Business words

Statistics → Data mining → predictive analysis → Data Science

Analysis vs Analytics

Analysis.

PAST

Estudio fraccionado de una información ya recolectada

How? Why?

Analytics

Exploracion de potenciales futuros. Identificando por ejemplo patrones

Tipos

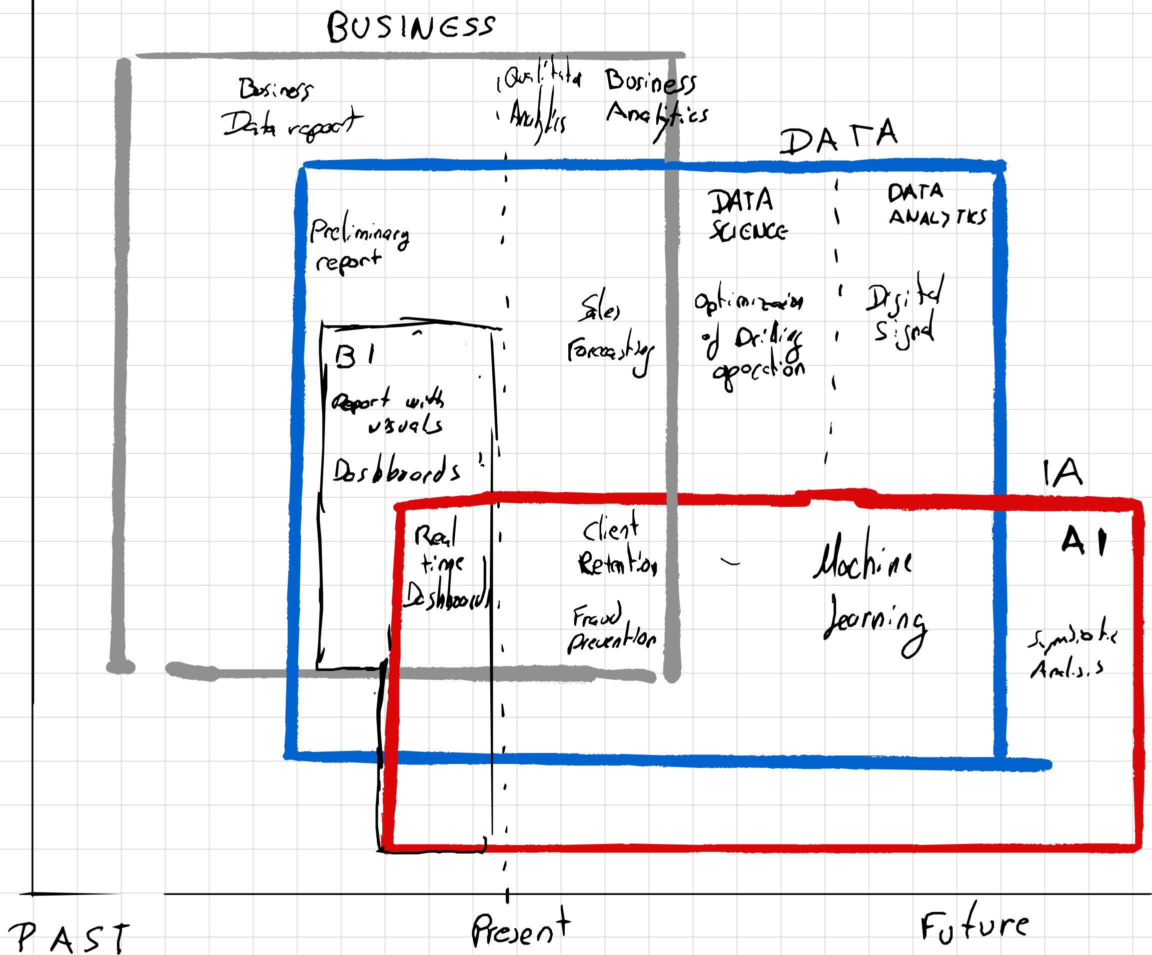
Qualitativo

Uso de la intuición y la experiencia

Quantitative

Uso de formulas y algoritmos

ESQUEMA TECNICISMOS - Analítica Avanzada



Business Intelligence (BI): Analisis y report de datos historicos de negocio

M1 - 1/ Probabilidad

A → Evento

P(A) → Probabilidad

$P(A) = \frac{\text{preferred (outcomes)}}{\text{all}}$

$\frac{\text{preferred}}{\text{all}} = \frac{\text{favourable}}{\text{sample space}}$

$P(A \text{ and } B) = P(A) \cdot P(B)$

Trial - intento

Experiment: conjunto de intentos

Experimental probability - probabilidad a base de pasado

Theorical probability - % teorica

- El complemento de un evento es todo lo que no es el evento, es decir la probabilidad restante

$A + A^c = \text{Sample space}$

$A + A' = 1$

$(A')' = A$

Combinatoria

- Permutaciones

Cantidad de maneras unicas de definir un set

Combinacion de de un set con n objetos

$$P_n = n \cdot (n-1) \cdot (n-2) \cdots 1 = n! \text{ (n factorial)}$$

- Factoriales

$$n! = 1 \cdot 2 \cdots \times n$$

$$3! = 1 \cdot 2 \cdot 3 = 6$$

- Los factoriales no tienen negativos
- $0! = 1$

Propiedades

$$n! = n \cdot (n-1)!$$

$$(n+1) = n! \cdot (n+1)$$

$$(n+k)! = n! \cdot (n+1) \cdots (n+k)$$

$$(n-k)! = \frac{n!}{(n-k+1) \cdot (n-k+2) \cdots (n-k+k)}$$

$\forall (n > k)$

$$\frac{n!}{k!} = (k+1) \cdot \ldots \cdot n$$

Combinatoria - Variaciones & Combinations

Variaciones con repeticion

Formula: $\overline{V}_p^n = n^p$

$n \to$ number of option
$p \to$ number of positions

Variaciones sin repetición

Formula $V_p^n = \frac{n!}{(n-p)!}$

n - num options
p - num positions

Combinations

Variaciones donde el orden no importa

$$C_p^n = \frac{n!}{p! \cdot (n-p)!}$$

Ej:

$$C_4^{10} = \frac{10!}{4! \cdot (10-4)!} = \frac{10!}{4 \cdot 6!} = \frac{1 \cdot \ldots \cdot 10}{1 \ldots 4 \quad 1 \ldots 6} = \frac{7 \cdot 8 \cdot 9 \cdot 10}{1 \cdot 2 \cdot 3 \cdot 4} = \frac{5040}{24} = 210$$

Combinatoria - Simetria

El punto simetrico es $n/2$

pick p-many of n = omit $n-p$

Conjunto de combinaciones

$$C_1 \cdot C_2 \cdot C_3 \cdots \cdot C_n$$

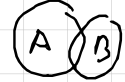
Basicamente calcula el numero de opciones para cada evento y multiplica

SIN REPE
$C = \frac{V}{P}$ $P_n = n!$ $V_p^n = \frac{n!}{(n-p)!}$
CON REPE
$\overline{V}_p^n = n^p$
$\overline{C}_p^n = \frac{(n+p-1)!}{p!\,(n-1)!}$
SIMETRIA
$C_p^n = C_{n-p}^n$

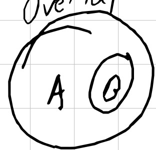
Sets - Espacios - Conjuntos

Diferentes

Intersección



Overlap



Intersect: Pueden ocurrir a la vez

Overlap: Solo pueden ocurrir B si ocurre A pero A puede ocurrir en B

INTERSECTIÓN: Conjunto perteneciente a ambos

- Si no tienen ningun elemento comun $A \cap B = \emptyset$
- Si tienen todos los elementos de B en comun (B es subset de A) $A \cap B = B$

UNIONS: Conjunto de todo A y B

$\boxed{A \cup B}$

Si $A \cap B = \emptyset \rightarrow A \cup B = A + B$

Si $A \cap = X \rightarrow A \cup B = (A + B) - X$

$A \cap B = B$

&

$A \cup B = A$

Mutually Exclusive Sets

No tienen ningun tipo de overlap

$A \cap B = \emptyset$ | $A \cup B = A + B$ |

Depencence and Independence Sets

$A \rightarrow Q\blacklozenge$

$B \rightarrow \blacklozenge$

$C \rightarrow Q$

$P(A|B) = \frac{1}{13}$ → Probabilidad de A en B

$P(A|C) = \frac{1}{4}$

Conditional Probability

Conditional Probability

$P(A) = P(A|B)$

Independientes → Si son independientes,

$P(A \cap B) = P(A) \cdot P(B)$ Solo si son independientes

Formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$ * Only true if $P(B) > 0$

El orden en $P(A|B)$ Importa

Formula Independientes: $P(A|B) = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$

LAW of Total Probability

$$A = B_1 \cup B_2 \cup \cdots \cup B_n$$

$$P(A) = [P(A|B_1) \cdot P(B_1)] + [P(A|B_2) \cdot P(B_2)] + \cdots [P(A|B_n) \cdot P(B_n)]$$

- ADITIVE LAW

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

- MULTIPLICATION RULE

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A|B) \cdot P(B) = \frac{P(A \cap B) \cdot \cancel{P(B)}}{\cancel{P(B)}}$$

$$\boxed{P(A|B) \cdot P(B) = P(A \cap B)}$$

Bayes' Rule - LAW - THEOREM

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(B|A) \cdot P(A)$$

$$\boxed{P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}}$$

Practical example

	Men	Women	
Full Time	886	1000	1886
Part Time	2	9	11
	888	1009	1897

Distributións

The possible values a variable can take and how frecuently they ocur

$Y \to$ Actual Outcome | $P(Y = y)$

$y \to$ Posible outcomes | $P(y)$

Ej: $Y \to$ marbles from bag

$y \to$ X number of marbles

Getting 5 marbles $P(Y=5)$ or $P(5)$

→ Mean μ

Averag value of a distribution

→ Variance σ^2

How spread the data is. How far values are from the mean

Population Data

Todos los Datos

mean: μ

variance: σ^2

Sample Data

Parte de los datos

anotaciones internas

mean de un sample: $\bar{X}$

variance .... : s^2

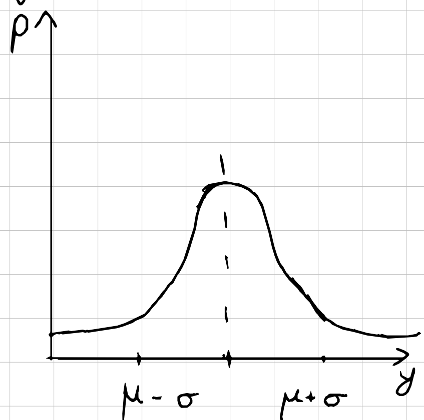
Standard Deviation

Square root of the variance $\sqrt{\sigma^2}$

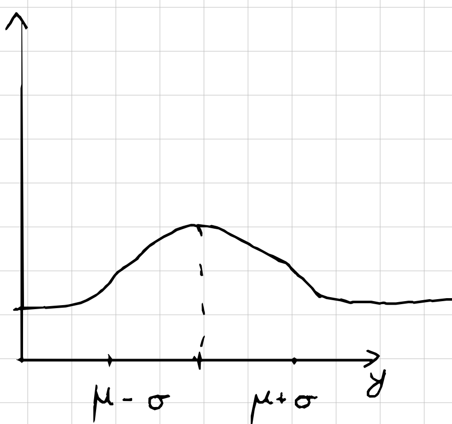
- ▹ Population: σ
- ▹ Sample: s

Se mide en las mismas unidades que la media

Ej



Mas congestión en el medio de la distribución = more data on it



Menos congestion

Los datos estaran más dispersos.

Hay relación constante entre mean y variance.

Expected value
↓

$$\sigma^2 = E\left((Y-\mu)^2\right) = E(Y^2) - \mu^2$$

TYPES OF DISTRIBUTIONS

▷ DISCRETE DISTRIBUTIONS

Finite number of outcomes

▷ CONTINUOUS DISTRIBUTIONS

Infinitely many outcomes

Anotaciones para expresar la distribución

$$\underbrace{X}_{\text{Variable}} \underbrace{\sim}_{\text{Tilde}} \underbrace{N}_{\text{Type of distribution}} \underbrace{(\mu, \sigma^2)}_{\text{Características (Pueden variar)}}$$

Discrete Distributions

▷ Distribuciones de casos finitos

▷ Uniform Distributions

▷ All outcomes tienen la misma probabilidad

Ej. flip coin ó sacar carta

Solo hay dos posibles respuestas $\begin{cases} \text{True} \\ \text{False} \end{cases}$

▷ Binomial Distribution

Solo dos outcomes por iteración pero puede haber varias iteraciones

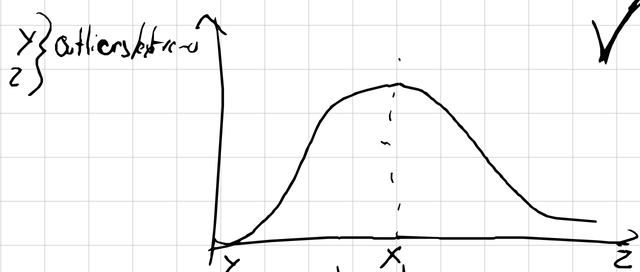
Ej. Lanzar una moneda 3 veces y calcular la probabilidad de que salga cara 2 veces seguidas

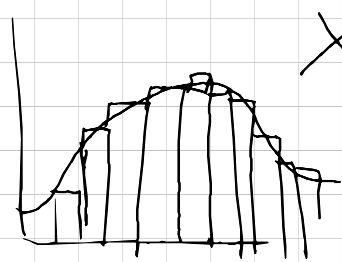
▷ Poisson Distribution

Como de inusual es la frecuencia de un evento en un intervalo específico.

CONTINUOUS DISTRIBUTIONS

Se representan como curvas en un grafico en lugar de las barras que estamos acostumbrados en las Discretas





- Normal Distribution

Suelen ser las que encontramos en la naturaleza

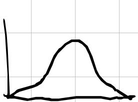
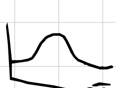
Ejemplo: el peso de un oso polar

Los outliers no suelen mostrarse en las distribuciones normales

- Student's-T Distribution

Tipo de distribución normal con datos limitados, solemos tratarla como una sample de una distribución normal

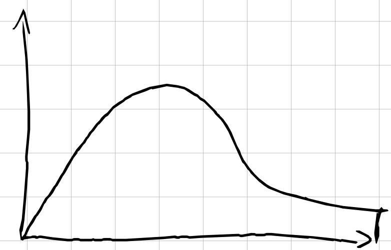
Los outliers de este tipo de distribución suelen estar elevados a diferencia de los de la normal

Normal  Student's-T 

CONTINUOUS DISTRIBUTIONS 2

Chi - Squared

- Asimetrica
- Solo valores no negativos
- Empieza en el 0 siempre
- No suele reflejar la realidad de eventos reales
- Se usa principalmente para testeo de Hypotesis



Exponencial Distribution

- Eventos que cambian rapidamente al comienzo
- Un ejemplo son los clicks en articulos, cuando son nuevos reciben mucha más atención

Logistic Distribution

- Forecast Analysis
- Usado para determinar el punto de corte de un outcome satisfactorio

Ej: Cuanta ventaja de oro hay que tener a minuto 10 en el lol para poder realizar predicciones de victoria

CARACTERISTICAS DISCRETE DISTRIBUTIONS

- Outcomes Finitos
- Se pueden mostrar con tablas, graficos o formulas
- Para esto solo tenemos que asignar la probabilidad de cada outcome
- Para calcular intervalos solo tenemos que sumar las probabilidades de todos los valores en ese rango.

$$P(Y \leq y) = P[Y < (y+1)]$$

Ejemplo:

$$P(\blacklozenge \leq 3) = P(\lozenge < 4)$$

Uniform Distribution

range of values

$U(a,b)$

Declaration of Uniform Distribution

Ejemplo declaración

$$X \sim U(3,7)$$

- Todos los resultados tienen la misma probabilidad

FLIP COIN

ROLL DICE

Ej dado: $P(1) = P(2) \cdots = P(6)$

- El expected value no nos da info al ser todos igual de probables

Así mismo el mean y la variance no tienen interpretación

Bernoulli Distribution