# Introducción a la Estadística

## Introducción

La estadística surge de estimar o sensar que tan probable es que ocurran eventos.

Los eventos probabilísticos generalemente son conjuntos en espacios de elementos, dónde cada elemento representa evento. Es decir, como tener un conjunto de eventos.

## Probabilidad

Una probabilidad es entonces un número entre 0 y 1 asociado a un subconjunto de eventos que podrían ocurrir del espacio de eventos.

Por ejemplo, si el espacio de eventos está constituido por todos los posibles eventos de lanzar dos dados, entonces, un subconjunto de espacio podría ser, todos los eventos dónde los dados suman una cantidad, por ejemplo, que sumen 5.

```
ESPACIO DE EVENTOS: { (1, 1), (1, 2), ..., (2, 1), (2, 2), ..., (6, 6) }

SUBCONJUNTO DE EVENTOS:
    A: DÓNDE AMBOS DADOS CAEN IGUAL { (1, 1), (2, 2), (3, 3), ..., (6, 6) }
    B: DÓNDE SON PARES LAS CARAS DE LOS DADOS { (2, 2), (2, 4), ..., (4, 2), (4,
4), ..., (6, 6) }
    C: DÓNDE LAS CARAS SUMEN 5 { (1, 4), (4, 1), (2, 3), (3, 2) }
```

La probabilidad `P` es una función que recibe un subconjunto de eventos y determina un número entre 0 y 1 de que el evento ocurra. Generalmente basta con dividir el número de eventos posibles entre el número de eventos totales. Siendo entonces la probabilidad, un conteo de cúantas veces podría ocurrir un subconjunto de eventos.

```
P(A) = 6 / 36 (1 / 6 ~ 0.166....)
P(B) = 9 / 36 (1 / 4 ~ 0.25)
P(C) = 4 / 36 (1 / 9 ~ 0.111....)
```

Dentro de la probabilidad hay más cosas que se pueden determinar, por ejemplo, la probabilidad de que dos eventos ocurran al mismo tiempo `P(A^B)`, la probabilidad de que un evento ocurra, dado que otro evento ya ocurrió `P(A | B)`, y determinar si los eventos son o no independientes.

## Muestras

La estadística se basa en la probabilidad, es decir, crea estimaciones, de acuerdo a que tan probable es que es ocurran las cosas. Generalmente la estadística se centra en eventos repetidos a lo largo de varias muestras y estas muestras se consideran las base del estudio.

Por ejemplo, una persona podría recolectar información acerca de las personas que transitan en la calle, preguntándoles su edad, su género, su nivel económico o académico, etc.

Cada muestra representa un paquete de datos con información multidimensional. Es decir, cada muestra es un vector de datos. Este vector conforma un registro de un evento ocurrido, por ejemplo, una medición sobre un experimento o un una medición sobre un individuo. Se podría decir que la muestra o el registro es lo que quedó plasmado del evento.

El preguntarnos que tan probable es que una muestra se repita, implica saber, qué tan probable es que un evento se repita. Sin embargo, surge el problema que las muestras son multidimensionales, es decir, en varias muestras se podría repetir la edad, el género, el nivel socioeconómico, etc. Pero qué podríamos decir de estas repeticiones.

## Frecuencias

La `frecuencia` es un conteo sobre qué tanto se repite un valor sobre un eje. El eje representa un eje datos o una columna de un vector de datos. Es decir, a partir de ahora estaremos pensando las muestras como vectores de datos, y cada columna del vector, representaría un eje de datos.

```
MUESTRA

REGISTRO: (Ana Ming, 23 años, Mujer, $23.000 , Doctora, Médico General) # REGISTRO

VECTOR: (3, 23, 4, 2, 23.000, 3, 5, 7) # VECTOR DE DATOS/ANÁLISIS

1: CATEGORÍA (1 - Nombres Nacionales, 2 - Nombres Mixtos, 3 - Nombres Extranjeros)
2: NOMINAL (0 ~ 100+)
3: CATEGORÍA (1 - INFANTE, 2 - ADOLESCENTE, 3 - JOVEN-ADOLESCENTE, 4 - JOVEN-
ADULTO, ...)
4: CATEGORÍA (1 - HOMBRE, 2 - MUJER, 3 - OTRO)
...
8: CATEGORÍA (1 - SIN-ESTUDIO, 2 - PRIMARIA, ..., 7 - POSGRADO)
```

Los ejes de datos generalmente pueden ser `CATEGÓRICOS`, `NOMINALES`, `REALES` y `GEOESPACIALES`. Sin embargo, podrían existir otros modelos y mediciones de datos.

La `frecuencia` viene a ser un conteo o el número de veces que un valor puede caer en un intervalor dentro de un eje de datos. A este también se le conoce como el *histograma*.

### Ejercicios

1. Da ejemplos de al menos dos muestras distintas
    - ¿Cómo se ven sus registros?
    - ¿Cómo se ven sus vectores de datos?
2. ¿Qué tipo de dato tienen los siguientes ejes?
    - Una edad
    - Un peso
    - Una estatura

    - Un género
    - Un nivel económico/académico
    - Lo que gana en pesos una persona
    - El monto total de una venta
    - El modelo de un celular
3. ¿Se puede dividir un dato del registro de la muestra en varios ejes del vector de datos?
    - SI: ¿Cómo y por qué?