Herramientas Estadísticas y Forecast

Clase 2 – Probabilidad y distribuciones

DATA SCIENCE UC
FACULTAD DE MATEMÁTICAS

FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

Contenido del curso

1. Introducción a la estadística y análisis descriptivo
2. Análisis descriptivo y gráfico
3. Probabilidad y Distribuciones
4. Muestreo
5. Inferencia estadística: Pruebas de hipótesis
6. Taller Práctico de inferencia
7. Introducción a los modelos estadísticos
8. Modelos predictivos I: Modelos de regresión lineal.
9. Modelos predictivos II: Regresión logística y otros modelos.
10. Modelos de Forecasting I.
11. Modelos de Forecasting II
12. Taller de Forecast

Conceptos de probabilidad

A la estadística descriptiva le concierne el resumen de datos recogido de eventos pasados.

Ahora presentaremos la segunda faceta de la estadística, a saber, el cálculo de la probabilidad de que algo ocurra en el futuro.

Esta faceta de la estadística recibe el nombre de **inferencia estadística** o estadística inferencial.

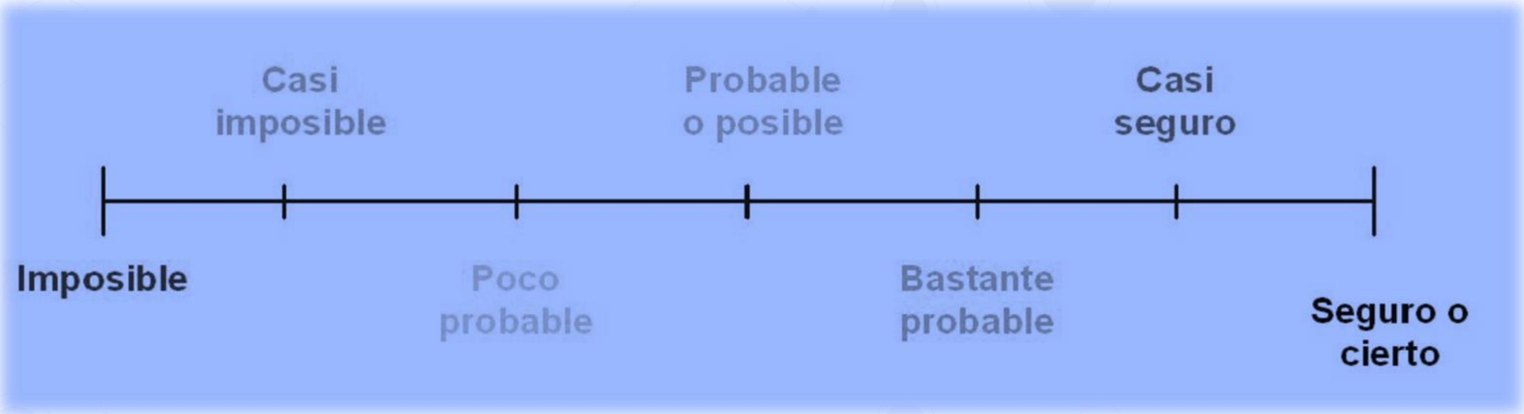
Conceptos de probabilidad

La inferencia estadística se relaciona con las conclusiones relacionadas con una población sobre la base de una muestra que se toma de ella.

Dada la incertidumbre existente en la toma de decisiones , es importante que se evalúen científicamente todos los riesgos implicados. La **teoría de la probabilidad** resulta útil para hace esta evaluación.

Conceptos de probabilidad

Se definirá **PROBABILIDAD** como una función numérica de ciertos eventos de algún experimento, tal que reportan valores entre 0 y 1 y representarán las posibilidades relativas que ocurra dicho evento.



Conceptos de probabilidad

La probabilidad tendrá sentido en experimentos que se consideren **aleatorios**, es decir, con una colección de resultados posibles, conocidos, pero sin certeza de la ocurrencia de ellos.

Por ejemplo, lanzar un dado, conocer si un producto al azar está defectuoso o no, el número de empleados que son necesarios para cierta operación, el monto total de las ventas al finalizar el mes o la rentabilidad mensual de los fondos de pensiones.

Conceptos de probabilidad

CLÁSICA

PROBABILIDAD

EMPÍRICA

SUBJETIVA

$$P(A) = \frac{\text{CASOS FAVORABLES}}{\text{CASOS TOTALES}}$$

La **Ley de los grandes** números indica que, luego de una gran cantidad de intentos, la probabilidad empírica se aproxima a la probabilidad clásica, y esta a la real.

Conceptos de probabilidad

GRUPOS DE EDAD	Hombres	Mujeres
0	0,005135	0,004381
1-4	0,000048	0,000040
5-9	0,000034	0,000022
10-14	0,000036	0,000033
15-19	0,000166	0,000064
20-24	0,000405	0,000089
25-29	0,000539	0,000090
30-34	0,000604	0,000087
35-39	0,000748	0,000146
40-44	0,001038	0,000311
45-49	0,001539	0,000746
50-54	0,002529	0,001307
55-59	0,004303	0,002288
60-64	0,007608	0,003985
65-69	0,013345	0,007077
70-74	0,022362	0,012250
75-79	0,041585	0,026313
80+	0,083178	0,066629

FUENTE: INE. Defunciones de 1998 – 2010, excluyendo 2007 y sus respectivas estimaciones y proyecciones de población

Las compañías de seguros de vida confían en datos empíricos para determinar la aceptabilidad de un solicitante, así como la prima. Las tablas de mortalidad incluyen una lista de las posibilidades de que una persona de determinada edad fallezca en el siguiente año.

Técnicas de probabilidad

Entre las herramientas más clásicas del cálculo de probabilidad, las primeras tienen relación con el **conteo** de datos.

La **regla de la multiplicación** indica que si un experimento se puede describir como una secuencia de ***k*** experimentos, con ***n1, n2,…, nk*** resultados posibles, respectivamente, entonces el número total de resultados posibles del experimento es ***n1•n2•…•nk***

Técnicas de probabilidad

El número de maneras de seleccionar ***k*** elementos de un conjunto de ***n*** distintos, sin importar el orden y sin reemplazo, se asocia a la **combinatoria**, y se calcula como:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Excel: COMBINAT(n,k)
R: choose(n,k)

El número de maneras de seleccionar ***k*** elementos de un conjunto de ***n*** distintos, cuando si importa el orden y es sin reemplazo, se asocia a la **permutación**, y se calcula como:

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

Excel: PERMUTACIONES(n,k)
R: factorial(n)/factorial(n-k)

Técnicas de probabilidad

¿Cuál es la probabilidad de ganar el Kino?

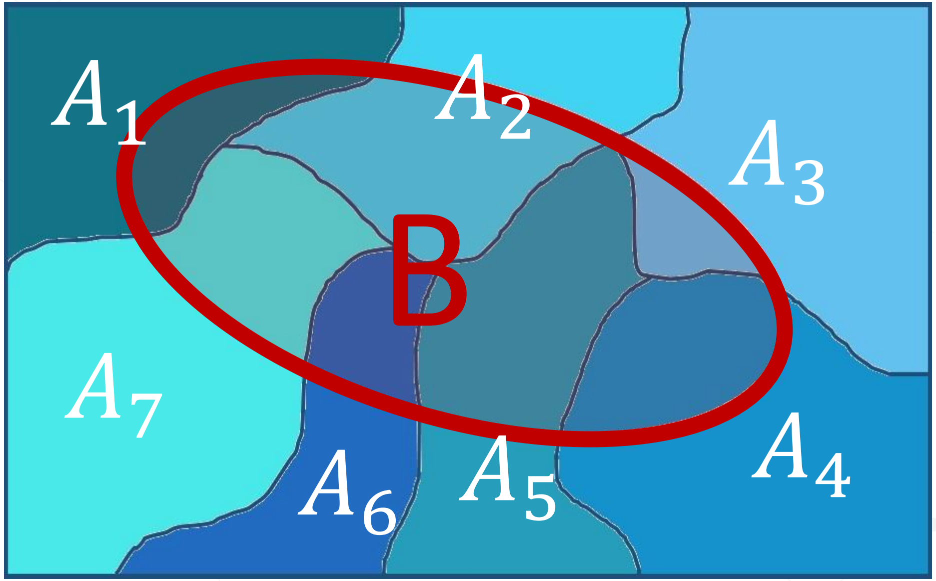
¿Cuál es la probabilidad que una patente (XXXXNN) tenga los dos últimos números iguales?

Técnicas de probabilidad

Una **probabilidad condicional** está pensada en cuando los resultados posibles de un experimento están acotados por la ocurrencia de otro evento.

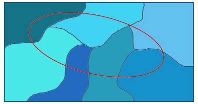
Además, se puede determinar la probabilidad de un evento a través de sumas de probabilidades condicionales, ponderadas por el evento condicionante, a lo que se le conoce como **Teorema de Probabilidades Totales**, y se puede conocer la probabilidad condicionada de forma inversa, a lo que se le denomina **Teorema de Bayes**.

Técnicas de probabilidad



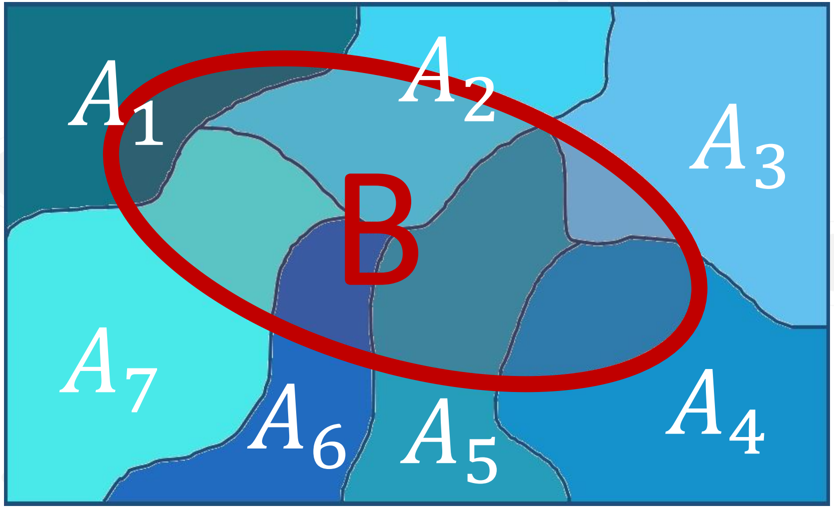
Probabilidad Condicional:
Probabilidad de $\boldsymbol{B}$ dado $A1$

$P(A_1) =$



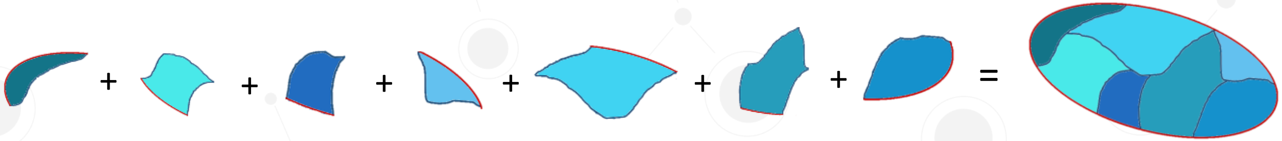
$$P(B|A_1) = \frac{P(B\ \cap A_1)}{P(A_1)} =$$

Técnicas de probabilidad

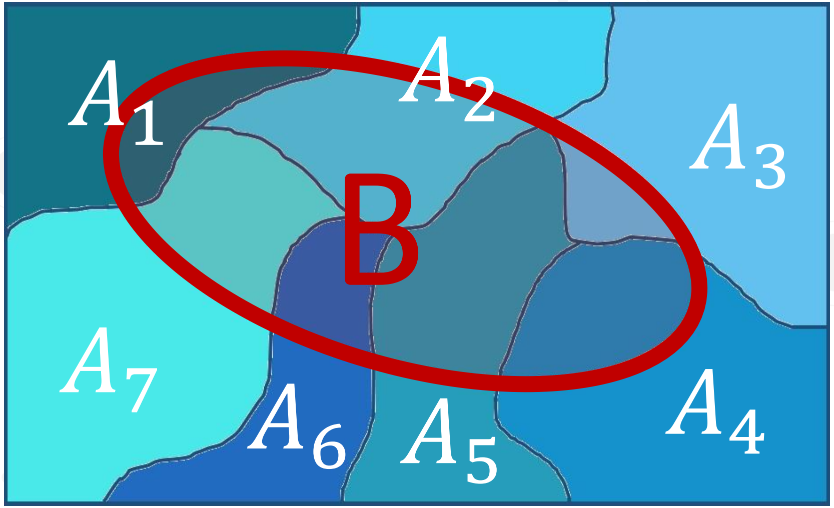


Teorema de Probabilidades Totales: Probabilidad de $\boldsymbol{B}$

$$P(B) = \sum_i P(B \cap A_i) =$$

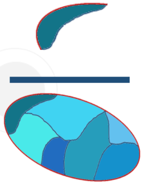


Técnicas de probabilidad



Teorema de Bayes: Probabilidad de A_1 dado B

$$P(A_1|B) = \frac{P(B \cap A_1)}{P(B)} = \frac{P(B|A_1)P(A_1)}{P(B)} =$$



Técnicas de probabilidad

En una empresa están determinando cuanto café comprar para sus empleados. Le consultaron a 300 de ellos, y estos fueron los resultados.

	Consumo de café			
Edad (años)	Bajo	Moderado	Alto	Total
Menos de 30	36	32	24	92
30 a 40	18	30	27	75
40 a 50	10	24	20	54
50 o más	26	24	29	79
Total	90	110	100	300

¿Cuál es la probabilidad de tener Menos de 30 años?

Técnicas de probabilidad

¿Cuál es la probabilidad de consumir bajo café, dado que el entrevistado es menor a 30 años?

Probabilidad condicional

	Consumo de café			
Edad (años)	Bajo	Moderado	Alto	Total
Menos de 30	36	32	24	92
30 a 40	18	30	27	75
40 a 50	10	24	20	54
50 o más	26	24	29	79
Total	90	110	100	300

$$P(Consumo\ de\ café\ bajo) = \frac{90}{300} = 0.3 = 30\%$$

$$P(Consumo\ de\ café\ bajo|Menos\ de\ 30) = \frac{36}{92} = 0.39 = 39\%$$

Técnicas de probabilidad

¿Cuál es la probabilidad de consumir bajo café?

	Consumo de café			
Edad (años)	Bajo	Moderado	Alto	Total
Menos de 30	36	32	24	92
30 a 40	18	30	27	75
40 a 50	10	24	20	54
50 o más	26	24	29	79
Total	90	110	100	300

Teorema de probabilidades totales

$$P(Consumo\ de\ café\ bajo) = \frac{90}{300} = 0.3 = 30\%$$

$$P(Consumo\ de\ café\ bajo|Menos\ de\ 30\ años) * P(Menos\ de\ 30\ años) +$$
$$P(Consumo\ de\ café\ bajo|30\ a\ 40) * P(30\ a\ 40) +$$
$$P(Consumo\ de\ café\ bajo|40\ a\ 50) * P(40\ a\ 50) +$$
$$P(Consumo\ de\ café\ bajo|50\ o\ más) * P(50\ o\ más)$$
$$= \frac{36}{92} * \frac{92}{300} + \frac{18}{75} * \frac{75}{300} + \frac{10}{54} * \frac{54}{300} + \frac{26}{79} * \frac{79}{300} = \frac{90}{300} = 0.3 = 30\%$$

Técnicas de probabilidad

¿Cuál es la probabilidad de tener menos de 30 años, dado que consume bajo café?

Teorema de Bayes

Edad (años)	Consumo de café: Bajo	Moderado	Alto	Total
Menos de 30	36	32	24	92
30 a 40	18	30	27	75
40 a 50	10	24	20	54
50 o más	26	24	29	79
Total	90	110	100	300

$$P(Menos\ de\ 30|Consumo\ de\ café\ bajo) = \frac{36}{90} = 0.4 = 40\%$$

$$= \frac{P(Consumo\ de\ café\ bajo|Menos\ de\ 30\ años) * P(Menos\ de\ 30\ años)}{P(Consumo\ de\ café\ bajo)} = \frac{{}^{36*92}/_{92*300}}{{}^{90}/_{300}} = \frac{36}{90} = 40\%$$

Técnicas de probabilidad

¿Cuál es la probabilidad de consumir bajo café, dado que el entrevistado es menor a 30 años?

Probabilidad condicional

¿Cuál es la probabilidad de consumir bajo café?

Teorema de Probabilidades Totales

¿Cuál es la probabilidad de tener menos de 30 años, dado que consume bajo café?

Teorema de Bayes

Actividad

Por la normativa vigente, la mayoría de los edificios posee alarma contra incendios, pero algunas veces se utiliza para algo "indebido" (ajeno a un incendio). A partir de las estructuras y cortafuegos actuales, la probabilidad que ocurra un incendio es de un 7%. Junto a esto, la probabilidad que suene una alarma cuando hay un incendio es de un 92%, mientras que la probabilidad que suene, dado que fueron asuntos indebidos es de un 12%.

a. Calcule la probabilidad que suene una alarma.

b. Calcule la probabilidad que, dado que sonó la alarma, esta sea por asuntos indebidos

Actividad

A través de la Encuesta Nacional de Salud, obtenga:

i. Variable categórica del rango del colesterol: Menor a 150, 150 a 180, 180 a 200 y mayor a 200.

ii. Obtenga una tabla de contingencia entre el rango de colesterol y el rango de educación (NEDU).

iii. Obtenga la probabilidad de tener más de 12 años de escolaridad

iv. Obtenga la probabilidad de tener más de 12 años de escolaridad, dado que se tiene de 150 a 180 mg/dL de colesterol. ¿Hay mucha diferencia con el resultado anterior?

v. Obtenga la probabilidad de tener entre 150 a 180 mg/dL de colesterol, dado que se tiene más de 12 años de escolaridad. Compare

Modelo de probabilidad

Una **distribución de probabilidad** muestra los posibles resultados de un experimento y la probabilidad de que cada uno se presente, de forma generalizada.

Una distribución puede estar definida sobre una **variable aleatoria discreta**, en el caso que los resultados sean contables (o numerables).

Una distribución también se puede definir en una **variable aleatoria continua**, si hay "infinitas" posibilidades, tal que no sea contable dos números consecutivos.

Modelo de probabilidad

EXPERIMENTO	VARIABLE	VALORES POSIBLES
OBSERVAR UN CLIENTE EN LA FILA DE UNA SUCURSAL DE COMIDA RÁPIDA	¿CUÁNTOS PLATOS, DE LOS 5 DISPONIBLES, PODRÍA PEDIR UN CLIENTE?	1, 2, 3, 4, 5
INSPECCIONAR UN LOTE DE 50 CELULARES	NÚMERO DE CELULARES CON ALGÚN DEFECTO	0,1,2,…,49,50
SUPERVISAR UN PEAJE EN LA AUTOPISTA	NÚMERO DE VEHÍCULOS DIARIOS	0,1,2,3,…
REALIZAR UNA CAMPAÑA PERSONALIZADA	SEXO DEL CLIENTE	0: SI ES HOMBRE 1: SI ES MUJER

Modelo de probabilidad

EXPERIMENTO	VARIABLE	VALORES POSIBLES
OPERAR UN BANCO	TIEMPO ENTRE LA LLEGADA DE DOS CLIENTES	$[0,\infty)$
RELLENAR UNA LATA	CANTIDAD DE ML	$[0,350]$
CONSTRUIR UN PROYECTO INMOBILIARIO	AVANCE DEL PROYECTO DURANTE UN PERÍODO	$[0,1]$
OBSERVAR EL MOVIMIENTO DE UNA ACCIÓN	RENTABILIDAD DE UN MES A OTRO	$(-\infty,\infty)$

Modelo de probabilidad

En el caso de las distribuciones, interesará conocer dos indicadores: **Valor Esperado** (o Esperanza) que representa el valor de la distribución más probable, mientras que la **Varianza** se relaciona a la dispersión de la distribución.

$$\sigma^2 = Var(X) = \begin{cases} \sum_x (x-\mu)^2 P(X=x) \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \end{cases}$$

$$\mu = E(X) = \begin{cases} \sum_x xP(X=x) \\ \int_{-\infty}^{\infty} xf(x)dx \end{cases}$$

$$Var(X) = E(X^2) - E(X)^2$$

Modelo de probabilidad

Un **modelo de probabilidad** responde a una generalización de diversos experimentos, cuyos posibles resultados se pueden adaptar a una sola función matemática, tomando valores sus **parámetros**. En el caso discreto, los más usuales son:

- Modelo Binomial
- Modelo Binomial Negativo
- Modelo Poisson

<u>Y en el caso continuo</u>

- Modelo Uniforme
- Modelo Normal
- Modelo Gamma

Modelo de probabilidad

- El modelo **Modelo Binomial** se adapta para calcular la probabilidad de ocurrencia de **x** éxitos sobre una muestra de tamaño ***n***. El parámetro es la probabilidad de ser un "éxito".

- Un Modelo **Binomial Negativo** corresponde cuando se desea obtener la probabilidad de que, al obtener ***k*** éxitos, la muestra sea de tamaño **x**. El parámetro es la probabilidad de ser un "éxito"

- El **Modelo Poisson** está pensado cuando se desee obtener la probabilidad de observar **x** veces la ocurrencia de un evento, en un cierto espacio o tiempo. El parámetro es la tasa histórica de ocurrencia.

Modelo de probabilidad

Usualmente un sábado en la tarde, el 35% de las personas que acuden al mall realizan una compra en una tienda específica. Si se observa a 50 personas en un instante de tiempo, ¿Cuál es la probabilidad de que compren al menos 20 personas?

En Excel: DISTR.BINOM.N

En R: pbinom, dbinom, qbinom.

Modelo de probabilidad

En la misma situación, al entrar a una tienda cualquiera el día sábado en la tarde, nos informan que han realizado 15 ventas. ¿Cuál es la probabilidad de que hayan circulado 50 personas?

En Excel: NEGBINOM.DIST
En R: pnbinom, dnbinom, qnbinom.

Modelo de probabilidad

Si la tasa en la que acuden personas al mall un día sábado en la tarde es de 80 personas por hora, ¿Cuál es la probabilidad de contar en la entrada, en 30 minutos, a lo más a 35 personas?

En Excel: POISSON.DIST
En R: ppois, dpois, qpois.

Modelo de probabilidad

- El modelo **Modelo Uniforme** se adapta para situaciones donde todas las situaciones son equiprobables. Los parámetros son los límites de la variable.

- Un Modelo **Normal** se utiliza cuando se observa una variable simétrica, cuyos datos se concentran en torno a la media. Los parámetros son dicha media, y la dispersión respecto a ella.

- El **Modelo Gamma** es útil cuando los datos no presentan simetría. Está reservado solo cuando la variable es positiva, y es flexible pensando en parámetros de forma y tasa.

Modelo de probabilidad

La espera de un pasajero en el metro, desde que llega hasta que el tren ingresa a la estación, puede ser descrito con una distribución uniforme entre 0 y 10 minutos. ¿Cuál es la probabilidad de esperar entre 2 a 6 minutos?

En Excel: x-min/(max – min)
En R: punif, dunif, qunif.

Modelo de probabilidad

El último estudio de obesidad escolar indicó que el peso (en kg) de alumnos entre 5° y 6° básico posee una distribución normal, con media 37kg y dispersión de 5 kg. ¿Cuál es la probabilidad de, al seleccionar un alumno al azar, este pese entre 36kg y 38kg?

En Excel: DISTR.NORM.N
En R: pnorm, dnorm, qnorm.

Modelo de probabilidad

El aeropuerto de Santiago está en constantes mejoras para la salida de sus aviones. En particular, el tiempo entre la salida de dos aviones posee distribución Gamma, con parámetro de forma igual a 3, y tasa 0.2. ¿Cuál es la probabilidad que un avión demore más de 10 minutos en salir, luego de la salida del anterior?

En Excel: DISTR.GAMMA.N(X;forma;1/tasa;ACUM)
En R: pgamma, dgamma, qgamma.