



BOURBAKI

COLEGIO DE MATEMÁTICAS

Índice

- 01. Introducción _____ pág. 02
- 02. Lectura de referencia: el rol de la estadística en la física de partículas ____ pág. 04
- 03. Test de Pearson para la comparación _____ pág. 05
- 04. Test de Kolmogorov-Smirnov _____ pág. 11

01 Introducción

Bienvenidos a nuestro curso de Matemáticas Avanzadas para la Ciencia de Datos, nuestro curso tiene tres módulos dedicados a estudiar las ideas matemáticas más útiles para comprender los algoritmos y modelos matemáticos más comunes en Ciencia de Datos. Los tres módulos son los siguientes

- Probabilidad y Estadística
- Álgebra Lineal
- Optimización y cálculo diferencial

Todos los módulos tienen una duración de 8 semanas. El curso está acompañado de ejercicios y tareas en Python para practicar y reforzar los conocimientos aprendidos así como las implementaciones en bases de datos de los algoritmos estudiados. Pueden consultar el repositorio de esta semana en [este link](#).

La estructura de cada una de las semanas es la siguiente:

1. Treinta minutos dedicados a estudiar un artículo de referencia que motivará los conceptos matemáticos de esta semana.
2. Dos horas dedicadas a estudiar el tema de la semana y algunos ejercicios.

3. Una hora y media dedicada a practicar lo aprendido utilizando Python.

El primer módulo de probabilidad consta de los siguientes temas:

1. Kolmogorov, independencia y condicionamiento
2. Variables aleatorias discretas y sus momentos
3. Ley de los grandes números y máxima verosimilitud
4. El teorema límite central y los intervalos de confianza
5. Tests estadísticos
6. Inferencia bayesiana
7. Cadenas de markov y muestreros de gibbs
8. Redes bayesianas

02 Lectura de referencia: el rol de la estadística en la física de partículas

En el área de la física que estudia la estructura de la materia los tests estadísticos son muy importantes para el análisis de los experimentos.

Recientemente uno de los descubrimientos más importantes de esta área es la comprobación en 2012 de la existencia del mítico Bosón de Higgs, el cual se descubrió utilizando los experimentos del Gran colisionador de hadrones que forma parte del proyecto de investigación europeo CERN.

La lógica detrás del razonamiento para poder concluir la existencia del Bosón de Higgs no sigue la estructura por ejemplo de la lógica aristotélica o proposicional, sino un razonamiento estadístico basado en los resultados de Neyman–Pearson los cuales serán el eje de esta semana.

Para conocer más sobre las implicaciones de los tests estadísticos en esta área de la física recomendamos al lector revisar [1].

03 Test de Pearson para la comparación

La probabilidad es la ciencia que después de fijar a una ley de probabilidad estudia las propiedades límites de la experiencia aleatoria asociada, por el contrario la estadística es el proceso inverso.

Definition 00.1. Sea (Ω, \mathbb{P}) un espacio de probabilidad y $S \in \Omega^n$, una estadística se define como un espacio de probabilidad (Ω, \mathbb{P}_S) obtenido a partir de S que pretende aproximar al espacio original.

Example 00.1. Supongamos que una compañía quiere modelar la satisfacción de un usuario utilizando la siguiente asignación, si el usuario está satisfecho se le asignará el valor uno, de otro modo se le asignará el valor cero.

Notemos que es posible modelar la satisfacción promedio de todos los usuarios mediante un solo valor $\beta^* \in [0, 1]$, sin embargo algunas veces podría ser complicado conocer la información de todos los usuarios y por tanto se considerará solo una muestra de tamaño N , digamos $S = \{x_1, \dots, x_N\}$. En ese caso es posible definir una estadística mediante el promedio empírico $\frac{1}{N} \sum_{i \leq N} x_i$.

Exercise 00.2. Determinar el espacio de probabilidad del ejemplo anterior así como la estadística que genera el promedio empírico:

- *El espacio de probabilidad asociado es un espacio de bernoulli:*

$$(\{\omega_{sat}, \omega_{insat}\}, \mathbb{P})$$

tal que $\mathbb{P}(\omega_{sat}) = \frac{n_{sat}}{n}$, $\mathbb{P}(\omega_{insat}) = \frac{n_{insat}}{n}$ donde n_{sat}, n_{insat}, n son las cantidades de clientes satisfechos, insatisfechos y totales respectivamente.

- *La estadística se define como*

$$(\{\omega'_{sat}, \omega'_{insat}\}, \mathbb{P}')$$

tal que $\mathbb{P}'(\omega'_{sat}) = \frac{N_{sat}}{N}$, $\mathbb{P}'(\omega'_{insat}) = \frac{N_{insat}}{N}$ donde N_{sat}, N_{insat} son las cantidades de clientes satisfechos e insatisfechos.

En algunas ocasiones será necesario identificar el espacio en el que deseamos realizar nuestra búsqueda, es decir la familia de probabilidades que pueden ser nuestra estadística. En el caso del test de pearson del cual hablaremos en esta sección la familia de probabilidades es muy grande pues no estamos pensando en un test paramétrico.

Definition 00.2. Sea $\Omega = \{\omega_1, \dots, \omega_n\}$ un conjunto, denotamos por Δ_n el subconjunto de elementos $(p_1, \dots, p_n) \in (0, 1)^n$ tales que $\sum_{i \leq n} p_i = 1$

Supongamos que tenemos una población total con T -individuos que pertenecen a n -clases diferentes, desafortunadamente no tenemos acceso a todos los individuos, solo a una muestra de tamaño $N < T$. Queremos saber si la distribución de nuestra población total distribuida en n -clases

diferentes (a la que llamaremos \mathbb{P}) sigue o no una distribución fija \mathbb{P}' , es decir queremos saber si nuestra muestra puede ser una muestra independiente e idénticamente distribuida que siga la ley de probabilidad \mathbb{P}' o no. Para este objetivo los tests χ^2 de Pearson pueden ser muy útiles, antes de comenzar definiremos qué entendemos por un test estadístico.

En esta sección nos concentraremos en el caso cuando tenemos una hipótesis nula simple $H_0 : \mathbb{P} = \mathbb{P}'$ y una hipótesis compuesta $H_0 : \mathbb{P} = \mathbb{P}' \vee \mathbb{P}'' \vee \dots$

Definition 00.3. (Paradigma de tests de Neymann-Person para hipótesis nula compuesta) Sea $\Omega = \{\omega_1, \dots, \omega_n\}$ un conjunto, \mathbb{P} una distribución, $\mathbb{P}' \in \Delta_n$ una distribución fija con quien deseamos comparar a nuestra \mathbb{P} y $\delta \subseteq \Delta_n$ una familia de distribuciones sobre Ω con r parámetros distintos. Supongamos que S es una familia X_1, \dots, X_N de variables aleatorias independiente e idénticamente distribuidas con probabilidad \mathbb{P} . Un test de comparación entre \mathbb{P} y \mathbb{P}' con nivel de significancia $\alpha \in (0, 1)$ consta de la siguiente información:

- Una estadística (Ω, \mathbb{P}_S) de (Ω, \mathbb{P})
- (Región de rechazo) Un subconjunto de posibles distribuciones $R_{N,\alpha} \subseteq \Delta_n$ tal que:

$$\mathbb{P}'(\mathbb{P}_S \in R_{N,\alpha}) \leq \alpha$$

Esta cantidad representa el riesgo de rechazar la hipótesis $H_0 : \mathbb{P} = \mathbb{P}'$, otra forma de decirlo es que la cantidad de falsos negativos está acotada

por α .

Que satisface alguna de lo siguiente:

- (Consistencia) Para cada $\mathbb{P}'' \in \delta$,

$$\lim_{N \rightarrow \infty} \mathbb{P}''(\mathbb{P}_S \in R_{N,\alpha}) = 1$$

Estas cantidades representan la probabilidad de rechazar la hipótesis cuando es falsa.

Para definir el test χ^2 de Pearson es necesario definir la siguiente distribución.

Exercise 00.3. Sean X_1, \dots, X_r un variables aleatorias independientes e idénticamente distribuidas cuya ley de probabilidad es $N(0, 1)$, definimos la variable aleatoria

$$X_1^2 + \dots + X_r^2$$

Llamaremos a su ley de probabilidad asociada χ_r^2 la ley de pearson de r grados de libertad.

Definition 00.4. Sea X una variable aleatoria y \mathbb{P}_X ley de probabilidad asociada. Para cada valor $p \in (0, 1)$ definimos el quantil de orden p de la variable aleatoria X de la siguiente manera:

$$q_X(p) = \inf\{x \in \mathbb{R} : p = \mathbb{P}(X \leq x)\}$$

Definition 00.5. Sean \mathbb{P}, \mathbb{P}' dos leyes de probabilidad sobre $\Omega = \{\omega_1, \dots, \omega_n\}$

$(\mathbb{P}(\omega_i) = p_i, \mathbb{P}'(\omega_i) = q_i)$, definimos la pseudo-distancia χ^2 entre \mathbb{P} y \mathbb{P}' de la siguiente manera:

$$\chi^2(\mathbb{P}, \mathbb{P}') = \sum_{i \leq n} \left(\frac{(p_i - q_i)^2}{q_i} \right)$$

Exercise 00.4. Fijemos \mathbb{P}' una ley de probabilidad tal que $\mathbb{P}'(\omega_i) = q_i$. Convencernos de que si $S \in \Omega_N^N$ es un muestreo independiente e idénticamente distribuido sobre Ω , si definimos $\mathbb{P}_S(\omega_i) = \frac{O_i}{N}$ donde O_i es la cantidad de veces que aparece ω_i en S y $E_i = N \cdot q_i$, entonces:

$$\chi^2(\mathbb{P}_S, \mathbb{P}') = \sum_{i \leq n} \frac{(O_i - E_i)^2}{E_i}$$

A continuación resumimos el test χ^2 cuando tenemos hipótesis H_1 comprobadas.

Proposition 00.5. Si tenemos un muestreo de N variables aleatorias independientes e idénticamente distribuidas con probabilidad \mathbb{P} . Digamos que \mathbb{P}' es una probabilidad y que δ es un conjunto de probabilidades con r parámetros.

En este caso diremos que $H_0 : \mathbb{P} \in \mathbb{P}$ y $H_1 : \mathbb{P} \in \delta$, si definimos:

- Una estadística \mathbb{P}_S igual al ejemplo 00.4.

$$\blacksquare \quad R_{N,\alpha} = \{\mathbb{P}_S \in \Delta_n : N\chi^2(\mathbb{P}_S, \mathbb{P}') \geq q_{\chi^2_{n-1}}(1-\alpha)\}$$

Obtenemos un test de comparación entre \mathbb{P} y \mathbb{P}' con nivel de significancia α .

Demostración. Esto es un corolario no-inmediato del teorema límite central.

□

04 Test de Kolmogorov-Smirnov

En el test χ^2 de Pearson supusimos que nuestras distribuciones \mathbb{P}, \mathbb{P}' son sobre un conjunto $\{\omega_1, \dots, \omega_n\}$, esta vez no lo supondremos, sino que ambas son sobre todo el conjunto de los números reales \mathbb{R} , en particular supondremos que la distribución acumulativa de \mathbb{P}' la denotaremos por $F(x) = \mathbb{P}(-\infty, x)$.

Definition 00.1. Sea $S\{x_1, \dots, x_N\}$ un muestreo de N variables X_i independientes e idénticamente distribuidas que siguen una ley de probabilidad \mathbb{P} .

Definimos la estadística de Kolmogorov-Smirnov como la ley de probabilidad sobre \mathbb{R} definida de la siguiente manera:

$$\mathbb{P}_S(-\infty, x) = \frac{1}{N} \sum_{i \leq N} 1_{(-\infty, x)}(x_i)$$

Proposition 00.1. *Dadas las condiciones anteriores, notemos que la cantidad*

$\hat{F}_N(x) = \mathbb{P}_S(-\infty, x)$ *es una variable aleatoria que satisface lo siguiente:*

- $N \cdot \hat{F}_N(x) \sim \text{Bin}(N, F(x))$
- $\mathbb{E}[\hat{F}_N(x)] = F(x)$
- *En probabilidad se tiene que $\hat{F}_N(x) \rightarrow F(x)$ (esto es gracias a la desigualdad de Chebychev).*

Para definir un test estadístico correctamente será necesario utilizar una

modificación de la estadística de Kolmogorov-Smirnov, a partir de ahora esta segunda definición guardará el nombre de estadística de Kolmogorov-Smirnov.

Definition 00.2. Dadas las definiciones anteriores, definimos la estadística de Kolmogorov Smirnov de la siguiente manera:

$$D_{N,S} = \sqrt{N} \left(\sup_{x \in \mathbb{R}} |\hat{F}_N(x) - F(x)| \right)$$

- Theorem 00.2.**
- Si $H_0 : \mathbb{P} = \mathbb{P}'$ entonces la distribución de $D_{N,S}$ es independiente de \mathbb{P} .
 - (Consistencia) Si la hipótesis nula no es cierta: $H_1 : \mathbb{P} \neq \mathbb{P}'$ entonces para todo $m > 0$, $\mathbb{P}(D_{N,S} > m) \rightarrow 1$.

B O U R B A K I

COLEGIO DE MATEMÁTICAS

Bibliografía

- [1] DavidA. van Dyk. The Role of Statistics in the Discovery of a Higgs Boson. <http://dx.doi.org/10.1146/annurev-statistics-062713-085841>, 1:41–59, jan 2014.



escuela-bourbaki.com