



MA34B – Estadística

Introducción

Prof. Rodrigo Abt B.
rabt@dim.uchile.cl

Algunas preguntas (1)

- Se lanza una moneda 10.000 veces, obteniéndose un total de 4.387 caras. ¿Está cargada la moneda?
- Un agricultor puede plantar papas o tomates. Si no llueve y planta papas, recibe \$1.000 por saco; pero si planta tomates, solo recibe \$600 por saco. En cambio si llueve y planta papas, recibe \$800, frente a los \$1.200 que recibiría si planta tomates. Según el pronóstico del tiempo, existe un 75% de probabilidad de que llueva. ¿Qué le conviene plantar al agricultor?
- El ejecutivo de un banco de un conocido banco observa preocupado que 23 de la 40 cuentas que maneja presentan sobregiros, lo cual comunica al gerente. El gerente lo tranquiliza indicando que históricamente las cuentas con sobregiro en el banco no superan el 50%?. ¿Debe preocuparse el ejecutivo con esta respuesta?
- Una barra de acero se somete a una prueba de calor y se mide su longitud (cm) para diferentes temperaturas ($^{\circ}\text{C}$), obtenéndose la siguiente tabla:

Algunas preguntas (2)

Temperatura (°C)	Logintud (cm)
30	25
40	25.2
50	25.7
60	27.1
70	27.9
80	28.1
90	28.6
100	?

¿Puede predecir la longitud de la barra a 100 °C ?

Algunas preguntas (3)

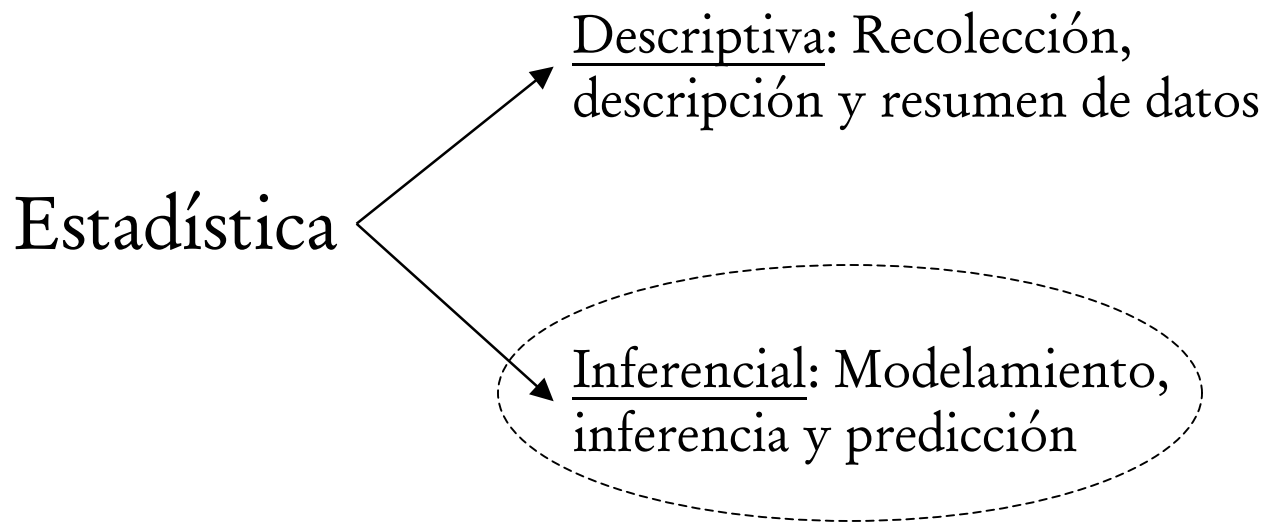
- Un estudio del Programa de las Naciones Unidas para el Desarrollo (PNUD), de 1998, revela algunos datos de ingreso y educación para comunas de Santiago

Comuna	Nº de Habitantes	%Alfabetización	Años escolaridad	%Matriculados	Ingreso Per Cápita (pesos)
Conchalí	380.016	98.7	9.26	87.3	73.950
Providencia	89.324	99.4	13.75	81.8	372.394
Las Condes	345.325	99.3	13.75	82.3	291.573
Ñuñoa	171.462	99.3	12.73	82.7	206.711
Renca	151.632	97.3	9.01	69	55.373
Qta. Normal	92.834	98.3	8.91	73.5	66.505
Maipú	305.140	99.5	10.54	73.4	94.061
La Florida	309.140	95	9.23	66	86.019

¿Qué puede decir al respecto?

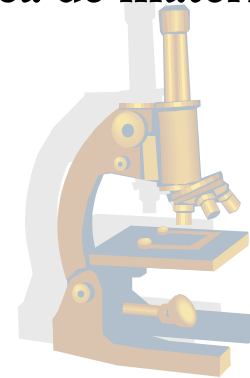
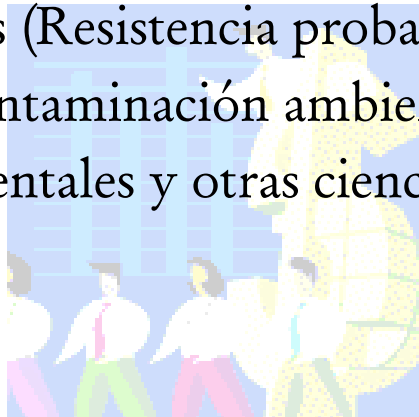
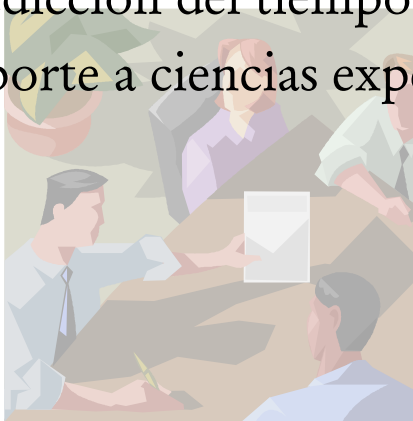
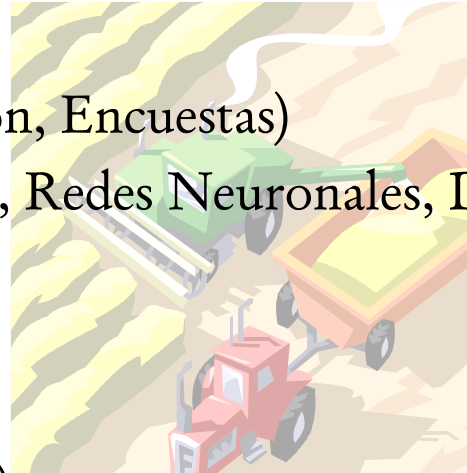
¿Qué es la Estadística?

- *“Es una disciplina de las matemáticas utilizada para describir, analizar e interpretar datos en los que interviene el fenómeno del azar”*



Aplicaciones

- Procesos productivos (Control de calidad)
- Estudios de mercado (Clusters, Segmentación, Encuestas)
- E-Commerce, Detección de patrones (OCR, Redes Neuronales, Data Mining)
- Búsqueda de yacimientos (Geoestadística)
- Genética, Biotecnología (Bioestadística)
- Economía (Series económicas, Econometría)
- Confiabilidad de Materiales (Resistencia probabilística de materiales)
- Predicción del tiempo y contaminación ambiental
- Soporte a ciencias experimentales y otras ciencias



El razonamiento estadístico

- Todo problema estadístico correctamente planteado se puede esquematizar en los siguientes pasos:

Paso 0: ∴ Definir objetivos y preguntas de estudio !!!!



1. Recolección de datos: definición de universo y marco de estudio, selección de variables y tipo de muestreo).
2. Descripción de los datos (Estadística Descriptiva): visualización, valores representativos, variabilidad, tendencias, valores extremos, datos perdidos, limpieza, etc.).
3. Análisis de los datos: estimación, elección de modelos, inferencia.
4. Decisión-Predicción: contraste de hipótesis iniciales, extensiones, validación

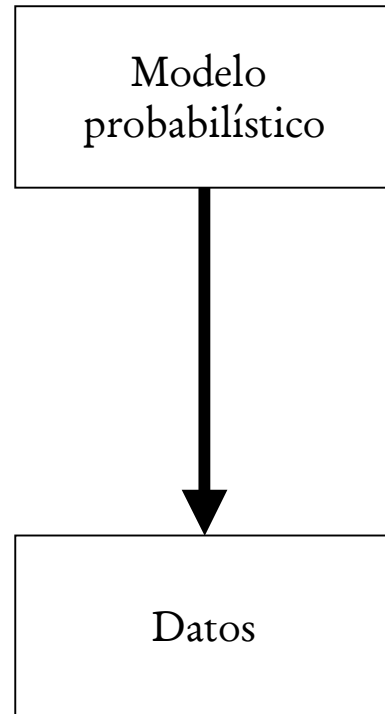
El azar

- Lo que diferencia un problema estadístico de cualquier otro problema con datos es la presencia del azar o incertidumbre. El fenómeno del azar se manifiesta en fluctuaciones de los datos, cuyo origen, en general, no depende del investigador.

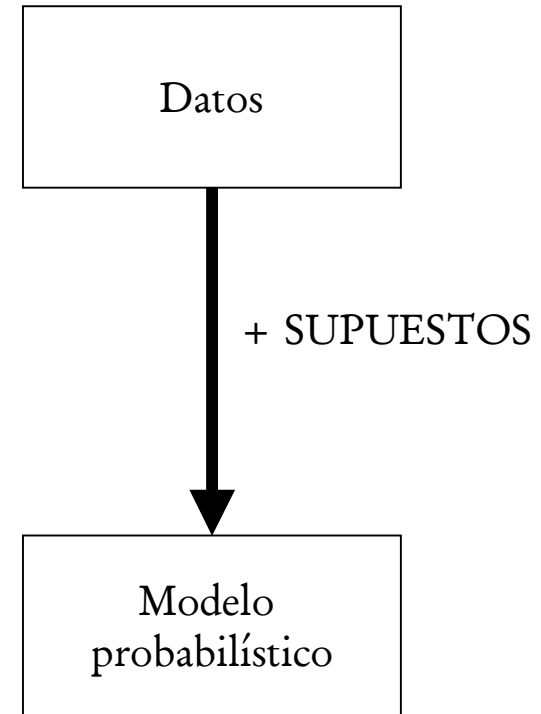
“El azar refleja la ignorancia que tenemos de las cadenas causales de la naturaleza. Los fenómenos fortuitos son, por definición, aquellos cuyas leyes ignoramos” – Henri, Poincaré (1854-1912)

- Incluso, cuando existen fenómenos que suelen ser muy complejos o irregulares, tienden a ser percibidos como fenómenos aleatorios.
- En las matemáticas, el tipo de variables que representan sucesos, observaciones o resultados de un experimento en los que está presente el fenómeno del azar reciben una denominación bastante conocida: variables aleatorias, lo cual sugiere la utilización de la disciplina de las **PROBABILIDADES**.
 - En Probabilidades, el punto de partida es generalmente un determinado modelo probabilístico, que proporciona la respuesta que teóricamente debiese esperarse de una variable aleatoria.
 - Mientras que en Estadística, se parte al revés. Una vez obtenidos los valores de una variable aleatoria, se intenta reconstruir el (o los) posibles modelos que generaron dichas observaciones.

Probabilidades vs. Estadística



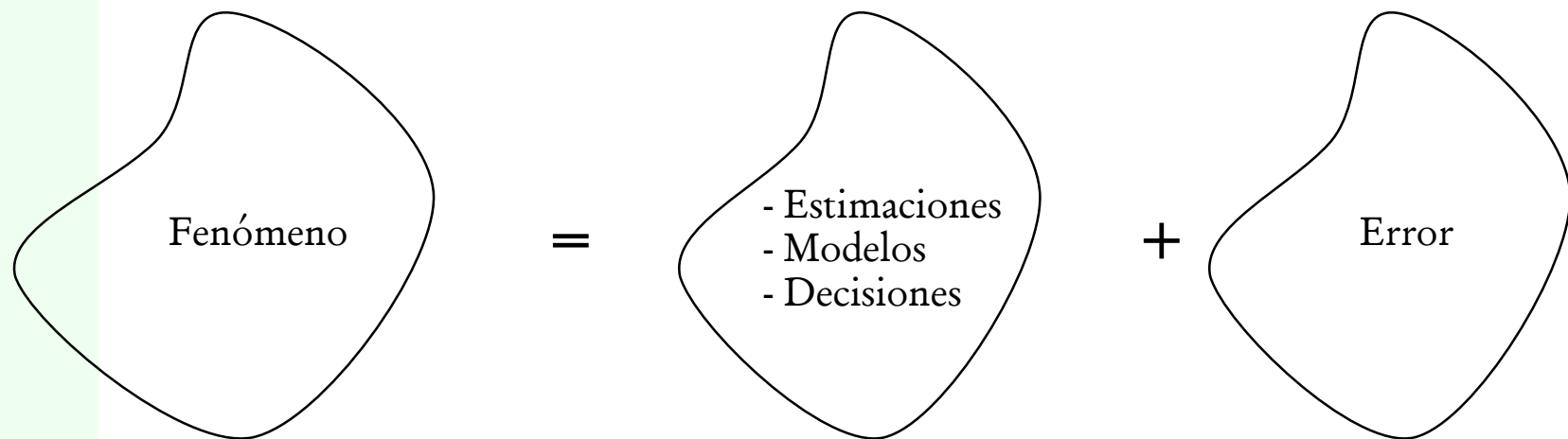
PROBABILIDADES



ESTADÍSTICA

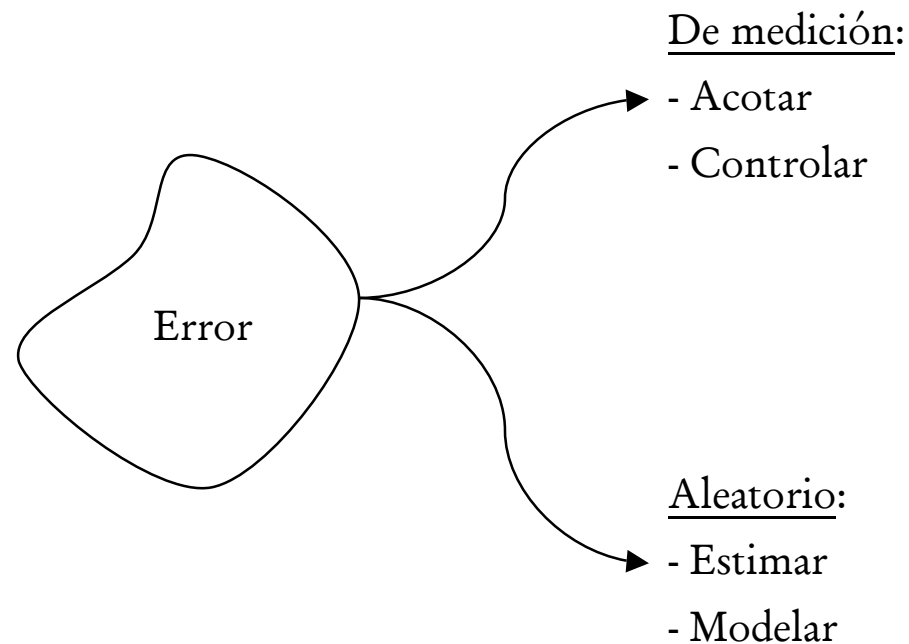
Presencia del error (1)

- Uno de los principales objetivos de la Estadística es generar un modelo de probabilístico que represente de *mejor* manera el fenómeno aleatorio en estudio. Dado que interviene el azar, no podemos garantizar un 100% de precisión en nuestros modelos, y por ello, *debemos estar dispuestos a tolerar un determinado margen de error.*



Presencia del error (2)

- En Estadística el error siempre está presente, y puede provenir de dos fuentes:

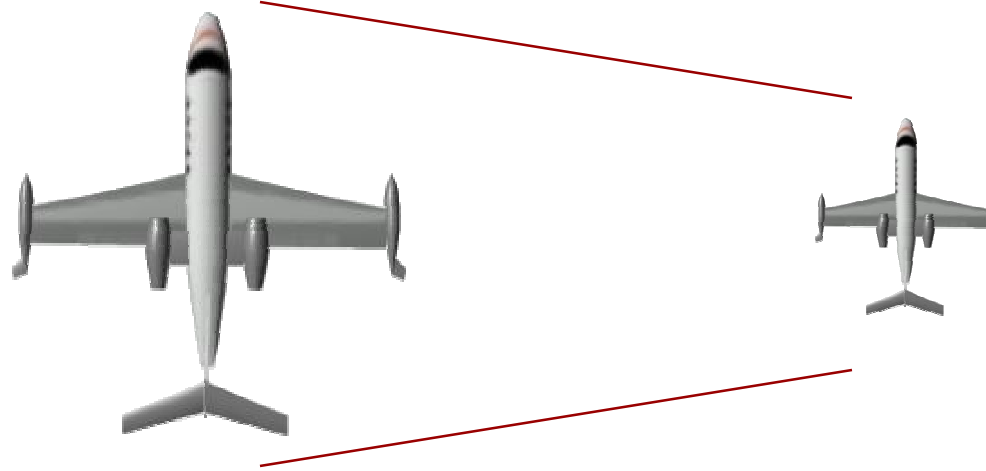


Población y Muestra

- Cuando un investigador trabaja con un fenómeno aleatorio, registra datos e información relacionados a un conjunto de elementos o individuos de estudio, al que se le denomina **Población**.
- La mayor parte del tiempo los investigadores trabajan con información limitada de la población, ya sea por tiempo, costo, disponibilidad y/o facilidad para obtener la información. En este caso se acostumbra a trabajar con un subconjunto de esta, denominado **Muestra**, que debe satisfacer determinadas características para garantizar buenos resultados.

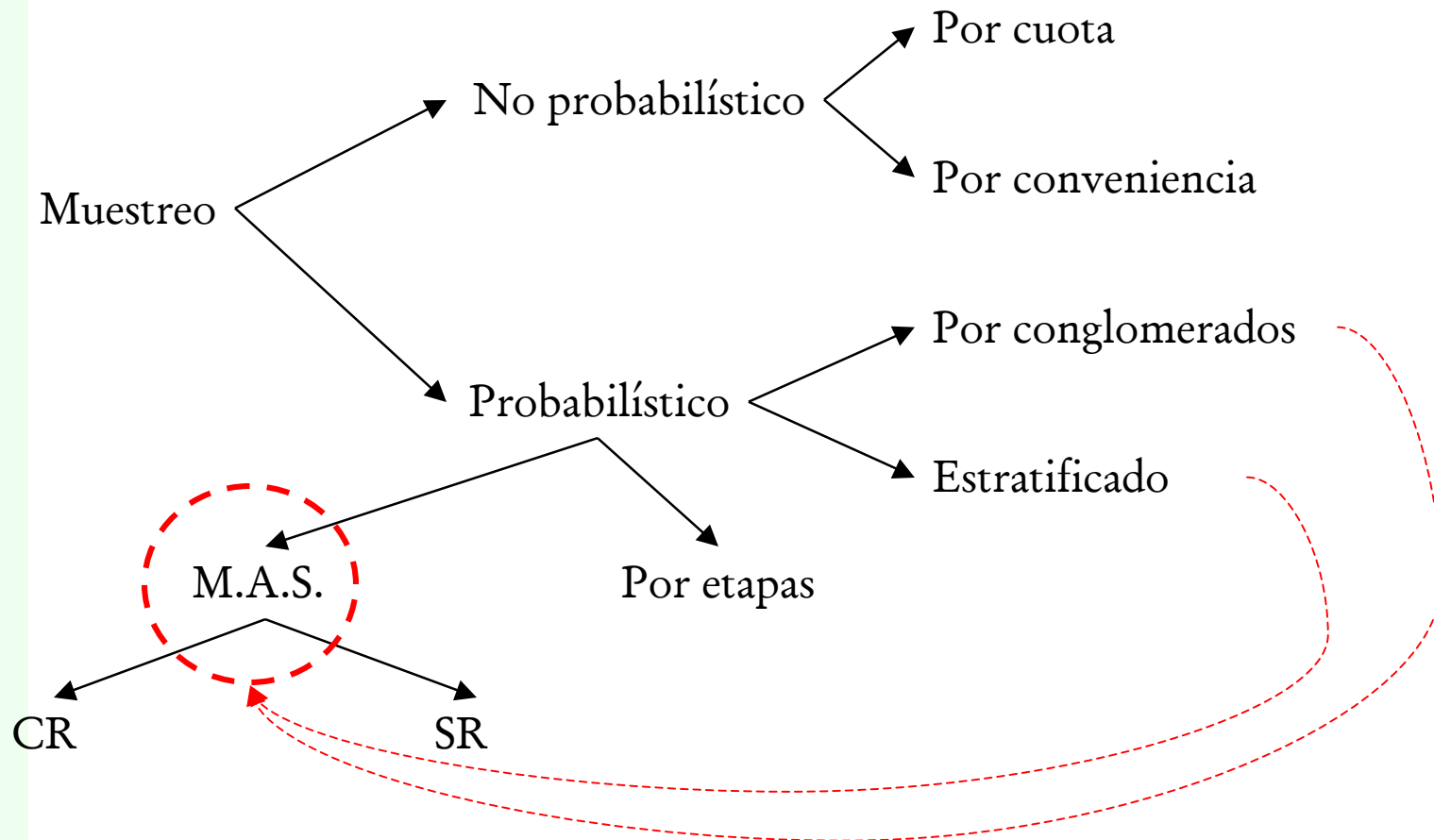
Características de una Muestra

- Una buena muestra debe ser idealmente *representativa* y *confiable*, esto es, debe reflejar de manera lo más fiel posible las características de la población en estudio, y además contar con la certeza de que los medios que se utilizaron para obtenerla garanticen dicha representatividad. Hay que pensar en cómo hace la mayoría de la gente para saber las características de una sopa.
- A modo de analogía, se puede pensar que la muestra es un modelo a escala de la población:



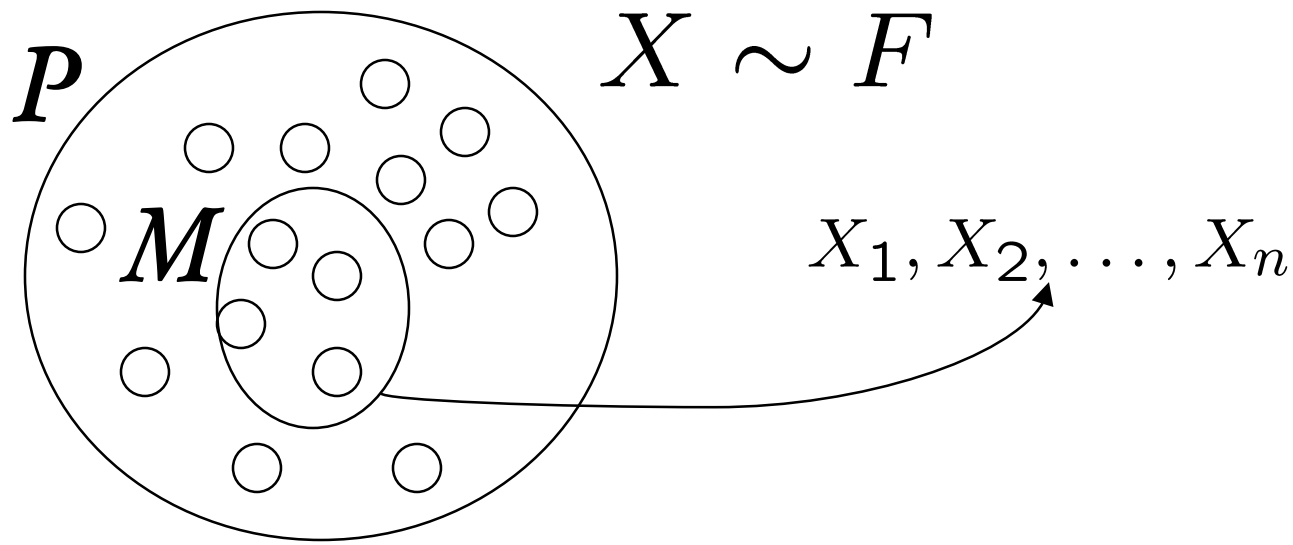
Muestreo

- La rama Estadística del **Muestreo** se encarga de las técnicas y métodos para determinar la muestra a utilizar. Los tipos de muestreo se pueden resumir en el siguiente esquema:



Supuestos básicos en Estadística

- Se tiene una población P , sobre la cual queremos estudiar una característica que medimos a través de una v.a. X .
- Se toma una muestra aleatoria simple (m.a.s.) M de tamaño n : X_1, X_2, \dots, X_n en que los X_i son i.i.d.
- La v.a. X sigue una distribución F no del todo desconocida
- Se pretende determinar F



Tipos de variables

- Si $X:\Omega \rightarrow Q$, dependiendo de Q , X puede ser:
 1. Cuantitativa
 - ❑ Continua ($Q \subseteq \mathbb{R}$)
 - ❑ Discreta ($Q \subseteq \mathbb{N}$)
 2. Cualitativa
 - Nominal (Q es un conjunto de atributos o categorías)
 - Ordinal (Q es un conjunto de atributos o categorías ordenadas)

Justificación del método

¿Cómo aseguramos que la muestra nos puede dar información para describir F ?

Caso cuantitativo (1)

- Si definimos:

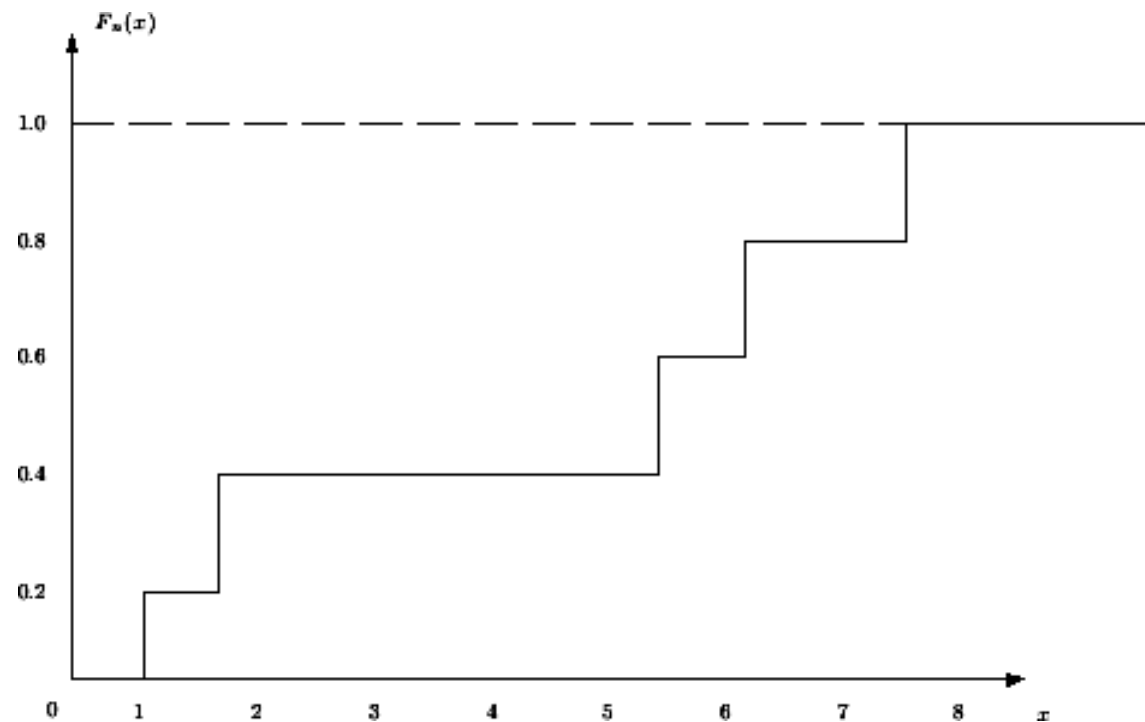
$$F_n(x) = \frac{\text{Card}\{x_i/x_i \leq x\}}{n}$$

Como la proporción de observaciones menores a x , entonces $F_n(x)$ corresponde a la **distribución empírica de X** .

- Ejemplo. Supongamos que observamos los siguientes valores:

Caso cuantitativo (2)

Obs.	X
1	1
2	2
3	1.7
4	5.5
5	6.2
6	7.6



Caso cuantitativo (3)

- La distribución empírica $F_n(x)$ tiene propiedades de una función distribución:
 - $F_n(-\infty) = 0$
 - $F_n(+\infty) = 1$
 - Si $x \leq y \Rightarrow F_n(x) \leq F_n(y)$
- Además, se puede notar que $nF_n(x)$ corresponde al número de observaciones menores o iguales a x , es decir, cuenta el número de “éxitos” entre n observaciones, por lo que $nF_n(x)$ se puede modelar como una distribución binomial, esto es: $nF_n(x) \sim \text{Bin}(n, P(X \leq x))$, siendo la última probabilidad igual a $F(x)$, la distribución teórica de X . De la ley de los grandes números se puede mostrar que, si n es grande, se debería esperar que $F_n(x)$ no difiera mucho de $F(x)$

Caso cualitativo

Si el conjunto $Q = \{q_1, q_2, \dots, q_n\}$ es un conjunto de atributos tal que:

$$P(X = q_j) = p_j \quad \forall j = 1, \dots, n$$

Representa la ley teórica de probabilidades de X , y dada una m.a.s. x_1, x_2, \dots, x_n , se define la ley empírica de proporciones como:

$$f_n(q_j) = \frac{\text{Card}\{x_i = q_i\}}{n} \quad j = 1, \dots, n$$

Siguiendo un razonamiento análogo se puede concluir que:

$$f_n(q_j) \rightarrow p_j$$

Notas

- Hemos probado que la distribución empírica de una m.a.s. de X nos acerca a la distribución real de X , lo que justifica esta forma de proceder.
- A pesar de que las distribuciones empíricas convergen a las teóricas, aumentar el tamaño de muestra no siempre es conveniente, ya que, si bien el error de muestreo decrece, los errores por causa de la población y medición aumentan. Lo ideal es tener un equilibrio entre ambos errores.
- Más que el tamaño de la muestra, importa el error que se espera respecto de la población.
- Para muestras grandes la diferencia entre usar muestreo con o sin reemplazo es despreciable.
- Los valores obtenidos de una m.a.s. son aleatorios, y por ende al repetir el experimento, los valores cambian.