

Business words

Statistics → Data mining → predictive analysis → Data Science

Analysis vs Analytics

Analysis.

PAST | Estudio fraccionado de una información ya recolectada
How? Why?

Analytics

Exploration de potenciales futuros. Identificando por ejemplo patrones

Tipos

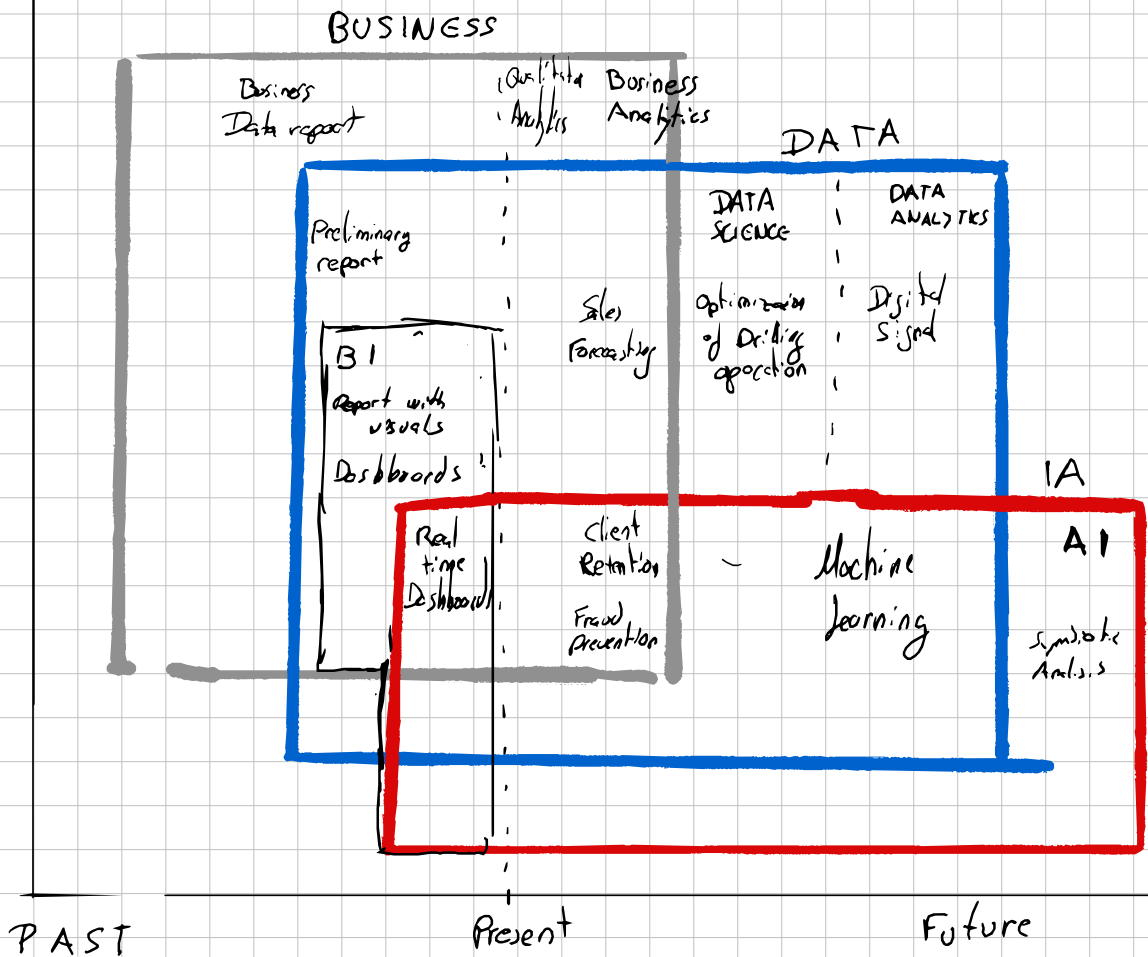
Qualitative

Uso de la intuición y la experiencia

Quantitative

Uso de fórmulas y algoritmos.

ESQUEMA TECNICO MOS # - Analítica Avanzada



Business Intelligence (BI): Analysis y report de datos históricos de negocio

$M = 1/\text{Probabilidad}$

$A \rightarrow \text{Evento}$

$P(A) \rightarrow \text{Probabilidad}$

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

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

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

Trial - intento

Experiment: conjunto de intentos,

Experimental probability - probabilidad a favor o favorable

Theoretical probability - % favorable

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

$$A + A^c = S = \text{sample space}$$

$$A + A' = 1$$

$$(A')' = A$$

Combinatoria

▷ Permutaciones

Cantidad de maneras únicas de definir un set
combinación de de un set con n objetos

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

▷ Factoriales

$$n! = 1 \cdot 2 \cdot \dots \cdot n$$

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

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

Propiedades

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

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

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

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

y $(n > k)$

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

Combinatoria - Variaciones & Combinations

▷ Variaciones con repetición

Formula: $\overline{V}_p^n = n^p$ $n \rightarrow$ number of options
 $p \rightarrow$ 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: $C_4^{10} = \frac{10!}{4! \cdot (10-4)!} = \frac{10!}{4! \cdot 6!} = \frac{1 \cdot \dots \cdot 10}{1 \cdot 4 \cdot 1 \cdot 5} = \frac{7 \cdot 8 \cdot 9 \cdot 10}{1 \cdot 2 \cdot 3 \cdot 4} = \frac{5040}{24} = 210$

Combinatoria - Simetria

El punto simétrico es $n/2$
pick p -many of n = omit $n-p$

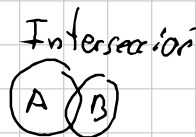
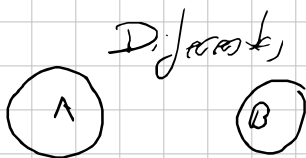
Conjunto de combinaciones

$$C_1 \cdot C_2 \cdot C_3 \dots \cdot C_n$$

Basicamente calcula el número 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 - Conjunto



Intersect: Pueden ocurrir a la vez

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

INTERSECCIÓN: Conjunto perteneciente a $a-b$

• Si no tienen ningún elemento común $A \cap B = \emptyset$

• Si tienen todos los elementos de B en común

(B es subset de A) $A \cap B = B$

UNIONS: Conjunto de todo A y B

$A \cup B$

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

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

$A \cap B = B$

+

$A \cup B = A$

Mutually Exclusive Sets

No tienen ningun tipo de overlap

$$\underline{A \cap B = \emptyset} \mid \underline{A \cup B = A + B} \mid \underline{\quad}$$

Dependence and Independence Sets

$$A \rightarrow Q \blacklozenge$$

$$P(A|B) = \frac{1}{23} \rightarrow \text{Probabilidad de A en B}$$

$$B \rightarrow \blacklozenge$$

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

Conditional Probability

$$C \rightarrow Q$$

Conditional Probability

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

Independientes \rightarrow Si son independientes

$$\underline{P(A \cap B) = P(A) \cdot P(B)} \quad \text{Solo si son independientes}$$

$$\text{Formula: } P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \left. \vphantom{\frac{P(A \cap B)}{P(B)}} \right\} \times \text{ only true if } P(B) > 0$$

Formula

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

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

LAW of Total Probability

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

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

- ADDITIVE 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

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

Practical example

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

Distributions

The possible values a variable can take and how frequently they occur

$Y \rightarrow$ Actual Outcome $\left| P(Y=y)$
 $y \rightarrow$ Possible outcomes $\left| P(y)$

Ej: $Y \rightarrow$ marbles from bag
 $y \rightarrow$ X number of marbles

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

\rightarrow Mean μ
Average value of a distribution

\rightarrow 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

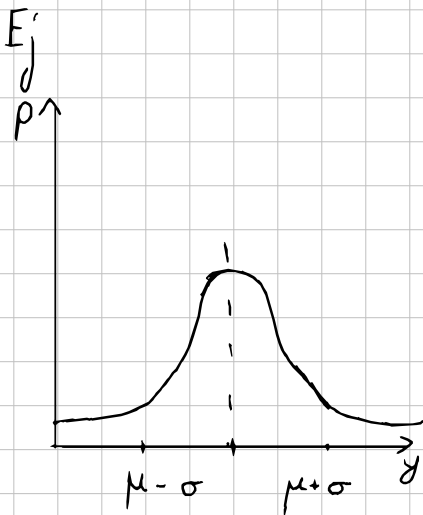
Sample Data

$\left\{ \begin{array}{l} \text{Parte de los datos} \\ \text{notación: } \bar{x} \\ \text{mean de un sample: } \bar{x} \\ \text{variance: } s^2 \end{array} \right.$

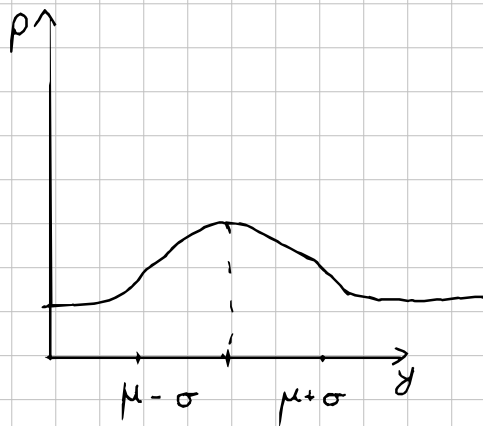
Standard Deviation

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

- ▷ Population: σ | Se mide en la misma unidades
- ▷ Sample: s | que la media



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



Menos congestión
Los datos estarán más
dispersos.

Hay relación constante entre mean &
variance.

Expected value \downarrow

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

TYPES OF DISTRIBUTIONS

▷ DISCRETE DISTRIBUTIONS

Finite number of outcomes

▷ CONTINUOUS DISTRIBUTIONS

Infinitely many outcomes

Notaciones para expresar la distribución

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

Discrete Distributions

▷ Distribuciones de cosas finitas

— ▷ Uniform Distributions

▷ All outcomes tienen la misma probabilidad

Ej. flip coin o 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

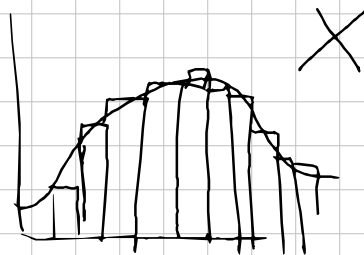
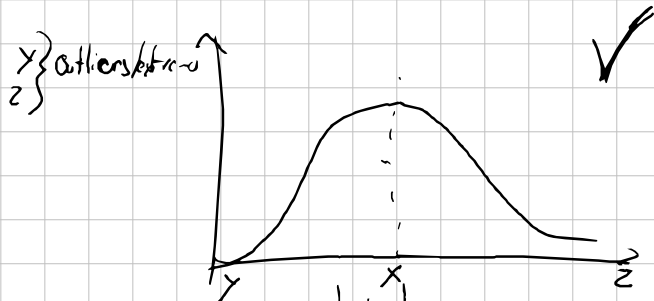
Ej. Lanzar una moneda 3 veces y calcular la probabilidad de que salga cara 2 veces seguidas

— ▷ Poisson Distribution

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

CONTINUOUS DISTRIBUTIONS

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



Normal Distribution

Suelen ser los que encontramos en la naturaleza

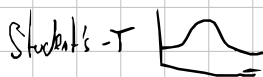
Ejemplo: el peso de un oso polar

los outliers no suelen mostrarse en las distribuciones normales

Student's -T Distribución

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



CONTINUOUS DISTRIBUTIONS 2

▷ Chi - Squared

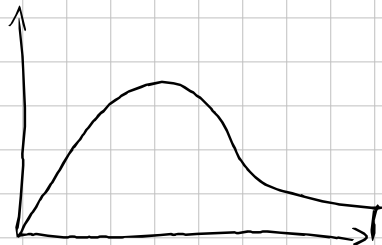
▷ Asimétrica

▷ Solo valores no negativos

▷ Empieza en el 0 siempre

▷ No suele reflejar la realidad de eventos reales

▷ Se usa principalmente para testeo de hipótesis



▷ Exponencial Distribución

▷ Eventos que cambian rápidamente al comienzo

▷ Un ejemplo son los clicks en artículos, 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 1º para poder realizar predicciones de victoria

CARACTERÍSTICAS DISCRETE DISTRIBUTIONS

- Outcomes Finitos
- Se pueden mostrar con tabla, gráficos o fórmulas
- 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:

$$P(\diamond \leq 3) = P(\diamond < 4)$$

Uniform Distribution

range of values
 $U(a, b)$
↑
Declaration
of
Uniform Distribution

Ejemplo declaración

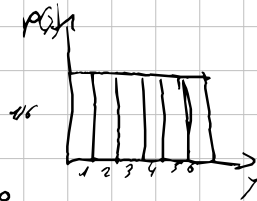
$$X \sim U(3, 7)$$

Todo los resultados tienen la misma probabilidad

FLIP COIN

ROLL DICE

Ej dato: $P(1) = P(2) = \dots = P(6)$



DE / expected value no no de info
al ser todos igual de Probables

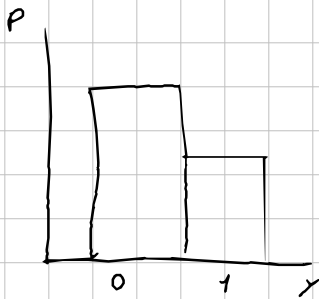
Así mismo el mean y la variance no tienen interpretación

Bernoulli Distribution

$$X \sim \text{Bern}(p)$$

▷ 1 trial and 2 possible outcomes

FLIP COIN | TRUE or FALSE test



$E(X)$ will be $\begin{cases} p \\ 1-p \end{cases}$

$$E(X) = p(1-p)$$

Binomial Distribution

Binomial Distribution

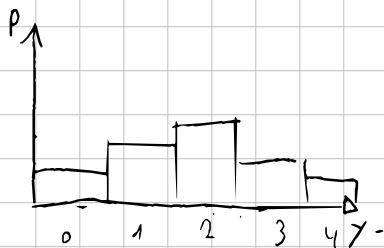
$$X \sim B \begin{matrix} \text{n trials} \\ \downarrow \\ (n, p) \\ \text{probability each} \end{matrix}$$

* Una Bern(p) se puede expresar también como una Binomial de un solo intento $B(1, p)$

Conjunto de trials E(Bern)

▷ E(Bern) - que usbr esperar de un solo intento

▷ E(Binomial Event) - el número de veces que experimentas para un caso concreto



El gráfico de una distribución Binomial representa el número de veces que obtenemos el resultado deseado

$$P(\text{desired}) = p$$

$y = \text{intentos deseados}$
 $n = \text{total intentos}$

$p = \text{prob. de solo intento}$

$$P(\text{alternative}) = 1 - p$$

$$P(y) = \binom{n}{y} \cdot p^y \cdot (1-p)^{n-y}$$

Ejemplo: calcular la probabilidad de que

n de los m intentos sean favorables
Incremento de un stock 5 días seguidos

$$C_3^5 = \frac{5!}{3! \cdot 2!} = \frac{5 \cdot 4 \cdot 3}{1 \cdot 2} = \frac{20}{2} = 10$$

$$P(4) = 0.6$$

$$P(4) = 0.4$$

$$P(3) = \binom{5}{3} \cdot 0.6^3 \cdot (1-0.6)^2 = 10 \cdot 0.216 \cdot 0.16 = 0.3456$$

Binomial Distribution 2

EXPECTED VALUE

$$E(X) = x_0 \cdot p(x_0) + \dots + x_n \cdot p(x_n)$$
$$Y \sim B(n, p)$$
$$E(Y) = p \cdot n$$

VARIANCE

$$\sigma^2 = E(y^2) - E(y)^2$$
$$= n \cdot p \cdot (1-p)$$
$$\sigma^2 = 1.2 \quad \sigma \approx 1.1$$

Example anterior (market $\frac{5}{3}$)

$$5 \cdot 0.6 \cdot 0.4 = 1.2$$

Poisson Distribution

$$Y \sim P_0(\lambda)$$

Frecuencia con la que un evento sucede en un intervalo específico.

Ej: un evento sucede 3 veces en 10 segundos

Posibilidad de 8 en 20?

Ejemplo: estudiante pasa de preguntar 4 a 7 preguntas en 1 día

$$\lambda = 4$$

Intervalo = 1 día

$$y = 7$$

e = euler number ≈ 2.72

$$x^{-n} = \frac{1}{x^n}$$

$$P(y) = \frac{\lambda^y \cdot e^{-\lambda}}{y!}$$

$$P(7) = \frac{4^7 \cdot e^{-4}}{7!} = \frac{16384 \cdot 0.0183}{5040} \approx 0.06$$

FORMULAS

$$E(y) = \lambda$$

$$P(y) = \frac{\lambda^y \cdot e^{-\lambda}}{y!}$$

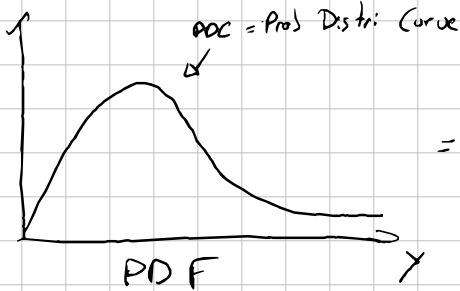
$$\sigma^2 = \lambda$$

$$\mu = \lambda$$

Características Distribuciones Continuas

Se representan con un grafico llamado

PDF (Probability Density Function)

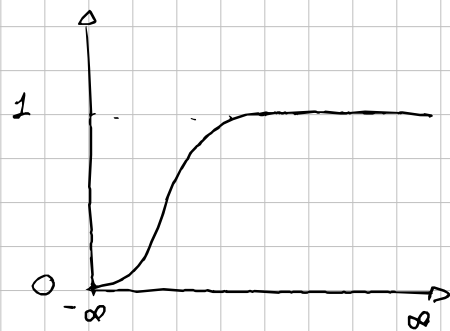


$$= f(y), y \in \text{Sample Space}$$
$$> 0$$

$$P(x) = \frac{1}{\infty} \approx 0$$

Sample Space

Cumulative Distribution function (CDF) = $F(y)$



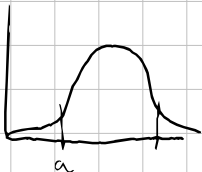
Probabilidad de un
valor random y
de ser $\leq Y$

$$F(-\infty) = 0$$

$$F(\infty) = 1$$

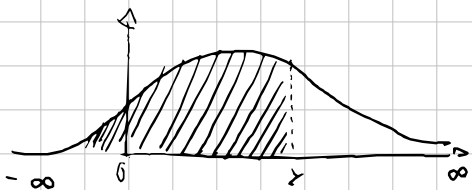
Integral

$$p(b > x > a) = \int_a^b p(x) dx$$



Ejemplo probabilidad intervalo

Probabilidad $(-\infty; y)$



$$\int_{-\infty}^y p(y) dy = F(y)$$

$$\begin{array}{ccc} p(y) & \xrightarrow{\text{Integral}} & F(y) \\ \text{PDF} & \xleftarrow{\text{Derivada}} & \text{CDF} \end{array}$$

$$p(y) = F(y) \frac{d}{dy}$$

Primera derivada de $F(y)$ con respecto a y

⊖

Normalmente cuando trabajamos con este tipo de distribuciones, solo nos daran sus PDF $f(y)$

Para poder dibujar el grafico necesitamos su Expected Value y Varianza

$$E(y) = \text{Integral} = \int_{-\infty}^{\infty} y p(y) dy$$

$$\int_{-\infty}^{\infty} \underbrace{y}_{\text{var}} \underbrace{p(y) dy}_{\text{diferencial sobre } y}$$

Product of any "y" and the $p(y)$ [PDF overall value] con respecto a y

El diferencial sobre y (dy) indica que la integraci3n se realiza con respecto a y

Distribuciones Continuas

Variance
Varianz

$$\text{Var}(y) = E(y^2) - E(y)^2$$

Normal Distributions

$$X \sim N(\mu, \sigma^2)$$

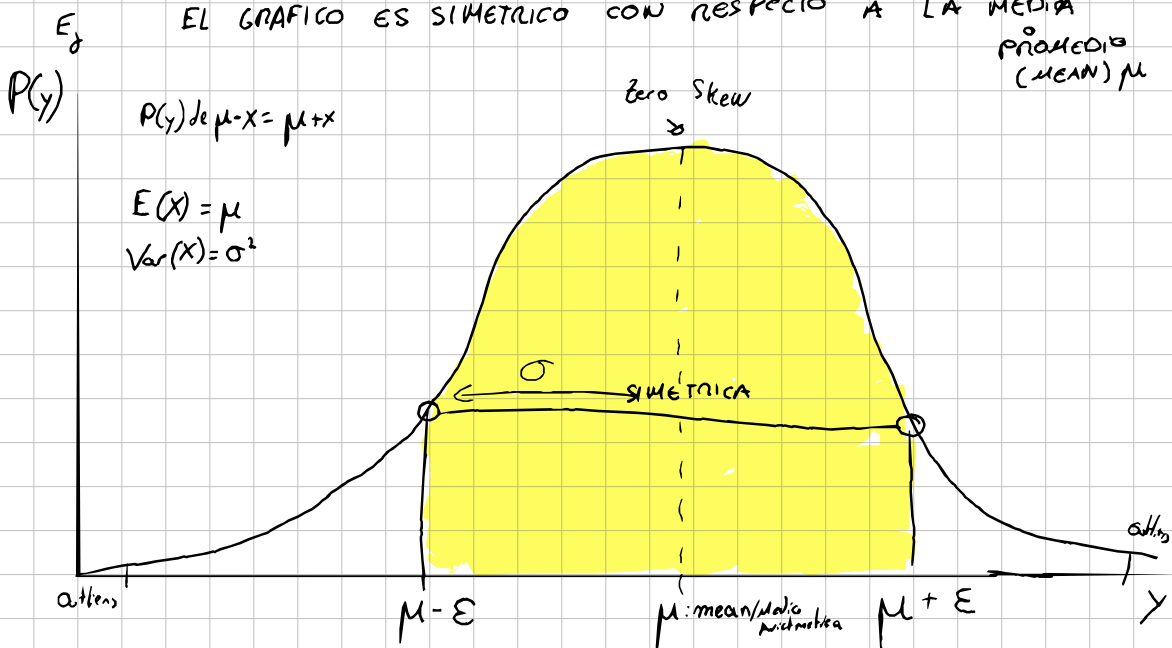
Normalmente
 μ y σ^2 usaran
 valores numéricos

Outliers: extremos excepcionales de los datos (inferiores y superiores)

Los gráficos de las distribuciones normales tienen forma de campana

EL GRÁFICO ES SIMÉTRICO CON RESPECTO A LA MEDIA

PROMEDIO
 (MEAN) μ



68, 95, 99.7 Law

- ▷ 68% of the elements between $\mu - \sigma$ y $\mu + \sigma$
- ▷ 95% $\mu - 2\sigma$ & $\mu + 2\sigma$
- ▷ 99.7% $\mu - 3\sigma$ & $\mu + 3\sigma$

Es importante tener
 volumen de datos por
 usar una distribución
 normal o se corre
 el riesgo de coger
 outliers que comprometen
 los datos. Si no hay
 muchos datos evita
 las distribuciones
 normales

Standardizing - Normal Distribución

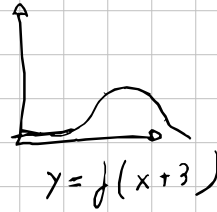
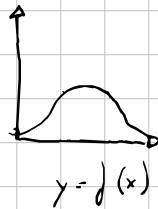
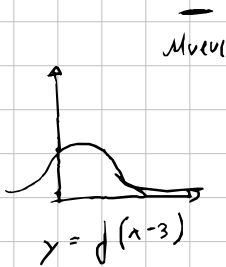
Transformación. Forma de alterar todo los elementos de una distribución para obtener otra distribución

En las distribuciones normales podemos usar
 $+$, $-$, \times , \div

sin alterar el tipo de distribución

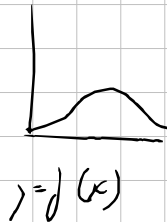
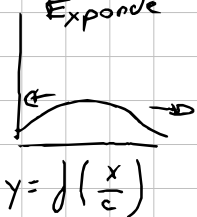
$$X \sim N(\mu, \sigma^2) \rightarrow X+3 \sim N(\mu, \sigma^2)$$

Seguira
siendo normal
+ $\mu + 3$



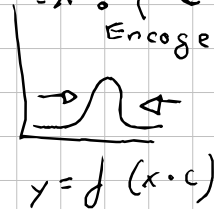
$$\begin{array}{l} \div X/(c > 1) \\ \cdot X \cdot (c < 1) \end{array}$$

Exponde



$$\begin{array}{l} \times (c > 1) \\ \div (c < 1) > 0 \end{array}$$

Encoge



Standardización Normales

Transformación especial en la que provocamos

$$E(x) = 0 \quad \text{Var}(x) = 1$$

La Distribución obtenida se llama

"Standard Normal Distribution"

A una distribución normal estandarizada se

le puede crear una tabla con los valores más comunes

que se llama CDF Table o Z-score table

Pasos estandarización

1. μ a 0 $\rightarrow y = f(x - \mu)$ con sumas o restas

2. dividir por σ para provocar $\sigma = 1$

$$y = f\left(\frac{x - \mu}{\sigma}\right)$$

ya estandarizada

$$Z \sim N(0, 1)$$

$$Y \sim N(\mu, \sigma^2)$$

$$Z = \frac{Y - \mu}{\sigma}$$

Ej: comparar valores

$$y = \mu + 2.3\sigma$$

$$z = 2.3$$

\downarrow
+0
Z

Student's T Distribution

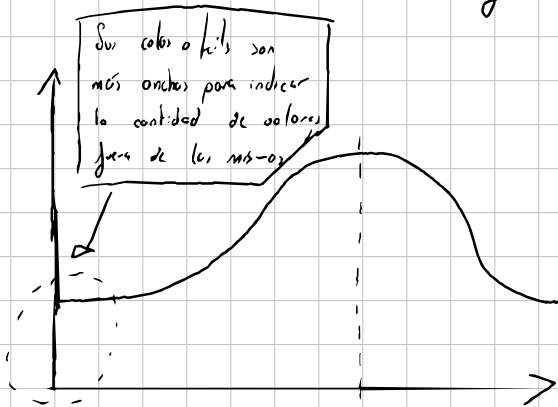
$$Y \sim t(k) \quad k = \text{degrees of freedom}$$

Small sample size approximation of normal distribution

Characteristics + sufficient data = Normal distribution

Characteristics + ~~sufficient data~~ = Student's T distribution

▷ Su curva también tiene forma de campana



$$if \quad k > 2 \quad \}$$

$$E(Y) = \mu$$

$$Var(Y) = \frac{s^2 \cdot k}{k-2}$$

▷ Frequently used when conducting statistical analysis

▷ Hypothesis testing with limited data
◦ Having a critical table (T-table)

$$\text{Formula} \rightarrow t_{n-1, \alpha} = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

n : size of sample
 α : confidence level

A partir de 50 se-pts
se suele usar la

z-table en lugar de la

T-table

Calculo umbrales de confianza

con T. Con variance unknown

sample mean

1. Calculo standard error: $\frac{s}{\sqrt{n}}$

2. Formula calculo intervalo sin conocimiento de variance

$$= \bar{x} \pm t_{n-1, \alpha/2} \cdot \frac{s}{\sqrt{n}}$$

2.1 Si suposieramos el population variance seria

$$= \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

n: sample size

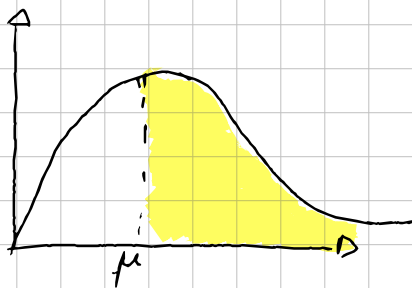
α : confidence level

CHI-SQUARED DISTRIBUTION

$$Y \sim X^2(k) \quad k = \text{degrees of freedom}$$

Aunque hay pocos casos de uso, sus usos principales, son:

- ▷ Testeo de hipótesis
- ▷ Computing confidence intervals
- ▷ "Goodness of fit" of categorical values



▷ Not symmetric!

▷ $Y \sim t(k) \rightarrow Y^2 \sim X^2(k)$
Es básicamente elevar students t al cuadrado

$$X \sim X^2(k) \Rightarrow \sqrt{X} \sim t(k)$$

$$\triangleright E(X) = k$$

$$\text{Var}(X) = 2k$$

Exponencial Distribution

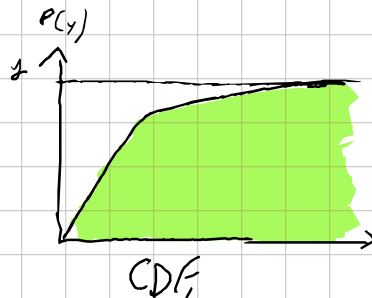
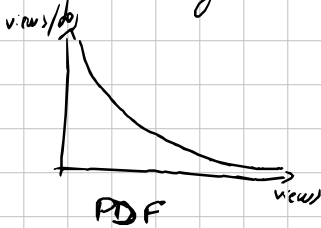
λ = scalar / scale / rate parameter

$$X \sim \text{Exp}(\lambda)$$

▷ Start high

▷ Comienza a descender hasta volverse plano

▷ Un ejemplo serían los views de un nuevo artículo o video de youtube



▷ La curva (PDF) parece un *boomarang* | ▷ La CDF es un *boomarang* invertido
en el 1 es donde se queda plana

λ o rate parameter:

▷ Nos indica lo rápido que llega la curva al punto de allanamiento

▷ Como de repartido está el gráfico

$$E(Y) = \frac{1}{\lambda} \quad \text{Var}(Y) = \frac{1}{\lambda^2}$$

No hay tabla de valores conocidos

TRANSFORMACIONES 2

Por ejemplo cuando trabajamos con distribuciones Exponenciales nos puede interesar transformar a normal para trabajar con ellas

$$\left. \begin{array}{l} Y \sim \text{Exp}(\lambda) \\ X = \ln(Y) \end{array} \right\} \Rightarrow X \sim N(\mu, \sigma^2)$$

LOGISTIC DISTRIBUTIÓN

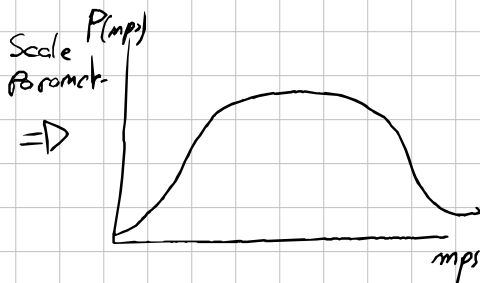
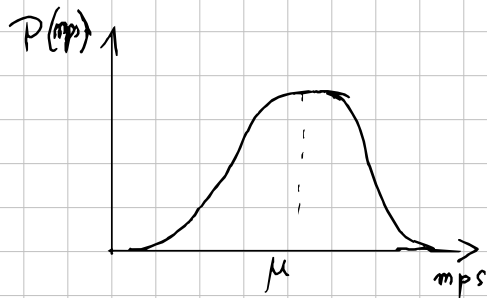
Logistic (μ, S)
 μ = Location
 S = scale parameter

▷ Usado para saber como variables externas continúan afectando a un output binario (2 valores)

Usado por ejemplo para predecir resultados de victoria o derrota en eventos deportivos.

Ej: Impacto en la velocidad y precisión

+ Velocidad significa + probabilidad de punto pero así mismo menos precisión. Por lo que buscamos la optima

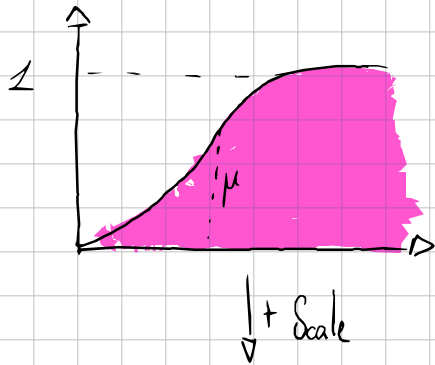


▷ Definida principalmente por su media (mean) y su escalar (Scale parameter)

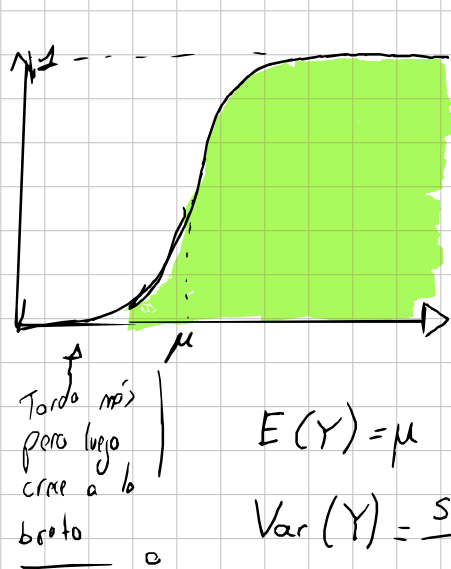
▷ μ : optimal value

▷ S : el "margen de fallo"

CDF de Logistic Distribution



- Curva que empieza lento pero que sube rápidamente hasta aplanarse en el valor 1
- Cuando llegamos al valor de la media el valor comienza un crecimiento drástico



- A más pequeño sea el scalar más tarda en comenzar a crecer de manera drástico. Pero así mismo alcanza 1 más rápido, es decir, obtiene una pendiente más pronunciada

$$E(Y) = \mu$$

$$\text{Var}(Y) = \frac{\pi^2}{3}$$

Ejemplo Práctica de Estadística

Desarrollo FIFA

▷ Balanceado

$n_{\text{jugadores buenos}} = n_{\text{jugadores malos}}$

▷ Divertido

Probability in Data science

