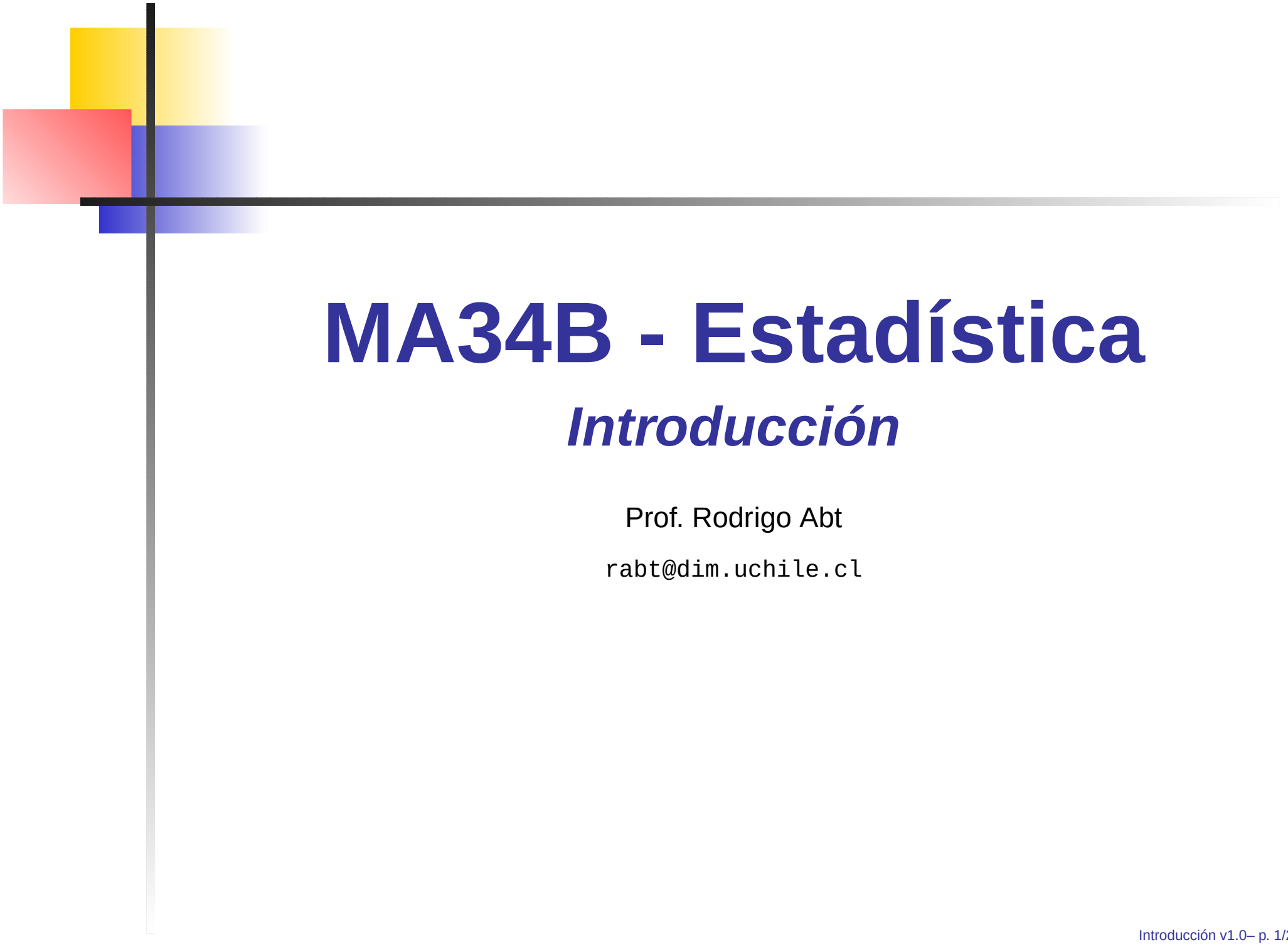


MA34B - Estadística

Introducción

Prof. Rodrigo Abt

rabt@dim.uchile.cl

Algunas cifras(1)

- Según las últimas encuestas en E.E.U.U. el senador John Kerry tiene un 49 % de apoyo de los votantes, mientras que Bush tiene el 47 %. ¿Puede aventurar quién ganará la elecciones presidenciales de ese país? (Fuente: La Tercera, Domingo 25 de julio 2003)
- Se lanza una moneda 10.000 veces, obteniéndose un total de 4387 caras. ¿Está cargada la moneda?
- Un agricultor puede plantar papas o tomates. Si planta papas y no llueve recibe $1.000 por saco, y si llueve recibe solo $600. En cambio si planta tomates, recibe $800 si no llueve, y $1.200 si llueve. Según el pronóstico del tiempo, existe un 75 % de probabilidad que llueva. ¿Qué le conviene plantar al agricultor?.
- Una barra de acero se somete a una prueba de calor y se mide su longitud (cm) para diferentes temperaturas (Celsius), obteniendo las siguientes observaciones: (30^oC,25cm),(40^oC,25.2cm),(50^oC,25.7cm),(60^oC,27.1cm),(70^oC,27.9cm), (80^oC,28.5cm),(90^oC,29.4cm).¿Puede decir cuánto se elongará la barra a 100^oC?

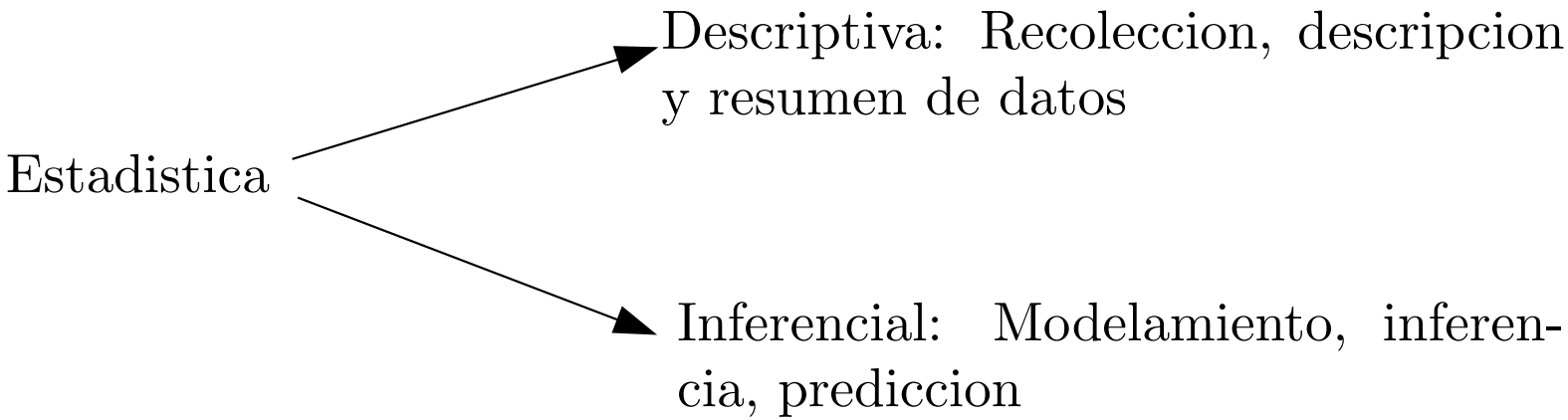
Algunas cifras(2)

- El ejecutivo de un banco observa preocupado que de 40 de sus cuentas bancarias, 23 tienen sobregiros, y se lo informa al gerente. El gerente lo tranquiliza mencionando que históricamente la proporción de cuentas con sobregiro es inferior al 50 %. ¿Tiene de qué preocuparse el ejecutivo?.
- Un estudio del PNUD para el año 1998 revela algunos datos de ingresos y educación para comunas del Gran Santiago. ¿Qué opina al respecto?

Comuna	Habs.	%Alfab.	Años escol.	% Matric.	Ing. percap.
Conchalí	380.016	98.7	9.26	9.26	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 es la Estadística?

"'Es una disciplina de las ciencias utilizada para describir, analizar e interpretar datos en los que interviene el fenómeno del azar, mediante el uso de procedimientos y técnicas de las matemáticas. La información obtenida a partir de los datos puede ser utilizada entonces para modelar, hacer predicciones y tomar decisiones respecto del fenómeno en estudio.'"



El razonamiento estadístico

El esquema de todo problema estadístico se puede resumir en lo siguiente:

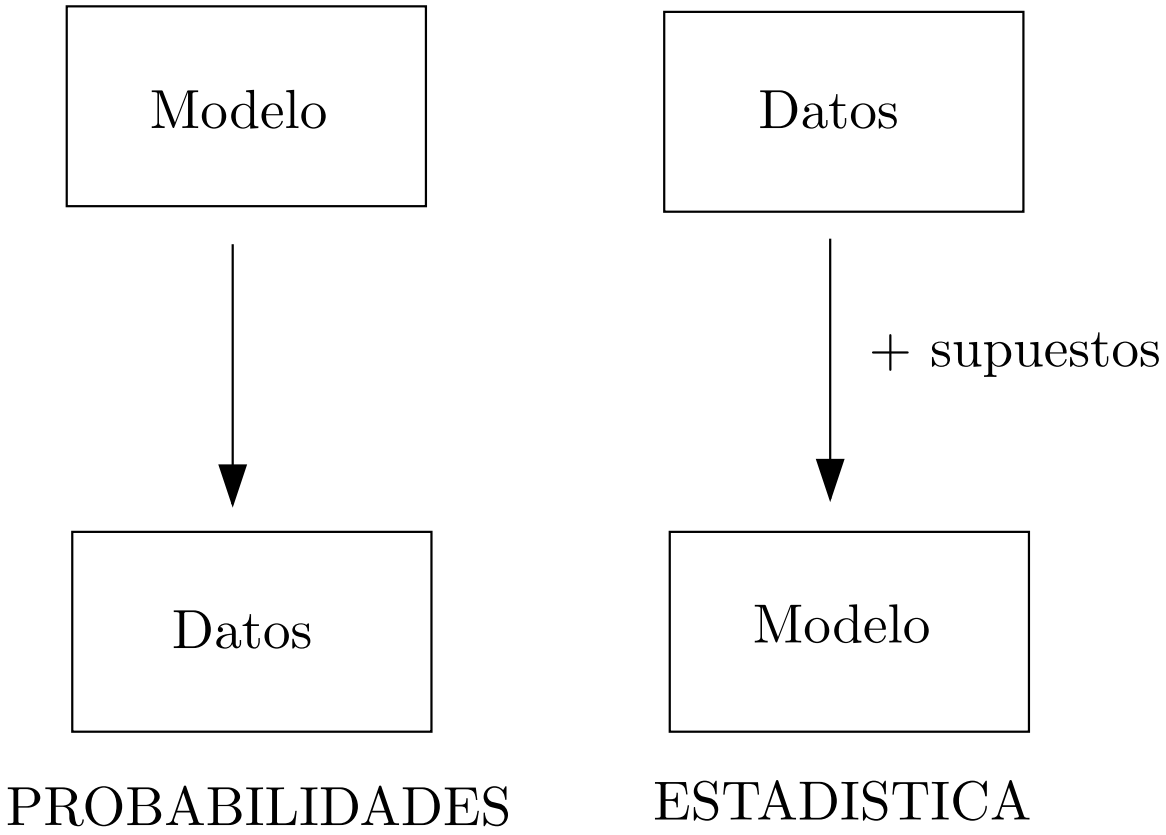
1. Recolección de datos (Tipos de muestreo, selección de variables, unidades de medida)
2. Descripción de los datos (Visualización, valores representativos, variabilidad)
3. Análisis de los datos (Estimación, modelamiento, inferencia)
4. Decisión-Predicción (Modelamiento, elección de criterios)

El azar

No todo problema con datos es susceptible de ser analizado desde el punto de vista estadístico, a menos que nos quedemos en un plano meramente descriptivo. 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 en los datos, cuyo origen, en general, no depende del investigador. 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 presencia 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

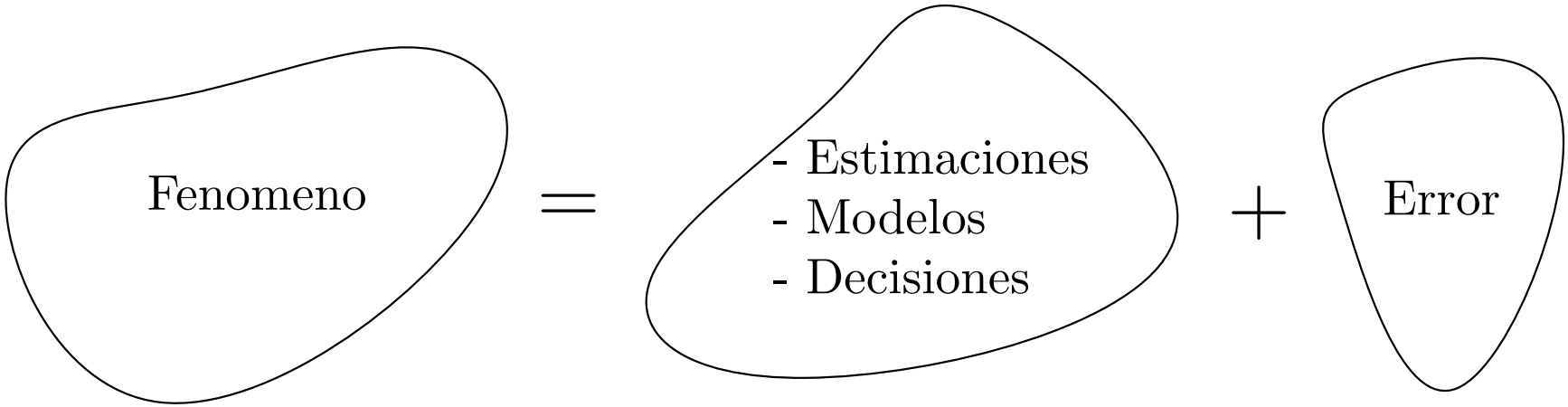


Aplicaciones

- Confiabilidad de materiales (Resistencia probabilística de materiales)
- Procesos productivos (Control de calidad)
- Estudios de mercados (análisis de clusters, análisis factorial)
- Algoritmos de clasificación (análisis discriminante, modelos logit-probit)
- Detección de patrones (OCR, Redes neuronales, Data Mining)
- Búsqueda de yacimientos (Geoestadística)
- Genética, biotecnología (Bioestadística)
- Sociología, sicología (Modelos de comportamiento, análsis de encuestas)
- Economía (Series económicas, Econometría)
- etcétera

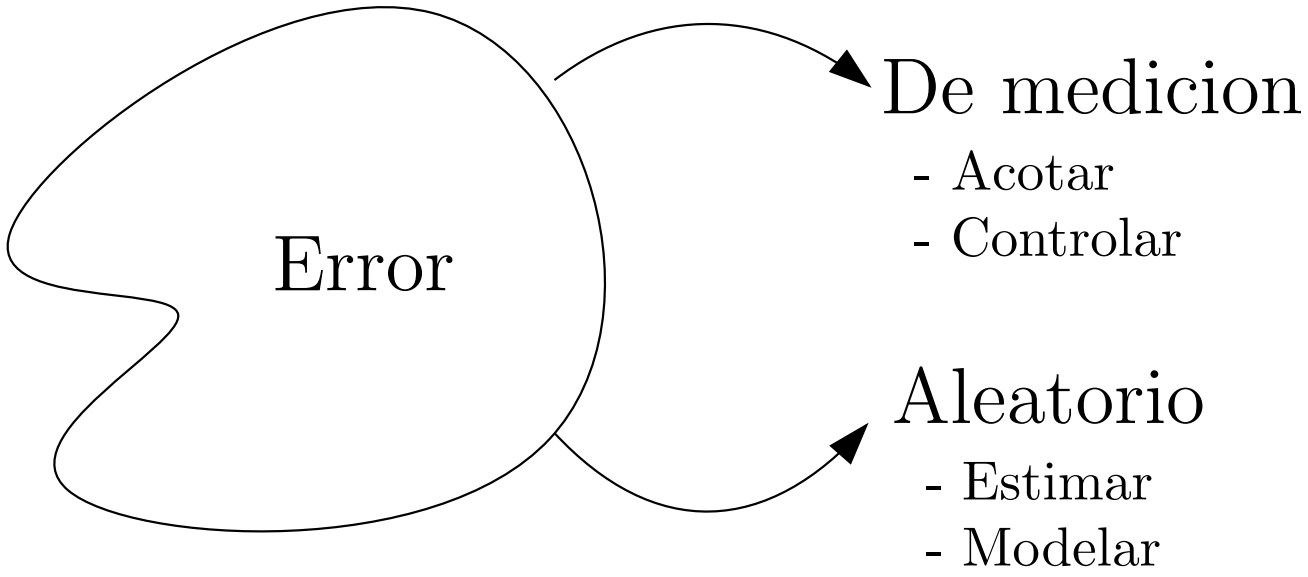
Presencia del Error(1)

Uno de los principales objetivos de la Estadística es generar un modelo de tipo probabilístico que represente de *mejor* manera el fenómeno en estudio. Dado que las observaciones provienen de un fenómeno en que interviene el azar, no podemos esperar que nuestros resultados sean %100 exactos. Siempre debemos estar dispuestos a aceptar o tolerar un determinado margen de error.



Presencia del Error(2)

En Estadística el error está siempre presente, y puede provenir, hablando de manera general de dos fuentes:

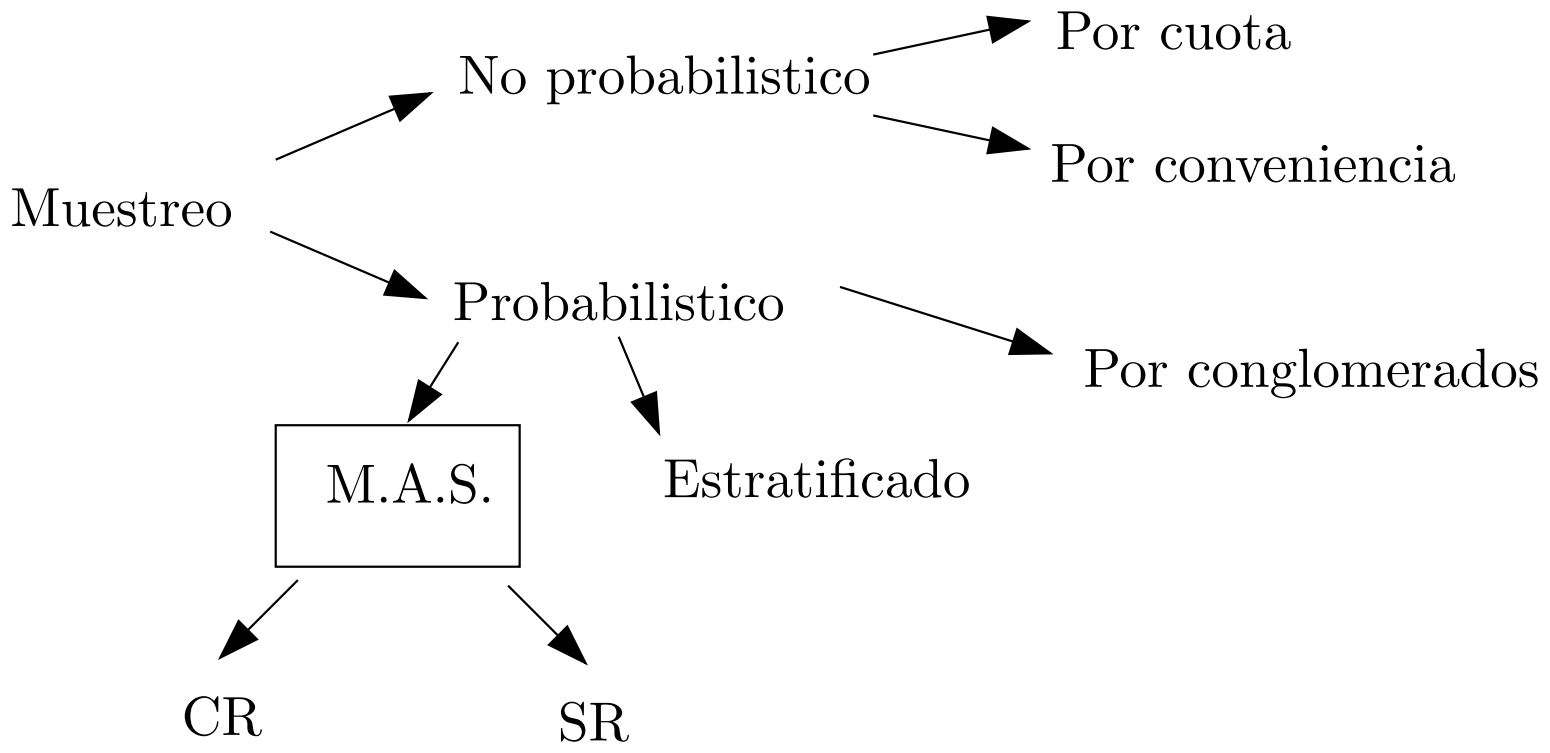


Población y Muestra

Además de suponer que los datos provienen de un fenómeno en que interviene el azar, el investigador la mayor parte del tiempo trabaja con información limitada, ya sea por costo, tiempo o facilidad de uso. En este caso, dada una población objetivo de estudio, se acostumbra a seleccionar un subconjunto de esta, denominado **muestra**, que sea lo más parecido posible al resto de la población. La muestra escogida deberá (idealmente) ser *representativa* y *confiable*.

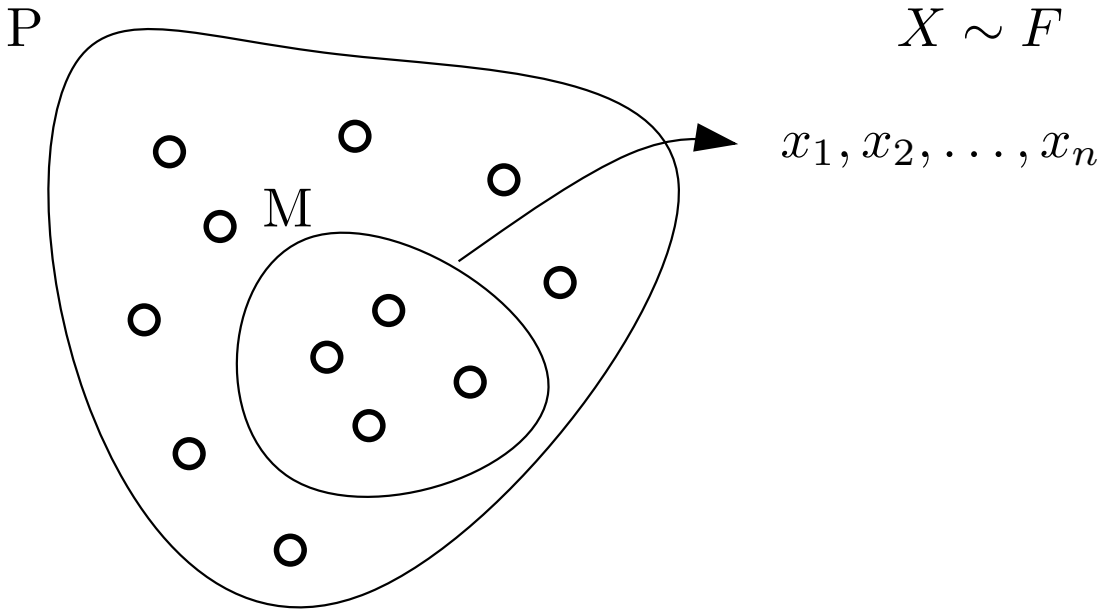
Muestreo

Una rama de la Estadística, denominada **Muestreo**, se encarga de las técnicas y métodos para determinar dicha muestra. A modo general se consideran los siguientes tipos de muestreo:



Supuestos en Estadística

- Se tiene una población **P**, sobre la cual queremos estudiar una característica que mediremos a través de una v.a. X.
- Se toma una muestra aleatoria simple (m.a.s.) **M** de tamaño n: $X_1, X_2, \ldots, X_n$, en que los X_i son i.i.d.
- La v.a. X sigue una distibución 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 ($Q \subseteq \mathcal{R}$)
   - Discreta ($Q \subseteq \mathcal{N}$)
2. Cualitativa
   - Nominal (Q es un conjunto de atributos o categorías)
   - Ordinal (Q es un conjunto de atributos o categorías ordenadas)

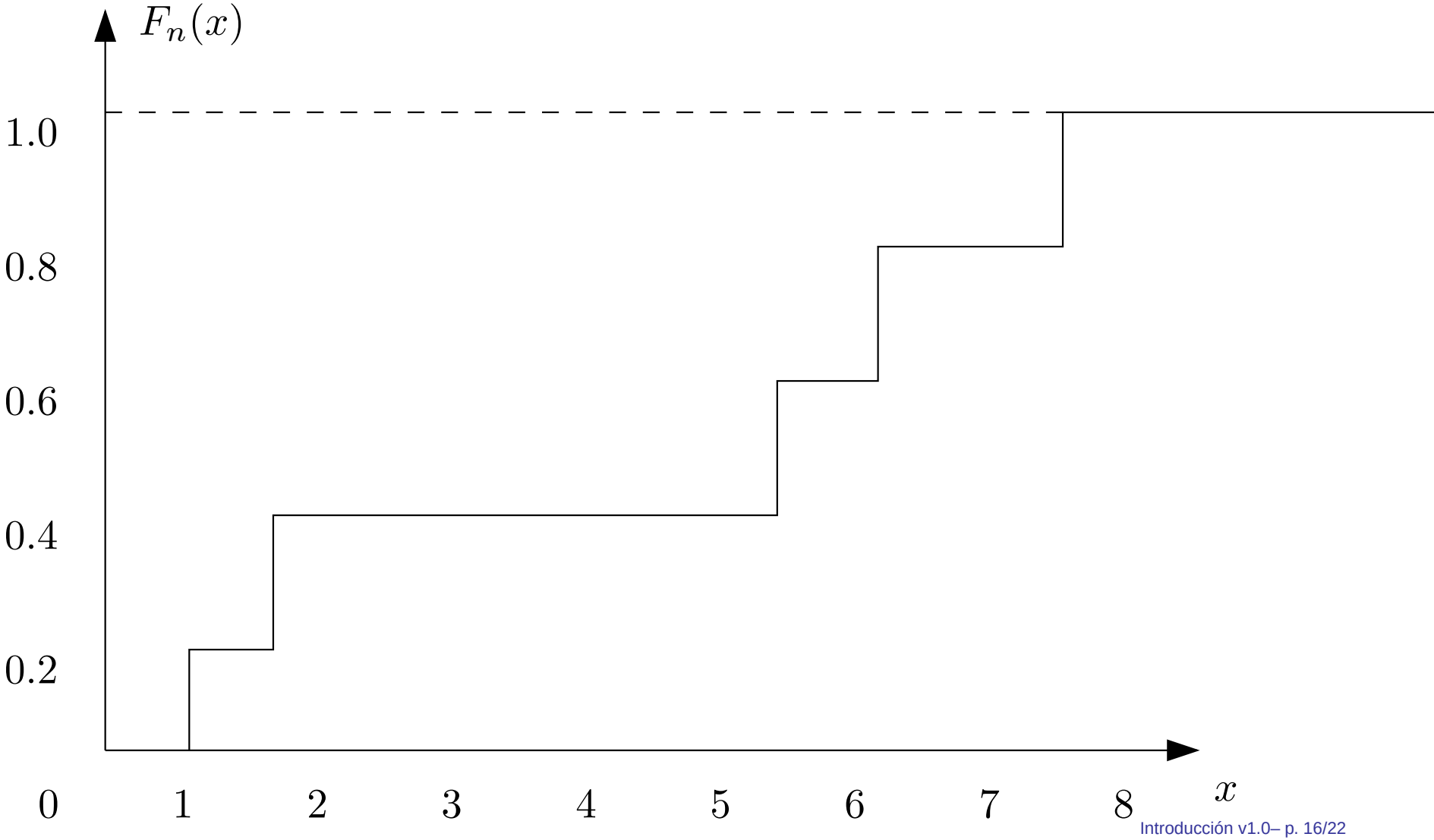
Caso cuantitativo(1)

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. Supongamos que observamos los siguientes valores: $1{,}2; 1{,}7; 5{,}5; 6{,}2; 7{,}6$, para una variable X. Entonces, el gráfico correspondiente de $F_n(x)$ será:

Caso cuantitativo(2)



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 binomial, esto es: $nF_n(x) \sim Bin(n, P(X \leq x))$, siendo la última probabilidad igual a $F(x)$, la distribución teórica de X.
De la ley de los grandes números, se puede demostrar que

$$P(\lim F_n(x) = F(x)) = 1 \qquad n \to \infty$$

Es decir, si n es grande, se debería esperar que $F_n(x)$ no difiera mucho de $F(x)$.

Caso cualitativo

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

$$P(X = q_j) = p_j \qquad \forall j = 1, \ldots, n$$

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

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

Entonces, con un razonamiento análogo se puede observar que $nf_n(q_j) \sim Bin(n, p_j)$, y que

$$P(\lim f_n(q_j) = p_j) = 1 \qquad n \to \infty$$

Respecto del muestreo

- Una población puede tener un número finito o infinito de elementos, por lo que trabajar con una muestra es casi siempre lo usual.
- A pesar de que las distribuciones empíricas convergen a las distribuciones teóricas, aumentar el tamaño de la muestra no siempre es conveniente, ya que, si bien el error de muestreo decrece con el tamaño de la muestra, los errores de observación y medición crecen con este tamaño. Por lo que es ideal buscar un equilibrio para ambos errores.
- Para muestras grandes la diferencia entre un muestreo con reemplazo y uno sin reemplazo es prácticamente despreciable. Por lo que se acostrumbra a usarse el primero.
- Los valores obtenidos en una muestra son aleatorios, ya que al repetir un experimento u obtener una nueva muestra, los valores cambian.

Estadísticos(1)

Una vez obtenida una muestra, se proceden a estudiar los valores muestrales obtenidos a partir de funciones denominadas **estadísticos**. Para un conjunto valores muestrales $X_1, X_2, \ldots, X_n$ un estadístico T es una función de los valores muestrales:

$$T = T(X_1, X_2, \ldots, X_n)$$

Algunos estadísticos usuales son:

- La media muestral: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- La varianza muestral: $S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Estadísticos(2)

- Los estadísticos de orden $X_{(k)}$,como el máximo $X_{(n)}$ y el mínimo $X_{(1)}$
- El rango: $W = X_{(n)} - X_{(1)}$
- Las cuantilas: $X_{[\alpha]}$, donde $\alpha = F^{-1}(X_{[\alpha]})$

Al ser funciones de variables aleatorias, los estadísticos son a su vez variables aleatorias, por lo que se pueden estudiar propiedades como esperanza, varianza, y por supuesto, sus distribuciones, denominadas *distribuciones en el muestreo*.

Ejemplo. La media muestral

Sea X una v.a. con media μ y varianza σ^2, y sea $X_1, X_2, \ldots, X_n$ una m.a.s. de X. Luego se tiene:

$$E(\bar{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E[\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$$

$$Var(\bar{X}) = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}Var[\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{\sigma^2}{n}$$