \text{y} \quad d_1(X, Y) = \inf E|X - Y|$$

donde el ínfimo es sobre todos acoplamientos (X, Y) con las distribuciones marginales dadas.

La elección de las funciones f medibles

$$\mathcal{H} = \{h : \mathbb{R} \to [0, 1]\}$$

da la distancia de variación total

$$d_{\mathrm{TV}}(X, Y) = \sup_{A \in \mathcal{B}} |P(X \in A) - P(Y \in A)|.$$

No es que todas las distancias se pueden escribir en esta manera, un ejemplo contrario es la distancia di Lévy-Prohorov, en el espacio $(\Omega, \mathcal{F})$,

$$d_{\mathrm{LP}}(P, Q) = \inf\{\epsilon > 0 : P(A) \leq Q(A) + \epsilon, Q(A) \leq P(A) + \epsilon, \forall A \in \mathcal{F}\}$$

Hasta ahora, nadie sabe como hacer que el metódo funcione para estas distancias.

De todos modos, sea que queremos calcular uno de estas tipos de distancas entre la distribución de X y la gaussiana $Z \sim \mathcal{N}(0, 1)$. La forma de la distancia da el lado derecho de

$$f'(x) - xf(x) = h(x) - Eh(Z)$$

y el lado izquierda viene de la lema de Stein. Era un idea profundo arreglar estas dos cosas asi.

Entonces, podemos sustituir la variable aleatoria X por x, y toma la media, obteniendo lo que necesitamos en el lado derecho, $E[h(X)] - Eh(Z)$, per uso en la forma de una distancia.

A primera vista, parece que hacemos el problema mas difícil, porque ahora tenemos resolver una ecuación diferencial, y ademas, como voy a mostrarles en un momento, tenemos ademas encontrar cotas sobre la solución de esta ecuación, y sus derivadas. Pero, en esta punto, solo observamos que hay solo una variable, X, en la ecuación, en contraste con la forma de las distancias, que tienen X y Z.

Para mostrarles un ejemplo que demuestra la utilidad escribiendo las formulas asi, voy a introducir una operación sobre distribuciones que se llama la 'sesgo cero' transformación.

Ahora, presento una idea similar, basado su el lema de Stein. Sabemos ahora que se

$$E[Xf(X)] = E[f'(X)] \quad \text{para cada } f \text{ tal que estas expresiones existen}$$

si y solo si $X \sim \mathcal{N}(0,1)$. Entonces, se X no es normal, no satisface esta identidad. Pero, tal vez satisface una identidad similar; vean [17].

Teorema 1.1. *Sea X una variable aleatoria con media cero y varianza $\sigma^2 \in (0,\infty)$. Entonces, exista una distribución única por una variable, X^*, que satisface*

$$E[Xf(X)] = \sigma^2 E[f'(X^*)] \quad \textit{para cada } f \textit{ tal que estas expresiones existen.}$$

Se llama la transformación de sesgo cero que envía $X \to X^*$. Podemos ahora re interpretare la lema de Stein come: La distribición normal es la único punto fijo de la tranformación de sesgo cero.

Para ver un ejemplo, sea $Y \in \{-1,1\}$ uniformemente. Entonces

$$E[Yf(Y)] = \frac{1}{2}[f(1) - f(-1)] = \frac{1}{2}\int_{-1}^{1} f'(u)du = E[f'(U)] \quad \text{donde} \quad U \sim \mathcal{U}[-1,-1],$$

mostrando $Y^* \sim \mathcal{U}[-1,1]$.

Sorprendentemente, la regla para formar X^* de la suma

$$X = \sum_{j=1}^{n} X_i$$

de variables independiente con media cero es casi igual de la regla que vimos con sesgo de tamaño. Pero aquí, reemplaza una variable elegido con probabilidad proporcional a su varianza,

$$P(I = i) = \frac{\sigma_i^2}{\sum_{j=1}^{n} \sigma_j^2}$$

con X_i^* independiente de $X_j, j \neq i$, y forma

$$X^* = \sum_{j \neq I} X_j + X_I^*,$$

que tiene la distribución de X-sesgo cero. Ves que aquí tenemos X y X^* en el mismo espacio.

Consideramos otra vez las distancias, y toma la class de funciones igual a la clase Lip_1, que resulta de la distancia L^1 o Wasserstein. Más, imaginamos que tenemos W y W^* sobre la misma espacio, quiero decir, un acoplamiento, (W, W^*). Entra la ecuación de Stein. Dado $h \in \text{Lip}_1$, hay una solución f.

Sea la varianza de W es igual a 1. Entonces, usando el teorema de los valores intermedios

$$|E[h(W) - Eh(Z)]| = |E[f'(W) - Wf(W)] = |E[f'(W) - f'(W^*)]| \leq \|f''\| E|W^* - W|.$$

Podemos mostrar, solo usando calcolo basico, que $\|f''\| \leq 2$ siempre, cuando $h \in \text{Lip}_1$, ejercicio, la única solución f acotada de

$$f'(w) - wf(w) = h(w) - Eh(Z) \quad \text{donde} \quad Z \sim \mathcal{N}(0,1)$$

satisface $\|f''\| \leq 2$, donde $\|g\|_\infty = \sup_{x \in \mathbb{R}} |g(x)|$. Por lo tanto, para cada $h \in \text{Lip}_1$,

$$|E[h(W) - Eh(Z)]| \leq 2E|W^* - W| \quad \text{o} \quad \sup_{h \in \text{Lip}_1} |E[h(W) - Eh(Z)]| \leq 2E|W^* - W|.$$

Pero, el lado izquierdo es exactamente la definición de la distancia Wasserstein, y por eso

$$d_1(W, Z) \leq 2E|W^* - W|.$$

Pero, esta desigualdad es verdad para todas las manieras en que podemos construir W^* y W en el mismo espacio, o, en otras palabras, para todo las distribuciones conjuntos che tiene las marginales como dadas. Por eso, podemos tomar el ínfimo sobre el lado derecho, y usando la tercera caracterización de la distancia Wasserstin, hemos logrado el teorema [14] que dice

$$d_1(W, Z) \leq 2d_1(W, W^*). \tag{1}$$

Con estas herramientas, ahora podemos escribir una prova sencilla y facil de la TLC, con una cota sobre el error, para la suma W, rescalado por σ,

$$W = \frac{1}{\sigma} \sum_{i=1}^{n} X_i$$

de variables independientes, con media cero, y varianzas $\sigma_1^2, \ldots, \sigma_n^2$, con

$$\sigma^2 = \sigma_1^2 + \cdots + \sigma_n^2.$$

Podemos construir una variable X^*, con la distribución X-sesgo de cero, y reemplazar una variable en la suma, escogida en proporción de su varianza, con una otra variable, independiente de las otras, con una variable teniendo la distribución de sesgo de cero de la variable que estamos reemplazando, o sea

$$W^* = W - X_I/\sigma + X_I^*/\sigma.$$

Bajo independencia, basta reemplazar solo una. Es fàcil verificar que $(aX)^* = aX^*$ por cada $a \neq 0$. Ahora podemos tomar el medio sobre el índice I, y obtener

$$E|W^* - W| = \frac{1}{\sigma} E|X_I^* - X_I| = \frac{1}{\sigma} \sum_{i=1}^{n} P(I = i) E|X_i^* - X_i| = \frac{1}{\sigma} \sum_{i=1}^{n} \frac{\sigma_i^2}{\sigma^2} E|X_i^* - X_i|.$$

Un poquito descuidadamente, podemos usar la desigualdad triangular, y obtener

$$E|X_i^* - X_i| \leq E|X_i^*| + E|X_i|. \tag{2}$$

Para encontrar una cota sobre el primer término, nos acordamos la definición de la distribución sesgo de tamaño, para una X con media cero

$$E[Xf(X)] = \sigma^2 E[f'(X^*)]$$

y sustituir $f(x) = x^2/2\mathbf{1}(x \geq 0) - x^2/2\mathbf{1}(x < 0)$, dandonos $f'(x) = |x|$, y entonces

$$\frac{1}{2} E[|X^3|] = \sigma^2 E[|X^*|] \quad \text{o, que} \quad E[|X^*|] = \frac{1}{2\sigma^2} E[|X^3|] \quad \text{y} \quad E[|X^*|]| \leq \frac{1}{2\sigma^2} E[|X|^3]$$

Además, a causa que $E|X/\sigma|^2 = 1$, usando la inegualidad de Lyapunov,

$$E|X/\sigma| \leq \sqrt{E|X^2/\sigma^2|} = 1 \leq (E|X|^3/\sigma^3)^{1/3} \leq E|X|^3/\sigma^3,$$

y, desenredando,

$$E|X| \leq E|X|^3/\sigma^2$$

Desde (2), ahora,

$$E|X_i^* - X_i| \leq \frac{3}{2} E|X_i|^3/\sigma_i^2$$

y

$$E|W^* - W| = \frac{1}{\sigma} \sum_{i=1}^{n} \frac{\sigma_i^2}{\sigma^2} E|X_i^* - X_i| \leq \frac{3}{2\sigma^3} \sum_{i=1} E|X_i|^3.$$

Vemos ahora el teorema de Berry Esseen, en la distancia Wasserstein, usando el teorema 1

$$d_1(W, Z) \leq 2d_1(W^*, W) \leq \frac{3}{\sigma^3} \sum_{i=1} E|X_i|^3$$

Cuando las variables $X_1, \ldots, X_n$ tienen la misma distribución, sea $\tau^2 = \mathrm{Var}(X_i)$, tenemos $\sigma^2 = n\tau^2$, y

$$d_1(W, Z) \leq \frac{3}{\sigma^3} \sum_{i=1} E|X_i|^3 = \frac{3}{n^{3/2}\tau^3} nE|X|^3 = \frac{3E|X|^3}{\sqrt{n}\tau^3}.$$

Esta forma del teorema es probablemente más reconocible, si lo han visto antes.

Pero el poder del método es más evidente en ejemplos que contienen dependencia. Por ejemplo, sea $A = (a_{ij})$ es una matriz, aleatoria o no, y π una permutación, decimos elegido uniformemente sobre toda las $n!$ permutaciónes, y formamos

$$W = \sum_{i=1}^{n} a_{i\pi(i)} \quad \text{y} \quad \frac{W_n - E[W_n]}{\sqrt{\text{Var}(W_n)}} \to Z.$$

En estadística, esta forma ocurre en algunas aplicaciones. Por ejemplo, en estadística no paramétrica, decimos que tenemos las parejas $(X_1, Y_1), \ldots, (X_n, Y_n)$. Tal vez pensamos que X y Y son relacionados, por ejemplo, la altura de un padre y su hijo. Si son relacionados, X_i y Y_i son vecino, y $a_{i,i}$ donde $a_{i,j} = |X_i - Y_j|$, es 'pequeña'. Para hacer una prueba de permutación, y ver si es la verdad que los valores $a_{i,i}$ son pequeñas, comparamos la suma

$$T = \sum_{i=1}^{n} a_{i,i}$$

a el valor de W, en que la altura de un padre es comparado a un hijo de otro padre, un hijo elegido uniformemente. En esta caso, también podemos formar un acoplamiento de sesgo cero, a demonstrar una cota similar a la cota por las variables independientes, vean [10].

2 El Caso de Poisson, y Applicaciónes

Recordamos que la distribución Poisson con media $\lambda > 0$, escrito como $\mathcal{P}(\lambda)$ esta definida por

$$P(Z = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

Podemos obtener la Poisson como un limite de Binomial $\text{Bin}(n, p)$, que es la distribución de la suma

$$X = X_1 + \ldots + X_n$$

donde cada $X_i, 1 = 1, \ldots, n$ tiene la distribución $P(X_i = 1) = 1 - P(X_i = 0) = p$, una variable 'Bernoulli'. La variable X cuenta el número de caras en n lanzamientos de una moneda. Y es facile mostrar que

$$\text{Bin}(n, \lambda/n) \to \mathcal{P}(\lambda),$$

a veces esta distribución esta llamada 'la ley del numeros pequeños,' o la ley de eventos raros. Hoy, vamos a usar el método de Stein para calcular una cota sobre el error en esta aproximación. Sencillamente, la Poisson ocurre, tal vez aproximadamente, cuando el numero de experimentos n es grande, pero la probabilidad de un éxito en cada experimento es pequeña.

Un ejemplo muy dramático viene de la bombardeo de Londres durante la segunda guerra mundial. Algunos de los ciudadanos pensaban que habia un orden en los lugares,

y encontrándolo fue una maniera para estar mas seguro, y vivo. Para examinar esta hipótesis, la ciudad era dividido en 576 cuadros, cada uno cuarto kilómetro por un cuatro kilómetro. 229 de estés cuadros no fueron bombardeado, 211 fueron bombardeado 'solo' una vez, etc, aquí esta una tabla con los datos

0	229	227
1	211	211.39
2	93	98.54
3	35	30.62
4	7	7.14
5+	1	1.57

Es fácil a ver que la aproximación es muy buena, y significa que los lugares del bombardeo fueron al azar.

Para $X \geq 0$ no trivial, y con media μ, introducimos la distriubución sesgo de tamaño de X, escrito X^s, o mejor, la transfomación de sesgo de tamaño $X \to X^s$, che define X^s por la caracterización

$$E[Xf(X)] = \lambda E[f(X^s)] \quad \text{donde } \lambda = E[X].$$

Empezamos con el ejemplo, el mas sencillo, el caso quando $X \in \{0, 1\}$, una variable 'Bernoulli'. Decimos que $P(X = 1) = p$, con $p \in (0, 1)$ para no ser trivial. Entonces

$$\begin{aligned} E[Xf(X)] &= pE[Xf(X)|X = 1] + (1 - p)E[Xf(X)|X = 0] \\ &= pf(1) = E[X]f(1) = E[X]E[f(X^s)] \quad \text{por } X^s = 1 \text{ con probabilidad 1.} \end{aligned}$$

Entonces, $X^s = 1$ en este caso.

Existe en general, pero, por ejemplo, se X tiene una densidad $p(x)$, la densidad de X^s es $xp(x)/\mu$, y se X es discreto,

$$P(X^s = x) = \frac{xP(X = x)}{\mu}.$$

En efecto, X^s escoge las valores de la variable X en proporción a el tamaño. Por un ejemplo, se hay reservas de petróleo baja la tierra, en lugares en donde no podemos ver, y exploramos sin idea donde las están, encontraremos las mas grandes con alta probabilidad. O, si marcamos un numero teléfono, aleatoriamente, habrá dos veces mas probable que contactamos una persona que tiene dos lineas telefónicas que una persona que tiene solo una. La transformación de sesgo de tamaño envia $X \to X^s$, mas precisamente, es una transformación entre distribuciones.

Ahora, sean variables aleatorias $X_1, \ldots, X_n$, no negativo y con media en $(0, \infty)$, decimos $\mu_1, \ldots, \mu_n$. Primero, suponemos que las variables son independientes, y sea W la suma,

$$W = \sum_{i=1}^{n} X_i$$

La pregunta es: como construir W^s? Esta pregunta hay una respuesta sorprendente. Elige una variable X_I, con probabilidad proporcional a su media, μ_i, así

$$P(I = i) = \frac{\mu_i}{\sum_{j=1}^{n} \mu_j},$$

con la índice I independiente de $X_1, \ldots, X_n$. Luego, reemplazar X_I con X_I^s, teniendo la distribución X con sesgo de tamaño, independiente de $X_j, j \neq I$. Ahora, la sum

$$W^s = \sum_{i \neq I} X_i + X_I^s$$

tiene la distribución de X^s. Tenemos,

$$W^s = W - X_I + X_I^s.$$

El caso $\mathrm{Bin}(n,p)$ es el caso especial quando $X_1, \ldots, X_n$ son Bernolli, y obtenemos

$$W^s = W - X_I + 1 \quad \text{o} \quad \mathrm{Bin}(n,p)^s = \mathrm{Bin}(n-1,p) + 1.$$

Entonces, la distribución sesgo de tamaño de $\mathrm{Bin}(n,p)$ es cerca a la distributución $\mathrm{Bin}(n,p)+1$, y entonces cerca de la Poisson. Tomando limites in n nos da que $Z_\lambda^s =_d Z_\lambda + 1$ cuando $Z_\lambda \sim \mathcal{P}(\lambda)$.

Como el Gaussiano, este argumento nos da un characterización del Poisson también, similar, en términos de funciones, en particular, $X \sim \mathcal{P}(\lambda)$ si y solo si

$$E[Xf(X)] = \lambda E[f(X+1)] \quad \text{para cada } f \text{ tal que estas expresiones existen.}$$

Y ademas el Poisson es la única distribución con esta propriedad. Podemos declarar esta hecho usando el lenguaje de transformaciones. Dada X, no trival y no negativo con media μ, la distribución Poisson es la única punto fijo de la trasformacin

$$X \to X^s - 1.$$

Ahora, usamos el método de Stein para cuantificar esta observación.b

Nos formamos la ecuación de Stein para la Poisson usando la expresión encima como il lado izquierdo,

$$\lambda f(w+1) - wf(w) = h(w) - \mathcal{P}_\lambda h \quad \text{donde } \mathcal{P}_\lambda h = h(Z_\lambda), Z_\lambda \sim \mathcal{P}_\lambda$$

Supongamos que, dada una variable $W \in \{0,1,2,\ldots\}$, podemos construir una otra variable W^s con la distribución la distribución W-sesgo de tamaño. (Es siempre posible, porque sencillamente podemos construir W^s independiente de W, pero esta construción sera inútil.) De todos modos, vedendo el lado izquierdo de la ecuación de Stein, tenemos

$$E[\lambda f(W+1) - Wf(W)] \;=\; E[\lambda f(W+1)] - \lambda f(W^s)] \;=\; \lambda E[f(W+1) - f(W^s)],$$

y definiendo

$$\Delta f = \sup_{x \in \{0,1,\ldots\}} |f(x+1) - f(x)| \quad \text{tenemos} \quad |f(b) - f(a)| \leq \Delta f \, |b-a|$$

y entonces

$$|Eh(W) - Eh(Z_\lambda)| = |E[\lambda f(W+1) - Wf(W)]| \leq \lambda \Delta f E|W + 1 - W^s|.$$

Con calculo, podemos probar que cuando f y h están relacionado por la ecuación de Stein,

$$\Delta f \leq \frac{1 - e^{-\lambda}}{\lambda} \quad \text{para toda} \quad h(x) = 1_A, A \subset \{0,1,2,\ldots\},$$

y por lo tanto

$$|Eh(W) - Eh(Z_\lambda)| \leq (1 - e^{-\lambda})E|W + 1 - W^s|,$$

y tomando el supremo sobre el lado izquierdo danos la distancia variación total, y el teorema

$$d_{TV}(W, Z_\lambda) \leq (1 - e^{-\lambda})E|W - (W^s - 1)|.$$

Se da cuenta de que no hay restriciónes sobre la forma de W, en particular, W puede ser una suma de variables dependente.

Usando $f(x) = x$ en

$$E[Wf(W)] = \lambda E[f(W^s)] \quad \text{da} \quad E[W^2] = \lambda E[W^s].$$

En el caso especial cuando $W^s - 1 \leq W$, podemos quitar el valore absoluto y obtener

$$\begin{aligned} E|W - (W^s - 1)| = E[W - (W^s - 1)] = \lambda - E[W^s] + 1 = \lambda - \frac{E[W^2]}{\lambda} + 1 \\ = 1 - \left(\frac{E[W^2]}{\lambda} - \lambda\right) = 1 - \frac{(E[W^2] - \lambda^2)}{\lambda} = 1 - \frac{\text{Var(W)}}{E[W]}, \end{aligned}$$

y

$$d_{TV}(W, Z_\lambda) \leq (1 - e^{-\lambda})E|W + 1 - W^s| = (1 - e^{-\lambda})\left(1 - \frac{\text{Var(W)}}{E[W]}\right).$$

Recordamos que la Poisson Z_λ satisface $\text{Var}(Z_\lambda) = E[Z_\lambda]$ y en este caso

En el caso clásico, la suma de Bernoullis independientes,

$$W^s = \sum_{i \neq I} X_i + 1 \quad \text{y en particular} \quad W^s - 1 = \sum_{i \neq I} X_i \leq W$$

y

$$\text{Var}(W) = \sum_{i=1}^{n} p_i(1 - p_i) \quad \text{and} \quad \lambda = \sum_{i=1}^{n} p_i$$

y

$$1 - \frac{\text{Var(W)}}{E[W]} = \frac{E[W] - \text{Var(W)}}{E[W]} = \frac{\sum_{i=1}^{n}(p_i - p_i(1 - p_i))}{E[W]} = \frac{\sum_{i=1}^{n} p_i^2}{\lambda}.$$

Cuando $p_i = \lambda/n$, la sum arriba es $\sum_{i=1}^{n} p_i^2 = n(\lambda^2/n^2) = \lambda^2/n$, y obtenemos convergencia a la Poisson, con la cota $(1 - e^{-\lambda})\lambda/n$. Haciendolo directamente, deberíamos mostrar una cota sobre

$$\sum_{k=1}^{n} |\binom{n}{k} p^k (1 - p)^{n-k} - e^{-\lambda}\lambda^k/k!| + \sum_{k \geq n+1} e^{-\lambda}\lambda^k/k!$$

Podemos usar estas formulas ademas en situaciónes con dependencia. Para esta proposito, debemos considerar como podemos constuir un variable W^s para una, por

ejemplo, $W = X_1 + \cdots + X_n$, una suma de variables Bernoulli, dependiente. Empezamos en la misma maniera, escogendo una variable en proporción de su media, que es en este caso, $E[X_i] = p_i$, la probabilidad que $X_i = 1$. Como la ultima vez, reemplazar X_i con una variable con la distribución X_i^s, o en este caso, 1.

Despues i esta escogido, para las otras variables, necesitamos constuirlas en la maniera en que la distribución conjunta satisface

$$P(X_1^{(i)} = x_1, \ldots, X_n^{(i)} = x_n) = P(X_1 = x_1, \ldots, X_n = x_n | X_i = 1).$$

Cuando las variables son independientes, no hay la necesidad cambiarlas, porque $P(X_1 = x_1, \ldots, X_n = x_n | X_i = 1) = P(X_1 = x_1, \ldots, X_n = x_n)$.

Un problema 'moderno' en que estas técnicas son útiles es en el campo de biología molecular. Supongamos que tenemos dos secuencias de DNA, compuestas de las letras A, T, G, C, de dos animales, decimos, un caballo e una vaca. La secuencia del caballo tiene n letras $A_1 \ldots A_n$, y de la vaca m, con $B_1 \ldots B_m$. Pero, ambos tiene una misma sub secuencia de k letras. Para un biólogo, tal vez la presencia de esta sub secuencia significa que es importante, que es un código para un gen. Pero, se la probabilidad para tener una emparejamiento así es 'grande', no significa nada. Podemos usar la aproximación Poisson. Define

$$W = \sum_{i=1}^{n-k+1} \sum_{j=1}^{m-k+1} X_{ij}$$

donde

$$X_{ij} = \mathbf{1}(A_i A_{i+1} \cdots A_{i+k-1} = B_j B_{j+1} \cdots B_{j+k-1}),$$

con $\mathbf{1}(A)$ la función característica de un evento A que toma el valor 1 cuando A es verdad, y 0 si no. Tomando la esperanza,

$$\begin{aligned} \lambda = E[W] &= (n-k+1)(m-k+1)P(A_i A_{i+1} \cdots A_{i+k-1} = B_j B_{j+1} \cdots B_{j+k-1}) \\ &= (n-k+1)(m-k+1)4^{-k}. \end{aligned}$$

Y W es aproximadamente $Z_\lambda \sim \mathcal{P}(\lambda)$, y la probabilidad que no hay una sub secuencia común de largo k es $P(W = 0) \approx P(Z_\lambda = 0) = e^{-\lambda}$. Podemos calcular un cota, vean [3], [4], pero vemos ahora un ejemplo mas sencillo.

El problema de cumpleaños. Probablemente, todos saben el problema de cumpleaños, cuantas personas necesitamos juntos para hacer una apuesta que hay dos che tienen el mismo cumpleaños. Porque el evento que yo y te tenemos el mismo cumpleaños es pequeña, pero en un grupo hay muchas posibilidades que tendríamos un coincidencia, la distribución debe estar cerca a la Poisson. Con $d = 365$ dias en el año, con $n = 23$, solo, la probabilidad es mas que un mitad $1/2$ que tenemos una coincidencia de cumpleaños, bajo la suposición que los fechas de cumpleaños son uniforme, y independiente.

Escriba, con n el numero de personas,

$$W = \sum_{\{\alpha,\beta\} \subset \{1,\ldots,n\}, |\{\alpha,\beta\}|=2} X_{\alpha,\beta},$$

donde $X_{\alpha,\beta}$ tiene valor 1 o 0 si la pareja α, β tiene, o no, la misma cumpleaños. Empezamos con el cálculo de la media $\lambda = E[W]$. Hay n sobre 2, $\binom{n}{2}$, parejas en la suma, y por cada sumando hay la probabilidad $1/d$ que esta pareja tiene el mismo dia, y por lo tanto,

$$\lambda = \frac{1}{d}\binom{n}{2}.$$

Con $d = 365$ y $n = 23$, $\lambda = 0.6931507 > 0.69314 > \log 2$, y la probabilidad Poisson que no hay una coincidencia es $e^{-\lambda} = 0.4999982$, bajo la mitad, correcto! En general, para hacer λ en orden como uno, necesitamos, aproximadamente

$$n^2/(2d) = 1 \quad \text{o} \quad n = \sqrt{2d} \quad \text{es 27, cuando } d = 365$$

Podemos usar el método de Stein para conseguir una cota, usando sesgo de tamaño? Escoge una pareja, en proporcón de la esperanza, pero todo tienen la misma esperanza, entonces escoge una uniformemente, decimos α, β. Si α, β tienen el mismo cumpleaño, no cambia nada. Si no, debemos construir las variables condicional sobre $X_{\alpha,\beta} = 1$. Entonces, solo escoge una de α, β, decimos β, y cambia su cumpleaños para ser el mismo día como α. Podemos verificar que esta produce variables con la correcta distribución condicional. Pero, dada que los cumpleaños son uniforme, son igual tambien cuando son igual, y todos los otros días no cambian, porque son independientes.

No hemos cambiado mucho en las variables, en particular, solo los cumpleaños en las parejas que contiene β pueden cambiar. Por tanto, con $X'_{\gamma,\beta}$ la función característica, después de haber cambiado el cumpleaños de β para ser la misma día de α, y notan que $X'_{\alpha,\beta} = 1$ por construcción, sustraendo 1 de la primera suma en la forma de $X'_{\alpha,\beta}$, tenemos

$$\begin{aligned} W^s - 1 - W = \sum_{\gamma\neq\beta} X'_{\gamma,\beta} - 1 - \sum_{\gamma\neq\beta} X_{\gamma,\beta} &= \sum_{\gamma\notin\{\beta,\alpha\}} X'_{\gamma,\beta} - \sum_{\gamma\neq\beta} X_{\gamma,\beta} \\ &= \sum_{\gamma\notin\{\alpha,\beta\}} (X'_{\gamma,\beta} - X_{\gamma,\beta}) - X_{\alpha,\beta}, \end{aligned}$$

y por eso, la esperanza de su valore absoluta, es acotada por

$$\begin{aligned} E\left[X_{\alpha,\beta} + \sum_{\gamma\notin\{\beta,\alpha\}} |X'_{\gamma,\beta} - X_{\gamma,\beta}|\right] &= E[X_{\alpha,\beta}] + \sum_{\gamma\notin\{\beta,\alpha\}} E\mathbf{1}(X'_{\gamma,\beta} \neq X_{\gamma,\beta}) \\ &= P(X_{\alpha,\beta} = 1) + (n-2)P(X'_{\gamma,\beta} \neq X_{\gamma,\beta}) = \frac{1}{d} + 2(n-2)\frac{1}{d}(1 - \frac{1}{d}) \leq \frac{2n}{d}, \end{aligned}$$

donde hemos usado que por $\gamma \notin \{\alpha, \beta\}$, tenemos $X'_{\gamma,\beta} = X_{\alpha,\beta}$, y por eso

$$\begin{aligned} P(X'_{\gamma,\beta} \neq X_{\gamma,\beta}) = P(X_{\gamma,\alpha} \neq X_{\gamma,\beta}) &= P(X_{\gamma,\alpha} = 1, X_{\gamma,\beta} = 0) + P(X_{\gamma,\beta} = 1, X_{\gamma,\alpha} = 0) \\ &= 2P(X_{\gamma,\beta} = 1, X_{\gamma,\alpha} = 0) = 2P(X_{\gamma,\beta} = 1)P(X_{\gamma,\alpha} = 0) \end{aligned}$$

porque $X_{\gamma,\beta}$ y $X_{\gamma,\alpha}$ son independiente cuando los cumpleaños son uniforme. Por eso

$$d_{\mathrm{TV}}(W, Z_\lambda) \leq 2(1 - e^{-\lambda})\frac{n}{d}.$$

Si $n \sim \sqrt{d}$ la cota decae en orden como $1/\sqrt{d}$.

Para un otro ejemplo, consideramos una permutación π de $\{1,\ldots,n\}$, elegido uniformemente de todas las permutaciónes, y contamos el numero de puntos fijos,

$$W = \sum_{i=1}^{n} X_i \quad \text{donde} \quad X_i = \mathbf{1}(\pi(i) = i).$$

Porque $P(\pi(i) = i) = 1/n$, tenemos $\lambda = n(1/n) = 1$, en la media, hay solo uno punto fijo.

Para construir W^s, escoge I, uniformemente de $\{1,\ldots,n\}$. Si $\pi(I) = I$, dejalo!

Si no, $\pi(I) \neq I$, tenemos $\pi(I) = k$ por alguno $k \neq I$, y debe existe un $j \neq I$, que satisface $\pi(j) = i$,

$$\begin{array}{cccccccc} 1 & 2 & \ldots & j & \ldots & I & \ldots & n \\ \pi(1) & \pi(2) & \ldots & I & \ldots & k & \ldots & \pi(n) \end{array}$$

En este caso, intercambiamos I y k arriba para formar π',

$$\pi(I) = k, \pi(j) = I \quad \text{cambia a} \quad \pi'(I) = I, \pi'(j) = k.$$

Podemos verificar que hemos logrado el distribución correcta.

Esta maniobra solo puede incrementar el numero de puntos fijos, y todos las variables $\pi(m)$, para m no igual a I o j tienen las mismas valores. Para la charla próxima, recuerdan que

$$W^s \leq W + 2.$$

Para obtener una cota entre la distribución de W y la Poisson, debemos acotar $E|W^s - 1 - W|$.

$$\begin{aligned} W^s - 1 - W &= \mathbf{1}(\pi'(I) = I) + \mathbf{1}(\pi'(j) = j) - 1 - \mathbf{1}(\pi(I) = I) - \mathbf{1}(\pi(j) = j) \\ &= \mathbf{1}(\pi'(j) = j) - \mathbf{1}(\pi(I) = I) - \mathbf{1}(\pi(j) = j) \\ &\leq \mathbf{1}(\pi'(j) = j) + \mathbf{1}(\pi(I) = I) + \mathbf{1}(\pi(j) = j) \end{aligned}$$

y

$$E|W^s - 1 - W| \leq \frac{2}{n} + \frac{1}{n-1}.$$

Entonces,

$$d_{TV}(W, Z_\lambda) \leq (1 - e^{-\lambda}) E|W + 1 - W^s| = (1 - e^{-1}) \left(\frac{1}{n-1} + \frac{2}{n} \right).$$

usando $\lambda = 1$. Muy explicita, pero en este caso, el método no puede producir cotas mas buenas, porque este problema es muy especial, la distancia a la Poisson es en realidad muy pequña, exponencial. Pero, recordamos para la próxima charla que en esta ejemplo $W^s \leq W + 2$. Vean [9] and [6].

3 Concentración de Medida

El teorema central da nos una approximación por la probabilidad que una variable toma valores en, decimos, un intervalo, que es buena cuando el tamaño de la muestra n es grande. Y es siempre una cuestión de cuanto error la approximación tiene, y si el tamaño de la muestra es suficiente grande. Las técnicas de la concentración de medida da nos una cota para la cola para una variable X con media μ que es correcta por todos n en la forma

$$P(X \geq \mu + x) \leq a(x)$$

con una función a explicita, y que decae rápidamente en $x \geq 0$. Ademas puede ser desigualdades para la cota abajo, y entonces para $P(|X - \mu| \geq x)$ también.

Un ejemplo bien conocido es la desigualad de Azuma-Hoeffding, sobre las funciones de diferencias cotadas. Decimos que $L : \mathbb{R}^n \to \mathbb{R}$ es un función de diferencias cotadas si existen $c_1, \ldots, c_n$ tal que

$$|L(x_1, \ldots, x_i, \ldots, x_n) - L(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i \quad \text{para cada } i = 1, \ldots, n.$$

Por ejemplo, si

$$L(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$$

y las variables x_i satisfacen $|x_i| \leq 1$, la función L tiene la propriedad de diferencias cotadas, con $c_i = 1$. La teorema de Azuma-Hoeffding dice que si las variables aleatorias $X_1, \ldots, X_n$ son independientes, y $X = L(X_1, \ldots, X_n)$, obtenemos

$$P\left(|X - EX| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}\right),$$

y vemos que la cota es en la forma de una variable Gaussiana.

Con esta desigualdad, podemos obtener una cota para la cola de L en el problema de la más larga secuencia común. Un ejemplo donde $L_n(X, Y) = 4$ por $n = 10$

$$\begin{array}{cccccccccc} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \end{array}$$

El teorema ergódico subaditivo dice que, si las variables son independientes L_n obedece la ley de números largos,

$$\lim_{n \to \infty} \frac{L_n}{n} \to_{\text{a.s.}} c,$$

pero il valore de c no es conocido, demostrando que el problema es difícil. Pero, (ejercicio!) L es una función de diferencias cotadas, y por eso, $L(X, Y)$ es concentrada.

El teorema de Azuma-Hoeffding, y otras, necesitan independencia de las variables. Pero, con algunas ideas de Stein, podemos obtener resultados que no la necesitan. Empezamos con le ecuación que caracteriza la distribución de X^s que tiene la distribución sesgo de tamaño de X, una variable no-negativa, y con una media μ, finito,

$$E[X f(X)] = \mu E[f(X^s)].$$

Recuerda que para nuestra usa del método de Stein necesita un acoplamiento (X, X^s). Decimos que uno acoplamiento es acota de un constante c cuando existe un constante c que

$$W^s \leq W + c. \tag{3}$$

Esta teorema es debido a [2], y ademas vean [11] para una generalización donde no necesitan que la suposición satisface (4) con probabilidad uno.

Teorema 3.1. *Sean (W, W^s) una acoplamiento de W y W^s, donde W^s tiene la distribución sesgo de tamaño de W, satisface*

$$W^s \leq W + c \quad \textit{por una constante } c > 0. \tag{4}$$

Entonces, $\mu = E[W]$ existe, y por todas $x \in \mathbb{R}$

$$P(W \geq x + \mu) \leq \exp\left(-\frac{\mu}{c} h\left(\frac{x}{\mu}\right)\right),$$

con $h(x) = (1+x)\log(1+x) - x$.

También,

$$\exp\left(-\frac{\mu}{c} h\left(\frac{x}{\mu}\right)\right) \leq \exp\left(-\frac{x^2}{2c\,(x/3+\mu)}\right).$$

La segunda desigualdad es más fácil usar como la primera, pero esconde el hecho que la cola decline como $\exp(-ax\log x)$, porque muestra solo la velocidad como $\exp(-ax)$. Ne es tan rápido come la gaussiana, que decade con la velocidad $\exp(-bx^2)$, pero es la velocidad de la cola de la Poisson.

Por ejemplo, si tenemos la suma

$$W = \sum_{i=1}^{n} X_i$$

de variables no-negativo, independientes, acotada de c, tenemos

$$W^s = W - X_I + X_I^s \leq W + X_I^s \leq W + c,$$

porque el suporte de X es un subconjunto del suporte de X^s, recuerde $\mu dF^s(x) = xdF(x)dx$. Inmediatamente, podemos declarar que W es concentrado. No hay más condiciones. Para el Poisson, $X^s =_d X + 1$, entonces es concentrada.

Para el numero de puntos fijos

$$W = \sum_{i=1}^{n} \mathbf{1}(\pi(i) = i).$$

en una permutación elegido uniformemente, obtenemos $W^s \leq W + 2$. Entonces W es concentrado.

Recordamos que para construir un acoplamiento, escoge I uniformemente, perche aquí todas las variables en la suma tiene la misma esperanza, y sobre el evento $I = i$, encuentra la variable k tal que $\pi(k) = i$ que y intercambiamos i y k abajo,

$$\pi(i) = k, \pi(j) = i \quad \text{cambia a} \quad \pi'(i) = i, \pi'(j) = k.$$

Creamos un punto fijo $\pi'(i) = i$, y posiblemente un otro a j. Inmediatamente, obtenemos que

$$W^s \leq W + 2.$$

Notan que aquí las variables son dependientes.

Ahora, deja S una muestra de n elementos de $\{1, \ldots, N\}$, sin reemplaza, y $\{y_1, \ldots, y_N\}$ nmeros positivos con $N > n$, y deja

$$W = \sum_{i=1}^{N} y_i \mathbf{1}(i \in S).$$

Usando nuestra receta, elegimos una i con probabilidad en proporción a y_i, reemplaza $y_i \mathbf{1}(i \in S)$ con y_i, y modificar las otras variables para tener la distribución condicional correcta. En otras palabras, debemos poner y_I en la muestra. Pero el tamaño de la muestra es limitado a n, entonces, si y_I ya no esta en la muestra, debemos quitar una variable que es, decimos y_J, donde J es elegido uniformemente sobre todas las variables que son elementos in S. Encontnces,

$$W^s = W - y_J + y_I \leq W + y_I \leq W + c \quad \text{donde} \quad c = \max_{i=1,\ldots,n} y_i$$

La distribución de W es cotada. h Aqui esta la prueba del teorema. Puedes mostrar que la función generadora de momentos

$$M(t) = E[e^{tW}]$$

existe solo bajo condición (3), y consecuentemente la media $\mu = E[W]$ existe también. Ahora, podemos tomar la derivada abajo la esperanza, y obtener que la derivada de M satisface, para cada $t \geq 0$,

$$M'(t) = E[We^{tW}] = \mu E[e^{tW^s}] \leq \mu E[e^{t(W+c)}] = \mu e^{tc} E[e^{tW}] = \mu e^{tc} M(t),$$

y hemos obtenido la desigualdad diferencial

$$\frac{M'(t)}{M(t)} \leq \mu e^{tc} \quad \text{o} \quad (\log M(t))' \leq \mu e^{tc} \quad \text{o} \quad \text{o} \quad M(t) \leq \exp\left(\frac{\mu}{c}(e^{tc} - 1)\right),$$

usando la condición de contorno que $\log M(0) = 0$. Ahora, aplicando la desigualdad de Markov, que dice que

$$P(|X| \geq t) = \int_{|x| \geq t} dF(x) \leq \int_{|x| \geq t} \frac{|x|}{t} dF(x) \leq \int \frac{|x|}{t} dF(x) = \frac{E|X|}{t},$$

y todavia tomando $t \geq 0$,

$$P(W \geq x + \mu) = P(e^{tW} \geq e^{t(x+\mu)}) \leq \frac{M(t)}{e^{t(x+\mu)}} \leq \exp\left(\frac{\mu}{c}(e^{tc} - 1) - t(x + \mu)\right). \tag{5}$$

Esta desigualdad es verdad para todo $t \geq 0$, substituyendo

$$t = \log(x/\mu + 1)/c$$

da nos la cota

$$P(W \geq x + \mu) \leq e^{x/c} \left(\frac{x}{\mu} + 1 \right)^{-(x+\mu)/c}.$$

Para que la parte primera en (5), hace

$$e^{tc} - 1 = x/\mu \quad \text{o} \quad \frac{\mu}{c}(e^{tc} - 1) = x/c \quad \text{y} \quad \exp\left(\frac{\mu}{c}(e^{tc} - 1) \right) = e^{x/c}$$

y para la otra parte,

$$\exp\left(-t(x + \mu)\right) = \exp\left(-\log(x/\mu + 1)(x + \mu)/c\right) = \left(\frac{x}{\mu} + 1 \right)^{-(x+\mu)/c}.$$

Para la suma arriba W, de variables independientes, acotada de c, hemos obtenido que $W^c \leq W + c$, y entonces la cola de W decade con al meno la misma velocidad de un Poisson, $\exp(-t \log t)$,

Para un otro ejemplo, tomamos un grafo G, con n vértices, y sobre cada vértice $v \in V$ ponemos una variable U_v, decimos, $\mathcal{U}[0, 1]$, independiente. Decimos que el grafo G es un grafo de grado acotado cuando el numero de aristas de cada vértice es cotada, decimos por d.

Declaramos un vértice v un máximo local cuando $U_v \geq U_w$ para cada vecino w de v. Queremos decir algo sobre la distribución del numero de máximos locales. Sea N_v el entorno de v, el conjunto de los vecinos de v.

$$W = \sum_{v=1}^{n} X_v, \quad \text{donde } X_v = \mathbf{1}(U_v \geq U_w, \forall w \in N_v).$$

En este caso, las variables son dependientes. Si $X_v = 1$, la variable U_v es 'grande', y hace imposible que un vecino w es también un máximo local.

El primero problema es como construir variables X_v que han la distribución condicional sobre $X_v = 1$. Si $X_v = 1$ el vértice v ya es un máximo local, no tocamos nada, y basta. Las otras variables ya tenemos la distrubución correcta. Pero, si $X_v = 0$, significa que existe al menos un vecino que ha un variable con tamaño mas grande de U_v. Toma la variable en el entorno de v con el tamaño mas grande,

$$U_w > U_y, y \in N_v,$$

y construimos el grafo G' con las variables U_v and U_w intercambiados. Esta maniobra hace v un máximo local, y deja que las otras variables tienen la distribución correcta, condicional de $X_v = 1$. Se da cuenta que no hemos cambiado mucho entre G y G', no mas que d, $W^s \leq W + d$, y la distribución esta concentrada. Vean [5] para esta construcción.

Para un otro ejemplo, decimos que tenemos n pelotas y las ponemos, cada una, en una de m cajas, (para recordar bien, tal vez dice nelotas y majas) uniformemente y

independiente. Sea W cuenta el numero de cajas vacías, esta es, el numero de cajas que no tienen pelotas. Es la distribución de W concentrada? Mas generalmente, podemos considerar el numero de cajas que tienen d o menos pelotas,

$$W = \sum_{i=1}^{m} \mathbf{1}(N_i \leq d),$$

donde $N_i, i = 1, \ldots, n$ cuentan cuantas pelotas hay in caja i. Toma $d = 0$, el caso que

$$W = \sum_{i=1}^{m} \mathbf{1}(N_i = 0),$$

que cuenta el numero de cajas vacias, para ilustrar el principio.

Nota que hay dependencia entre los numeros N_i. Si N_i es grande, es más probable que N_j es pequeña, para cada $j \neq i$. Ademas, es fácil construir una no-cotada sesgo de tamaño acoplamiento. Escoge una caja I, uniformemente. Si no hay pelotas en esta caja, $N_I = 0$, dejalo. Si $N_i \geq 1$, debemos vaciar la caja, entonces ponga todas las pelotas in esta caja uniformemente en otras cajas. Esta nueva variable tiene la correcta distribución condicional, pero hay una probabilidad que sean muchos pelotas en la caja, y ellas son puestas en otras cajas vacías, destruyendolos, cambiando W por mucho, y no podemos satisfacer la desigualdad $W^s \leq W + c$ con un constante c, independiente de n.

La primera truca para construir un acoplamiento acotado es de reemplazar W por $m - W$, el numero de cajas non-vacias, y W es concentrada si y solo si $m - W$ es concentrada, invirtiendo las colas arriba y abajo. Entonces, consideramos

$$W = \sum_{i=1}^{m} \mathbf{1}(N_i \neq 0).$$

Elije una caja I, uniformemente. Recordamos que necesitamos arreglar las pelotas para lograr $P(\cdot|N_I \neq 0)$. Si caja I no es vaciá, $N_i \neq 0$, la dejamos. Si $N_I = 0$ tenemos que hace no vacia, para obtener la distribucón condicional sobre el evento $N_I \neq 0$. Imaginamos distribuyendo la masa probabilidad del evento $\{N_i = 0\}$ sobre el eventos $N_i = k, k \geq 1$ para lograr esta distribución condicional. Podemos distribuir la masa a una vez a estes otros eventos. Pero, podemos ademas distribuirla en una maniera menos 'caótico'. Primeramente, movimos toda la masa de $N_i = 0$ a $N_i = 1$. Pero ahora sera demasiado masa sobre el evento $N_i = 1$, y tenemos demasiado cajas con una sola pelota. Entonces, retenemos solo la masa necesaria para lograr $P(N_i = 1|N_i \neq 0)$, y mueva el resto al evento $N_i = 2$, etcétera. Cada vez que movimos la masa, habrá menos y menos para mover. Por eso, al fine de todas, tenemos tener una probabilidad para mover una pelota en todos los casos, excepto cuando la caja tiene todas las pelotas.

Ahora convertimos esta idea a una acoplmaniento. Elige una caja I y una pelota, J en una caja diferente de I, uniformemente y independiente. Mueve pelota J de su caja a la caja I con una probabilidad π_{N_I} que depende en N_I,

$$\pi_k = \begin{cases} \frac{P(N>k|N>0)-P(N>k)}{P(N=k)(1-k/n)} & 0 \leq k \leq n-1 \\ 0 & k = n. \end{cases}$$

Podemos verificar que la función π_k decae en k. Por ejemplo, si caja I esta vacía, $N_I = 0$ y la formula nos da

$$\pi_0 = \frac{P(N > 0|N > 0) - P(N > 0)}{P(N = 0)} = \frac{1 - P(N > 0)}{P(N = 0)} = 1$$

que dice tenemos mover pelota J, haciendo la caja I no vacía. Y si ya hay una pelota, añadimos una otra pelota con probabilidad un poquito menos de 1, y hay n pelotas, no seria posible añadir nada, y vemos que en este caso la formula nos da $\pi_n = 0$. Porque estamos moviendo al máximo solo una pelota, no cambia mucho el valore de W, y el acoplamiento esta cotada. Aquí tenemos

$$W^s \leq W + 1,$$

porque solo podemos crear a máximo una nueva caja con una pelota non solada, en el caso $N_I = 0$.

Podemos usar esta técnica en otras situaciones. Por ejemplo, consideramos el grafo de Erdős-Rényi con n vértices, y una arista conectando cada pareja de vértices distintas con probabilidad p, independiente de todas las otras aristas. Podemos estudiar las grados de las vértices, y preguntar, por ejemplo, si el numero W de las vértices soladas, que non estan conectado, es concentrada. El argumento para mostrar que lo es continua sobre las misma lineas como en el caso de pelotas en cajas. Si encontramos una vértice solada, añadimos una arista con probabilidad 1, y en general, si la vértice tiene grado k, añadimos una arista con probabilidad π_k, una función decreciente en kg; vean [16], [7] para este, y otros, ejemplos.

Hay generalizaciónes de esta técnica que obtiene concentración usando sesgo de tamaño en sin el requerido que la acoplamiento seria sesgado, bajo la condición, para obtener la cota sobre la cola arriba,

$$P(W \geq x + \mu/p) \leq \exp\left(-\frac{\mu}{cp} h\left(\frac{px}{\mu}\right)\right),$$

abajo la condición que existe $p \in (0, 1]$ tal que

$$P(X^s \leq X + c|X^s \geq x) \geq p \quad \text{para cada } x \geq 0,$$

y con una condición similar para la cola abajo. El caso de que hablamos hoy es el caso especial de que $X^s \leq X + c$ con probabilidad uno, y podemos tomar $p = 1$ en esta desigualdad. Vean [11] para el resultado, y aplicaciones sobre los auto-valores de la matriz adyacencia de grafos 'regular' aleatorias.

4 Otras distribuciónes, y la Dickman

Hasta ahora, hemos visto el método de Stein para el Gaussiano y la Poisson. Recuerde, para aplicar el método debemos encontrar una ecuación que caracteriza la distribución. No hay sola una ecuación que funciona, y si tenemos una, podemos construir infinitamente mas, usando transformaciones. Debemos elegir una ecuación que funciona para el propósito. Anche no hay una algorítmica definida para producir una una ecuación que

funciona, una maniera para construir una ecuación caracterstica es por la densidad $p(x)$, si lo, y su derivada existen y son regular. Sea $p(x) > 0$ sobre $(-\infty, \infty)$. Fácilmente, bajo condiciones de regularidad, por ejemplo, su la comportamiento de la densidad a $\pm\infty$, X ha densidad p si y solo si

$$E[f'(X)] = -E\left[\frac{p'(X)}{p(X)}f(X)\right] \quad \text{para cada } f \text{ tal que existen estas esperanzas.}$$

Es facile verificar esta enunciado, usando integración por partes, porque, bajo regularidad,

$$E\left[\frac{p'(X)}{p(X)}f(x)\right] = \int_{-\infty}^{\infty}\frac{p'(x)}{p(x)}f(x)p(x)dx$$
$$= \int_{-\infty}^{\infty}p'(x)f(x)dx = -\int_{-\infty}^{\infty}f'(x)p(x)dx = -E[f'(X)].$$

Y esta observación nos conduce a la ecuación de Stein

$$f'(x) + \frac{p'(x)}{p(x)}f(x) = h(x) - E[h(X)].$$

En el caso Gaussiano,

$$p(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2), \quad \frac{p'(x)}{p(x)} = -x, \quad \text{y obtenemos} \quad f'(x) - xf(x) = h(x) - E[h(X)].$$

Hay algo similar para las distribuciones discretas.

Prefiero presentar algo diferente hoy, un caso interesante en cual estas formulas no nos ayudan. Espero en esta maniera mostrar que el campo de distribuciones a las cuales el método funciona es grande, y desconocido ya. Que es, no sabemos exactamente el ámbito en que el método funciona.

Entonces, he elegido hablar sobre la distribución Dickman, un ejemplo un poquito inusual, y no bien conocido. Tomamos variables $B_1, \ldots, B_n$, con $B_k \sim \text{Bern}(1/k)$, y formamos el promedio ponderado

$$W_n = \frac{1}{n}\sum_{k=1}^{n} kB_k. \tag{6}$$

Es fácil ver que $E[W_n] = 1$, y

$$\text{Var}(W_n) = \frac{1}{n^2}\sum_{k=1}^{n} k^2(1-1/k)/k = \frac{1}{n^2}\sum_{k=1}^{n} k(1-1/k) = O(1).$$

Se piensas que la distribución W_n seria fácil expresar en una forma sencillo, trata lo. No es. Hay una aproximación? Es una suma de variables independiente, pero (ejercicio), no podemos usar el teorema de limite central.

El limite es la distributión Dickman, que fue descubierto en 1930 en conexión con los números primos [13]. Recordamos el teorema de numeros primos, deja que $\pi(n)$ es el numero de primos n o menos. Por ejemplo,

$$\pi(12) = 5 \quad \text{contando los numeros } 2, 3, 5, 7 \text{ y } 11.$$

Entonces

$$\pi(x) \sim x/\log x.$$

Pero, ademas estamos interesados en los tamaños de las factores de números compuestos. Para x y y sea $\Phi(x, y)$ el numero de enteros menos que x que son compuestos de factores primos menos o igual a y; numeros con factores primos 'pequeñas'. Por ejemplo

$$\Phi(12, 3) = 7 \quad \text{contando los siete numeros } 2, 3, 4, 6, 8, 9, 12$$

Y tenemos el teorma

$$\Phi(x, x^{1/a}) = x\rho(a) + O(x/\log x),$$

donde ρ, la función Dickman, satisface la ecuación diferencial con retardo

$$u\rho'(u) + \rho(u-1) = 0, \quad \text{con la condición de contorno } \rho(u) = 1 \text{ por } 0 \le u \le 1.$$

Y la variable D tiene la densidad hecho por rescalando ρ para hacerla una densidad.

Nuestro objetivo hoy sera un cota por la aproximación de la variable W_n en (6) por D. Pero primera, da una ojeada sobre un problema, relacionado pero mas difícil, en el teoría de números aleatorios. La pregunta es: si elegimos un numero aleatoriamente, podemos decir algo sobre la factorización en un producto de números primos? Dejan una enumeración de los números primos, $p_1 < p_2 < \ldots$, por ejemplo $p_1 = 2, p_2 = 3$. Supongamos que elegimos un numero M_n in Ω_n, los numeros cuyas mas grande factor prima es no mas que p_n, con la probabilidad Π_n, donde

$$P(M_n = m) = \frac{1}{\pi_n m} \quad \text{lo que es, proporcional a } 1/m.$$

Ahora, consideramos el logarítmico de M_n, rescalando apropiadamente,

$$S_n = \frac{\log M_n}{\log p_n}.$$

Tenemos

$$S_n \to D \quad \text{pero, con el método, podemos decir más:} \quad d_{1,1}(S_n, D) \le \frac{C}{\log n},$$

donde, con α, β positivo,

$$\mathcal{H}_{\alpha,\beta} = \{h : h \in \mathrm{Lip}_\alpha, h' \in \mathrm{Lip}_\beta\},$$

y

$$d_{1,1}(X, Y) = \sup_{h \in \mathcal{H}_{1,1}} |Eh(X) - Eh(Y)|,$$

el métrico Wasserstein 2.

Tal vez es sorprendente, pero podemos representar la distribución de M_n en un otra maniera. Escribimos $X \sim \mathrm{Geom}(p)$ cuando $P(X = m) = p(1-p)^m, m \ge 0$. Entonces

$$M_n =_d \prod_{k=1}^{n} p_k^{X_k} \quad \text{con} \quad X_k \sim \mathrm{Geom}(1 - 1/p_k) \quad \text{y independiente.}$$

El método empieza con una ecuación que caracteriza la distribución, que en el caso de gaussiano es

$$E[f'(x) - xf(x)] = 0,$$

o con una transformación de cual la distribución es un punto fijo, en el caso de gaussiano,

$$X \to X^* \quad \text{donde } X^* \text{ satisface} \quad \sigma^2 E[f'(X^*)] = E[Xf(X)].$$

Para el Dickman, tenemos al meno dos ecuaciones características. Pero el método usando la densidad es inútil, porque no podemos esribir $\rho'(u)/\rho(u)$ explcitamente.

Sin embargo, una de las ecuaciones características es una ecuación integral, (voy a llamarla la ecuación g)

$$E[g(X) - A_{X+1}g] = 0 \quad \text{donde} \quad A_x h = \frac{1}{x}\int_0^x h(u)du \quad \text{el operador integral de media}$$

que nos lleva a la ecuación

$$g(x) - A_{x+1}g = h(x) - E[h(D)]$$

y otra una ecuación diferencial con anticipo (llamada la ecuación f),

$$E[Xf'(X) + f(X) - f(X+1)] = h(x) - E[h(D)].$$

Sin perdiendo generalidad, por reemplazando $h(x) - E[h(D)]$, podemos suponer que $E[h(D)] = 0$.

Mas comprensible, es la transformación

$$X \to X^* \quad \text{donde} \quad X^* = U(X+1) \quad \text{donde } U \sim \mathcal{U}[0,1], \text{ independiente de } W,$$

que tiene D como el único punto fijo, que es, $W \sim D$ si y solo si $W^* =_d W$. Cuando comparamos los casos normal y el Dickman, una, el gaussiano, tiene una densidad sencillo y una transformación complicada y la otra, el Dickman, el contrario.

Obtenemos la cota, sin el uso del método de Stein, hay una maniera mas directa,

$$d_1(W, D) \leq 2d_1(W, W^*).$$

Podemos usar esta cota en las situaciones en cual podemos construir un acoplamiento de W y W^*.

Los dos ecuaciones son relacionadas, si g es la solución de la ecuación integral, $f(x) = A_x g$ resuelve la otra. Parece que no es posible encontrar la solución de la ecuación g en una maniera fácil, pero podemos escribir una solución como una serie infinito

$$g(x) = \sum_{n\geq 0} A_{x+1}^{(n)} h \quad \text{donde} \quad A_x^{(0)} h = h, A_x^{(n+1)} h = A_x(A_x^{(n)} h), n \geq 0.$$

Tenemos

$$g(x) - A_{x+1}g = \sum_{n\geq 0} A_{x+1}^{(n)} h - \sum_{n\geq 1} A_{x+1}^{(n)} h = A_{x+1}^{(0)} h = h(x)$$

De esta observación, podemos demostrar que cuando $h \in \text{Lip}_1$ con $E[h(D)] = 0$, tenemos la solución $g \in \text{Lip}_2$. Ahora, usando $f = A_x g$, obtenemos que para cada $h \in \mathcal{H}_{1,1}$ con $E[h(D)] = 0$,

$$\|f'\| \leq 1 \quad \text{and} \quad \|f''\| \leq 1/2.$$

que podemos usar para obtener cotas en el metrico Wasserstein 2.

Regresamos a la suma arriba,

$$W_n = \frac{1}{n}\sum_{k=1}^{n} kB_k \quad \text{donde} \quad B_k \sim \text{Bern}(1/k),$$

y recordamos la ecuación de Stein para el Dickman, en la forma f,

$$xf'(x) - f(x+1) - f(x) = h(x) - E[h(D)],$$

y sustituimos W_n por x, y tomamos la media, dando,

$$E[W_n f'(W_n) - f(W_n + 1) - f(W_n)] = E[h(W_n)] - E[h(D)].$$

Deja que

$$W_n^{(k)} = W_n - \frac{k}{n}B_k,$$

y para el primero término de la ecuación, tenemos

$$\begin{aligned} E[W_n f'(W_n)] &= E\left[\frac{1}{n}\sum_{k=1}^{n} kB_k f'(W_n)\right] = \frac{1}{n}\sum_{k=1}^{n} E\left[kB_k f'\left(W_n^{(k)} + \frac{k}{n}B_k\right)\right] \\ &= \frac{1}{n}\sum_{k=1}^{n} E\left[kf'\left(W_n^{(k)} + \frac{k}{n}\right)\right] P(B_k = 1) = \frac{1}{n}\sum_{k=1}^{n} E\left[f'\left(W_n^{(k)} + \frac{k}{n}\right)\right]. \end{aligned}$$

Entonces, el lado izquierda es

$$\begin{aligned} &\frac{1}{n}\sum_{k=1}^{n} f'\left(W_n^{(k)} + \frac{k}{n}\right) - \int_0^1 f'(W_n + u)du \\ &\qquad = \frac{1}{n}\sum_{k=1}^{n}\left(f'\left(W_n^{(k)} + \frac{k}{n}\right) - f'\left(W_n + \frac{k}{n}\right)\right) \\ &\qquad\qquad + \left(\frac{1}{n}\sum_{k=1}^{n} f'\left(W_n + \frac{k}{n}\right) - \int_0^1 f'(W_n + u)du\right), \end{aligned}$$

sumando y sustraendo el mismo termino.

Para la esperanza del primero término, usando $\|f''\| \leq 1/2$, obtenemos la cota

$$\frac{\|f''\|_\infty}{n}\sum_{k=1}^{n} E|W_n^{(k)} - W_n| \leq \frac{1}{2n}\sum_{k=1}^{n} E\left[\frac{k}{n}B_k\right] = \frac{1}{2n}.$$

Y para el segundo, usando la misma cota sobre la segunda derivada, tenemos

$$\left| \frac{1}{n}\sum_{k=1}^{n} f'\left(W_n + \frac{k}{n}\right) - \int_0^1 f'(W_n+u)du \right| \leq \sum_{k=1}^{n} \int_{\frac{k-1}{n}}^{\frac{k}{n}} \left| [f'(W_n+k/n) - f'(W_n+u)] \right| du$$
$$\leq \frac{1}{2}\sum_{k=1}^{n} \int_{\frac{k-1}{n}}^{\frac{k}{n}} (k/n-u)du = \frac{1}{2}\left(\frac{1}{n^2}\sum_{k=1}^{n} k - \int_0^1 udu\right) = \frac{1}{4n}.$$

Combinando estas dos cotas,

$$|E[h(W_n)] - E[h(D)]| \leq \frac{3}{4n}$$

Y tomando el supremo sobre $\mathcal{H}_{1,1}$ y recordando la definición de $d_{1,1}$ tenemos la cota

$$d_{1,1}(W_n, D) \leq \frac{3}{4n}.$$

Regresando al problema de la factorización de números aleatorias, para este, usamos una otra caracterización de D, que dice D es la única variable que satiface

$$D^s =_d D + U,$$

con $U \sim \mathcal{U}[0,1]$, independiente de D. Entonces, dada W, si podemos construir una acoplamiento $(W^s, W+T)$ donde W^s tiene la distribución sesgo de tamaño de W, y la distribución de T esta cerca de $\mathcal{U}[0,1]$ con U independiente de W, debe ser que la distribución de W esta cerca de D. Aquí esta un teorema que hace esta idea mas preciso:

Teorema 4.1. *Dada W una variable no negativo con media μ y varianza finito y no cero. Si existe una variable T que satisface $W+T>0$ con probabilidad uno, y*

$$E[W\phi(W)] = \mu E[\phi(W+T)] + R_\phi \quad \textit{for all } \phi \in \mathrm{Lip}_{1/2}.$$

entonces

$$d_{1,1}(W,D) \leq |\mu - 1| + \frac{1}{2}\inf_{(T,U)} E|T-U| + \sup_{\phi \in \mathrm{Lip}_{1/2}} |R_\phi|$$

donde el ínfimo es tomado sobre todas las parejas (T,U) construidos sobre el mismo espacio como W, con U y W independientes.

En nuestro caso, tenemos

$$W_n = \frac{\log M_n}{\log p_n} = \frac{1}{\log(p_n)}\sum_{i=1}^{n} X_i \log(p_i).$$

Para el medio, encontramos que

$$\mu_n = E[W_n] = \frac{1}{\log(p_n)}\sum_{k=1}^{n} \frac{\log(p_k)}{p_k - 1}.$$

Podemos aplicar la recete usual para construir W_n^s, que es, eligiendo una variable $I \in \{1, \ldots, n\}$ con la probabilidad $P(I = k)$ en proporción a la media del sumando k, que significa

$$P(I = k) = \frac{\log(p_k)}{(p_k - 1)\log(p_n)\mu_n} \quad \text{for } k \in \{1, \ldots, n\}.$$

con I independiente de W_n.

Usando la receta, tenemos

$$E[W_n \phi(W_n)] = \mu_n E[\phi(W_n + T_n)] + R_{n,\phi},$$

con una largo expresión para $R_{n,\phi}$ que no consideramos hoy, y

$$T_n = \frac{\log(p_I)}{\log(p_n)} \quad \text{and} \quad \mu_n - 1 = O\left(\frac{1}{\log n}\right).$$

Al fine de cuentos, podemos demonstrar que

$$E|T_n - U| = O\left(\frac{1}{\log n}\right).$$

$U \sim \mathcal{U}[0, 1]$ independiente de W_n, usando resultados del teoría de numeros como el teorema de numeros primos $\pi(n) \sim n/\log n$. Con esta, obtenemos la cota

$$d_{1,1}(W_n, D) = O(1/\log n).$$

Para más información, vean [8], [12], [13], [15], [18], [19] y [20].

References

[1] Arras, B., Mijoule, G., Poly, G. and Swan, Y. (2016). Distances between probability distributions via characteristic functions and biasing. `https://arxiv.org/abs/1605.06819v1`.

[2] Arratia, R. and Baxendale, P. (2015) Bounded size bias coupling: a Gamma function bound, and universal Dickman-function behavior. Probability Theory and Related Fields 162.3-4: 411-429.

[3] Arratia, R., Goldstein, L. and Gordon, L. (1989) Two moments suffice for Poisson approximations: the Chen-Stein method." The Annals of Probability: 9-25.

[4] Arratia, R., Goldstein, L. and Gordon, L. (1990) Poisson approximation and the Chen-Stein method. Statistical Science: 403-424.

[5] Baldi, P., Rinott, Y. and Stein, C. (1989) A normal approximation for the number of local maxima of a random function on a graph. Probability, statistics, and mathematics. 59-81.

[6] Barbour, A., Holst, L. and Janson, S. (1992) Poisson Approximation.

[7] Bartroff, J., Goldstein L. and Işlak, Ü. (2014) Bounded size biased couplings, log concave distributions and concentration of measure for occupancy models. *Bernoulli*, to appear. https://arxiv.org/abs/1402.6769

[8] Bhattacharjee, C. and Goldstein, L. (2018) Dickman approximation in simulation, summations and perpetuities, https://arxiv.org/abs/1706.08192.

[9] Chen, L.H.Y. (1975). Poisson approximation for dependent trials. *Ann. Prob.*, 534–545.

[10] Chen, L.H.Y., Goldstein, L, and Shao, Q.M. (2010). Normal approximation by Stein's method. Springer.

[11] Cook, N., Goldstein, L. and Johnson, T. (2018) Size biased couplings and the spectral gap for random regular graphs. The Annals of Probability 46.1: 72–125.https://arxiv.org/abs/1510.06013

[12] Devroye, L. and Fawzi, O. (2010). Simulating the Dickman distribution. *Statistics and Probability Letters*, 80(03), 242-247.

[13] Dickman, K. (1930). On the frequency of numbers containing prime factors of a certain relative magnitude. *Ark. Mat. Astr. Fys.*, 22(10), 1-14.

[14] Goldstein, L. (2004) Normal approximation for hierarchical structures. The Annals of Applied Probability 14.4: 1950-1969. https://arxiv.org/abs/math/0503549

[15] Goldstein, L. (2017). Non asymptotic distributional bounds for the Dickman approximation of the running time of the Quickselect algorithm. https://arxiv.org/abs/1703.00505

[16] Goldstein, L. and Penrose, M. (2010) Normal approximation for coverage models over binomial point processes. The Annals of Applied Probability 20.2 (2010): 696–721 https://arxiv.org/abs/0812.3084

[17] Goldstein, L. and Reinert, G. (1997). Stein's method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Prob.*, 7(04), 935-952. https://arxiv.org/abs/math/0510619

[18] Hwang, H. K. and Tsai, T. H. (2002). Quickselect and the Dickman function. *Combinatorics, Probability & Computing*, 11(04), 353-371.

[19] Pinsky, R. (2016). A Natural Probabilistic Model on the Integers and its Relation to Dickman-Type Distributions and Buchstab's Function. https://arxiv.org/abs/1606.02965.

[20] Pinsky, R. (2016). On the strange domain of attraction to generalized Dickman distributions for sums of independent random variables. https://arxiv.org/abs/1611.07207.

[21] Ross, N. (2011). Fundamentals of Steins method. *Probability Surveys*, 8, 210-293.https://arxiv.org/abs/1109.1880

[22] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. Proc. Sixth Berkeley Symp. Math. Stat. Prob., 583-602.

[23] Stein, C. (1986). Approximate Computation of Expectations. Institute of Mathematical Statistics, Hayward, CA.