DATA SCIENCE

Javier Lezama

Incluit

Mayo - Junio de 2019

Ingesta de Datos

Formatos de Datos

Formatos de Datos

* Tabulares: como una planilla, con filas y columnas.
    * Formatos de Archivos: CSV, TSV, XLS
    * Estructura de Datos: Dataframe
* Jerárquicos: con valores anidados dentro de otros valores.
    * Formatos de Archivos: JSON, XML
    * Estructura de Datos: Lista de Objetos
* Crudos: sin estructura específica
    * Formato de Archivos: TXT
    * Estructura de Datos: String

CSV - Comma Separated Values

* Archivos de texto delimitado que usa coma para separar valores.
* Cada línea es un registro con uno o más campos.
* No está formalmente especificado!

```
latitud,longitud,Nombre
-54.832543,-68.3712885,SAN SEBASTIAN ( USHUAIA )
-54.8249379,-68.3258626,AERO PUBLICO DE USHUAIA
-54.8096728,-68.3114748,PUERTO USHUAIA (PREFECTURA)
-54.8019121,-68.3029511,PUERTO USHUAIA
-51.6896359,-72.2993574,PASO LAURITA CASAS VIEJAS
-51.5866042,-72.3649779,PASO DOROTEA
-51.2544488,-72.2652242,PASO RIO DON GUILLERMO
-53.3229179,-68.6063227,PASO SAN SEBASTIAN
-53.78438,-67.7173342,TERMINAL RIO GRANDE
```

https://tools.ietf.org/html/rfc4180

Lectura de CSV

https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html

```
In [1]: data = 'col1,col2,col3\na,b,1\na,b,2\nc,d,3'
In [2]: pd.read_csv(StringIO(data))
Out[2]:
col1 col2  col3
0    a    b    1
1    a    b    2
2    c    d    3
```

Bases de Datos

Una base de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.
Clasificadas según
* Variabilidad de los datos: estáticas o dinámicas
* Contenidos: bibliográficas, texto completo, directorios, etc
* Modelo de administración: [no] transaccionales, [no] relacionales, distribuidas, etc

¿De qué hablamos cuando decimos que algo es probable?

La Teoría de Probabilidades estudia los llamados experimentos aleatorios.
Un experimento aleatorio tiene las siguientes características:
1- Se lo puede repetir bajo las mismas condiciones tantas veces como se desee.
2- No se puede predecir con exactitud el resultado de dicho experimento, pero se puede decir cuáles son los posibles resultados del mismo.

¿De qué hablamos cuando decimos que algo es probable?

3- A medida que el experimento se repite, los resultados individuales parecen ocurrir en forma caprichosa. Pero si el experimento se repite un gran número de veces, y registramos la proporción de veces que ocurre un determinado resultado, veremos que esa proporción tiende a estabilizarse en un valor determinado a medida que aumenta el número de veces que se repite el experimento.

No pretendamos que las cosas cambien si siempre hacemos lo mismo
Albert Einstein

¿Ejemplos?

a) tirar un dado y observar el número en la cara de arriba.
b) El pronóstico meteorológico.
En los experimentos no aleatorios o deterministas se puede predecir con exactitud el resultado del experimento, es decir, las condiciones en las que se verifica un experimento determinan el resultado del mismo

Definición axiomática

Sea ϵ un experimento aleatorio y S un espacio muestral asociado con ϵ. Con cada evento A asociamos un número real llamado probabilidad de A, que anotamos P(A), el cual satisface las siguientes propiedades básicas o axiomas

1- $0 \leq P(A) \leq 1$

2- $P(S) = 1$

3- Si A y B son eventos mutuamente excluyentes entonces $P(A \cup B) = P(A) + P(B)$

4- Si $A_1, A_2, \ldots, A_n, A_{n+1}, \ldots$ es una secuencia de eventos tales que

$A_i \cap A_j = \varnothing$ si $i \neq j$, entonces $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

¿Ejemplos?

A veces sucede que un experimento no es aleatorio estrictamente, pero resulta mucho más sencillo estudiarlo como si fuera aleatorio
¿Qué variables deberíamos conocer con anterioridad para predecir de qué lado cae la moneda?

Tipos de datos

* Numéricos (discretos y continuos)
* Categóricos
* Ordinales

Datos Cuantitativos

Los datos cuantitativos son datos que miden o calculan un algo para llegar a un punto en su investigación. Estos datos nos dicen a través de números una explicación para alguna tendencia o resultados de algún experimento. Con los datos cuantitativos, se puede hacer todo tipo de tareas de procesamiento de datos numéricos, tales como sumarlos, calcular promedios, o medir su variabilidad.

Datos Cuantitativos

Los datos discretos solo van a poder asumir un valor de una lista de números específicos.
Representan ítems que pueden ser contados; todos sus posibles valores pueden ser listados.
Suele ser relativamente fácil trabajar con este tipo de dato.
Los datos continuos representan mediciones; sus posibles valores no pueden ser contados y sólo pueden ser descritos usando intervalos en la recta de los números reales.

Datos Cualitativos o Categóricos

Si los datos nos dicen en cual de determinadas categorías no numéricas nuestros ítems van a caer, entonces estamos hablando de datos cualitativos o categóricos; ya que los mismos van a representar determinada cualidad que los ítems poseen

Datos Ordinales

Una categoría intermedia entre los dos tipos de datos anteriores, son los datos ordinales. En este tipo de datos, va a existir un orden significativo, vamos a poder clasificar un primero, segundo, tercero, etc. es decir, que podemos establecer un ranking para estos datos, el cual posiblemente luego tenga un rol importante en la etapa de análisis. Los datos se dividen en categorías, pero los números colocados en cada categoría tienen un significado..
Ej: Puntuación de estrellas

Nos acercamos a los datos

```
dataset = pandas.read_csv('../violencia-institucional-2018-01.csv', encoding='utf8')
```

```
dataset[:3]
```

	area	organismo_origen	via_acceso	year	provincia	contexto	contexto1	circunstancia	alojamiento	violencia_fisica	violencia_psiquica
0	DNPCVI	SECRETARIA DE DDHH	Telefónica	2017.0	Buenos Aires	Situaciones de Detención	Penal / Complejo Penitenciario PROVINCIAL	NaN	NaN	NaN	NaN
1	DNPCVI	SECRETARIA DE DDHH	Telefónica	2017.0	Buenos Aires	Situaciones de Detención	Penal / Complejo Penitenciario PROVINCIAL	NaN	Malas condiciones de alojamiento (higiene), Hu...	NaN	NaN
2	DNPCVI	SECRETARIA DE DDHH	Telefónica	2017.0	Buenos Aires	Situaciones de Detención	Penal / Complejo Penitenciario PROVINCIAL	NaN	Malas condiciones de alojamiento (higiene)	NaN	NaN

Datos categóricos - Ordinales?

Leemos el dataset

```
dataset = pandas.read_csv('../violencia-institucional-2018-01.csv', encoding='utf8')
```

```
dataset[:3]
```

	area	organismo_origen	via_acceso	year	provincia	contexto	contexto1	circunstancia	alojamiento	violencia_fisica	violencia_psiquica
0	DNPCVI	SECRETARIA DE DDHH	Telefónica	2017.0	Buenos Aires	Situaciones de Detención	Penal / Complejo Penitenciario PROVINCIAL	NaN	NaN	NaN	NaN
1	DNPCVI	SECRETARIA DE DDHH	Telefónica	2017.0	Buenos Aires	Situaciones de Detención	Penal / Complejo Penitenciario PROVINCIAL	NaN	Malas condiciones de alojamiento (higiene), Hu...	NaN	NaN
2	DNPCVI	SECRETARIA DE DDHH	Telefónica	2017.0	Buenos Aires	Situaciones de Detención	Penal / Complejo Penitenciario PROVINCIAL	NaN	Malas condiciones de alojamiento (higiene)	NaN	NaN

Datos numéricos - Continuos o discretos?

Leemos el dataset

```
In [41]: poblacion[:3]
```

Out[41]:

	Provincia	Población 2001	Población 2010	Variación absoluta	Variación relativa (%)
0	Ciudad de Buenos Aires	2.776.138	2.890.151	114.013	4,1
1	Buenos Aires	13.827.203	15.625.084	1.797.881	13,0
2	Catamarca	334.568	367.828	33.260	9,9

Datos categóricos

Datos numéricos

Discretos

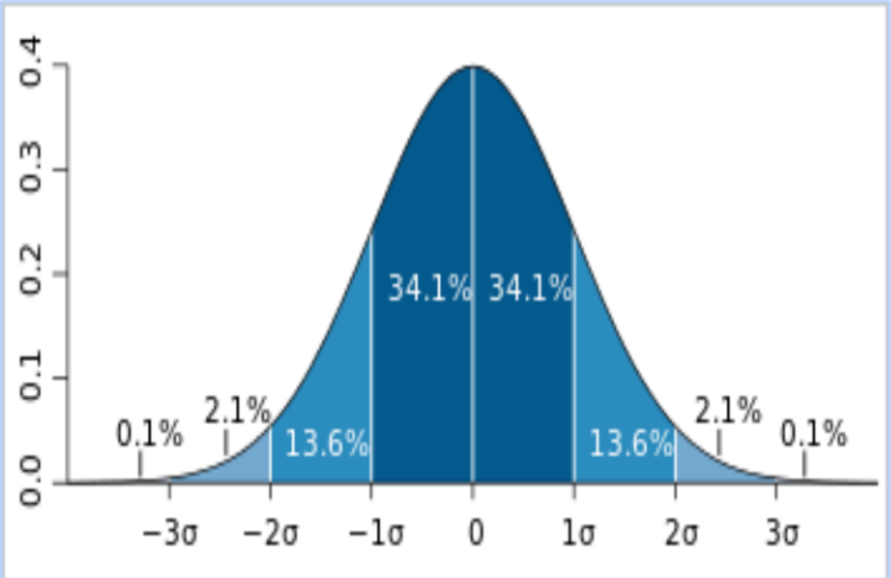
Datos numéricos

Continuos

Frecuencia de Distribución de Probabilidad

La distribución de probabilidad de una variable aleatoria es una función que asigna a cada suceso definido sobre la variable la probabilidad de que dicho suceso ocurra. La distribución de probabilidad está definida sobre el conjunto de todos los sucesos y cada uno de los sucesos es el rango de valores de la variable aleatoria.

Frecuencia de Distribución de Probabilidad



Probabilidad Condicional

Supongamos el experimento aleatorio de extraer al azar sin reemplazo dos bolillas de una urna que contiene 7 bolillas rojas y 3 blancas.
Consideramos los eventos
A: la primer bolilla extraída es blanca
B: la segunda bolilla extraída es blanca.

Probabilidad Condicional

Calculamos la p(A) = 3 / 10

calculamos p (B)... aunque ahora ya no es tan directo.

¿Que cambió?

Podemos calcular la probabilidad de B sabiendo que A ocurrió :es igual a 2/9, ya que si A ocurrió, entonces en la urna quedaron 9 bolillas de las cuales 2 son blancas. La probabilidad anterior la anotamos P (B / A) y se lee:

“probabilidad condicional de B dado A. Es decir P (B / A) = 2/9”

Teorema de la multiplicación

Si A y B son dos eventos entonces

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

si $P(A) \neq 0$

Análogamente

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

si $P(B) \neq 0$

Si ambos se cumplen se cumple el teorema de la multiplicación

Si A 1 , A 2 , A 3 son tres eventos entonces

$P(A_1 \cap A_2 \cap A_3) = P(A_1)\, P(A_2 / A_1)\, P(A_3 / A_1 \cap A_2)$

Se lee como la probabilidad que pase A1 y A2 y A3 es igual a la probabilidad que ocurra A1 por la probabilidad que ocurra A2 dado que ocurrió A1 por la probabilidad que ocurra A3 dado que ocurrieron A1 y A2

Teorema de la multiplicación

Esta regla es importante porque a menudo se desea obtener $P(A \cap B)$, en tanto que $P(B)$ y $P(A \mid B)$ pueden ser especificadas a partir de la descripción del problema.
La regla de multiplicación es más útil cuando los xperimentos se componen de varias etapas en sucesión. El evento condicionante B describe entonces el resultado de la primera etapa y A el resultado de la segunda, de modo que $P(A \mid B)$, condicionada en lo que ocurra primero, a menudo será conocida. La regla es fácil de ser ampliada a experimentos que implican más de dos etapas.

Independencia

→ Dados dos eventos A y B , puede ocurrir que $P(B/A)$ y $P(B)$ sean diferentes, eso significa saber que A ocurrió y modifica la probabilidad de ocurrencia de B

→ Entonces, dos eventos A y B son independientes si $P(B/A) = P(B)$, y son dependientes de otro modo

→ Notar que por el teorema de la multiplicación $P(A \cap B) = P(B/A)\,P(A)$ si $P(A) > 0$

→ Entonces A y B son independientes $\Rightarrow$

$$P(A \cap B) = P(B/A)\,P(A) = P(B)\,P(A)$$

Ejemplo de independencia

1- Las probabilidades de que tres hombres peguen en el blanco son, respectivamente, $\frac{1}{6}$, $\frac{1}{4}$, y $\frac{1}{3}$.

Cada uno dispara una vez al blanco.

a) ¿Cuál es la probabilidad de que exactamente uno de ellos pegue en el blanco?

b) Si solamente uno pega en el blanco, ¿cuál es la probabilidad de que sea el primer hombre?

<u>Solución:</u>

a) consideremos los eventos A_i: "el hombre i-ésimo pega en el blanco" $\quad i = 1, 2, 3$

$$P(A_1) = \frac{1}{6} \qquad P(A_2) = \frac{1}{4} \qquad P(A_3) = \frac{1}{3}$$

Sea el evento B: "exactamente un hombre pega en el blanco"

Entonces $B = \left(A_1^C \cap A_2^C \cap A_3\right) \cup \left(A_1^C \cap A_2 \cap A_3^C\right) \cup \left(A \cap A_2^C \cap A_3^C\right)$

Por lo tanto $P(B) = P\left(A_1^C \cap A_2^C \cap A_3\right) + P\left(A_1^C \cap A_2 \cap A_3^C\right) + P\left(A \cap A_2^C \cap A_3^C\right)$

Y por independencia $P(B) = P\left(A_1^C\right)P(A_2^C)P(A_3) + P\left(A_1^C\right)P(A_2)P)(A_3^C) + P\left(A_1\right)P(A_2^C)P)(A_3^C) =$

Ejemplo de independencia

$$=\left(1-\frac{1}{6}\right)\left(1-\frac{1}{4}\right)\frac{1}{3}+\left(1-\frac{1}{6}\right)\frac{1}{4}\left(1-\frac{1}{3}\right)+\frac{1}{6}\left(1-\frac{1}{4}\right)\left(1-\frac{1}{3}\right)=\frac{1}{12}+\frac{5}{36}+\frac{5}{24}=\frac{31}{72}$$

b) Se pide calcular $P(A_1/B)$

$$P(A_1/B)=\frac{P(A_1\cap B)}{P(B)}=\frac{P(A_1\cap A_2^C\cap A_3^C)}{P(B)}=\frac{\frac{1}{6}\left(1-\frac{1}{4}\right)\left(1-\frac{1}{3}\right)}{\frac{31}{72}}=\frac{6}{31}$$

Variables Aleatorias Continuas

Una variable aleatoria continua es una variable aleatoria con un conjunto de valores posibles (conocido como el rango) que es infinito y no se puede contar.

Sea X una v.a.. Decimos que es continua si existe una función no negativa f , definida sobre todos los reales $x \in (-\infty, \infty)$, tal que para cualquier conjunto B de números reales

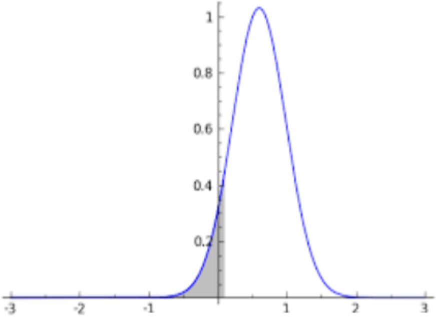
$$P(X \in B) = \int f(x)\,dx$$

O sea que la probabilidad de que X tome valores en B se obtiene al integrar la función f sobre el conjunto B. A la función f la llamamos función densidad de probabilidad (f.d.p.).

Variables aleatorias continuas importantes Distribución normal o gaussiana

Se llama distribución normal, distribución de Gauss, distribución gaussiana o distribución de Laplace-Gauss, a una de las distribuciones de probabilidad de variable continua que con más frecuencia estadística aparece.

Gráfica de una distribución normal



$$\text{fdp}=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^{2}}dt$$

Características

- Es una distribución que tiene forma de campana, es simétrica y puede tomar valores entre menos infinito y más infinito
- Media, mediana y moda son iguales
- Es simétrica

Los parámetros de la normal

Cuando μ varía la gráfica de la función se traslada, es un parámetro de posición.
Cuando σ aumenta, la gráfica se "achata", cuando σ disminuye la gráfica se hace más "puntiaguda", se dice que es un parámetro de escala.

¿Quienes son μ y σ?

- μ es la media
- σ es la desviación típica

Distribución exponencial

Cuando μ varía la gráfica de la función se traslada, es un parámetro de posición.
Cuando σ aumenta, la gráfica se "achata", cuando σ disminuye la gráfica se hace más "puntiaguda",
se dice que es un parámetro de escala.

¿Quienes son μ y σ?

- μ es la media
- σ es la desviación típica

¿Dudas? ¿Comentarios?

Bibliografia

http://diplodatos.famaf.unc.edu.ar/

Alberto Cairo. 2016. The Truthful Art: Data, Charts, and Maps for Communication (1st ed.). New Riders Publishing, Thousand Oaks, CA, USA.

Devore J. 2008. Probabilidad y Estadística para ingeniería y ciencias (7ma Edición) Cengage Learning.

Martín Gardner 2007 ¡Ajá! Paradojas que hacen pensar España, RBA Coleccionables, S.A.