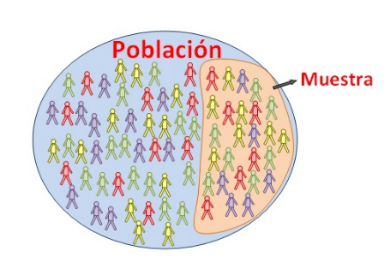
**Estadística Inferencial I**

**Población – Muestra:**



**Población:** conjunto de objetos o individuos. Se define en relación al problema a investigar.

**Población Finita:** Población cuyo número de individuos es finito. IE: Número de habitantes de una ciudad; producción de unidades por día de una planta.

**Población Infinita:** Población cuyo número de individuos es infinito. IE: Conjunto de números positivos; lanzamiento de un dado. Pueden considerarse como infinitas aquellas poblaciones con números extremadamente grandes (granos de arena de una playa; átomos del universo).

**Población Real:** Grupo de **elementos concretos**.

**Población Hipotética:** Conjunto de **situaciones posibles imaginables** en que puede representarse un suceso. IE: posibles reacciones de una persona ante una catástrofe.

**Muestra:** Se selecciona a partir de la población; subconjunto de unidades seleccionadas de una población definida. "**n**” unidades de una **muestra**. “**N**” unidades de una **población**.

**Censo:** Cuando se muestra a toda la población.

Al muestrear, podemos encontrarnos con **sesgos muestreales.** Estos pueden ser:

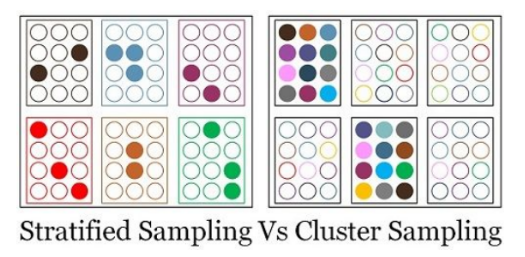
* **Por Conveniencia:** Es más fácil que seleccionemos a los individuos más fáciles de acceder.
* **No Respuesta:** Individuos que no responden la encuesta pueden dar lugar a falta de información de una porción no aleatoria de la muestra. Entonces la muestra deja de representar a la población.
* **Respuesta Voluntaria:** La contracara del caso anterior. Muestra compuesta solamente por personas que se ofrecen a responder (que suelen ser outliers con opiniones fuertes bien marcadas).

**Tipos de Muestras:**

* **Probabilísticas:** Podemos calcular qué probabilidad de selección tiene cada unidad de la muestra. Entonces podemos calcular una **medida del error**.

1. **Muestreo Aleatorio Simple**:Todos lo individuos tienen la misma probabilidad de ser elegidos.
2. **Muestreo Aleatorio Estratificado:** Los individuos se dividen en grupos o estratos. Se elige la muestra tomando individuos de cada estrato por muestreo aleatorio simple.
3. **Muestreo por Clusters**
4. **Muestreo Aleatorio por Etapas Múltiples** (Mezcla de 3 y 2).

****

****

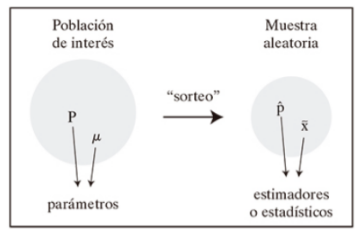
* **No Probabilísticas:** No salen de un proceso de selección aleatoria; En general sus sujetos se seleccionan a partir de la facilidad para acceder a los mismos o del criterio y/o intención del investigador. IE: Prueba de una nueva vacuna recién patentada en los pacientes de un hospital.

1. **Muestreo por Conveniencia**
2. **Muestreo Discrecional**

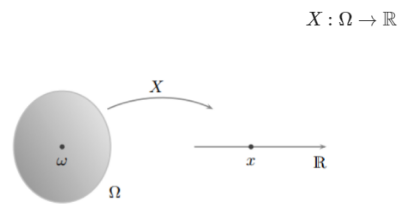
**Parámetros, Estimaciones y Estimadores**

El **objetivo** es **estimar** alguna **característica** que se supone **fija (no aleatoria)** de la **población**, que puede ser **simple** (media/proporción/varianza) o **más compleja** (coeficientes de una regresión/asociación entre variables). Dicha característica la denominamos **Parámetro.**

Un **Estimador** es un estadístico (en función de una muestra) que usamos para **estimar** un **parámetro desconocido** de la **población**. IE: la **media muestral** es un estimador de la **media poblacional**. Los estimadores son **variables aleatorias.**



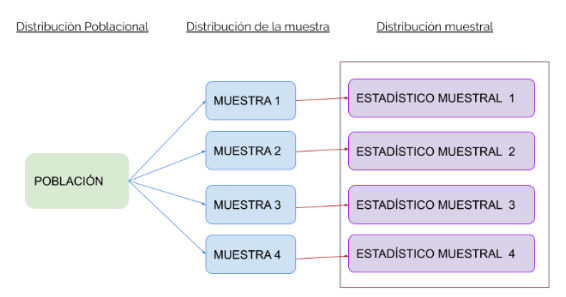
**Variable Aleatoria:** Transformación X del espacio de resultados al conjunto de números reales. Su valor está asociado al resultado de un experimento aleatorio. Su valor lo determina la realización del experimento.

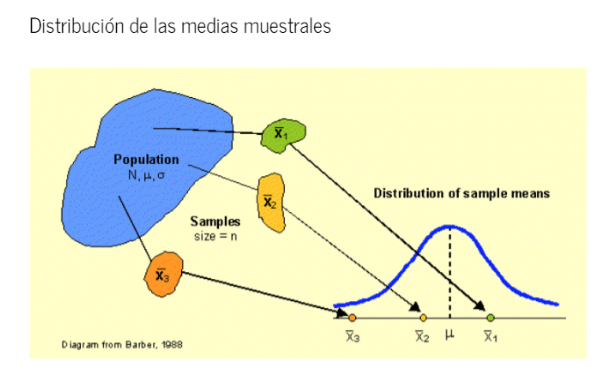
****

En general vamos a considerar que los parámetros son constantes en el tiempo y el espacio, a diferencia de los **estimadores.**

* **Estimación Puntual:** Un solo valor obtenido a partir de un estadístico. Es usar el valor de un estadístico (estimador) para calcular el valor de un **parámetro** desconocido de una población, el cual se considera fijo (no aleatorio).
* **Estimación por Intervalos de Confianza:** En vez de un valor fijo, se define un rango de valores posibles, con un límite inferior (Li) y superior (Ls) con una **probabilidad conocida** de que el valor verdadero se encuentre dentro de dicho rango.

**Distribuciones Muestrales**:





Dada una población, independientemente de su distribución, si tomamos muestras grandes y a cada muestra le calculamos su media muestral, la distribución de estas medias muestrales va a tener la forma de una distribución normal (Introducción al **Teorema del Límite Central**).

**Ley de Grandes Números:** A medida que el **tamaño de la muestra aumenta**, la **media muestral** se **acerca** a la **media poblacional**.

Planteos de **Kolmogorov:** dadas extracciones X1,…,Xn; si son observaciones **independientes** e **idénticamente distribuidas** de forma tal que E(Xi) sea igual a una constante µ finita. El promedio de Xi tiende a E(Xi) cuando n tiende a infinido.

**Teorema del Límite Central:**

Si X1,…,Xn son un conjunto de **variables aleatorias**, **independientes**, e **idénticamente distribuidas** de una distribución con media µ y varianza σ2 distinta de 0. Si n es lo suficientemente grande, entonces la media muestral tiene aproximadamente una distribución normal con µ = µ y σ2 =

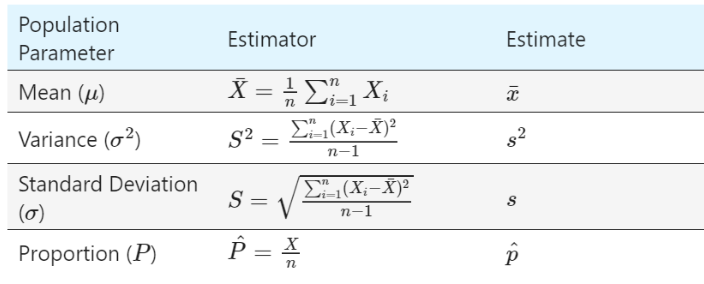
**Condiciones del Teorema del Límite Central:**

* **Independencia:** Las observaciones deben **proceder de un muestreo aleatorio** y deben ser **independientes**. Si el muestreo es sin reemplazo, entonces n > 10% de la población.
* En caso de tener una **distribución muy asímetrica,** **n debe ser muy grande**. (En algunos casos se habla de por lo menos 30; pero depende de la asimetría de la población). Con una **población normal**, **no hay condiciones sobre n**.
* Da una **distribución para nuestro estimador de la media poblacional** a partir de la **media muestral.**
* El **desvío estándar** de la distribución muestral, **también llamado error estándar (SE)** es σ =
* **Si no tenemos** el **desvío estándar de la población**, podemos aproximarlo con el **desvío estándar muestral (s)**

σ =

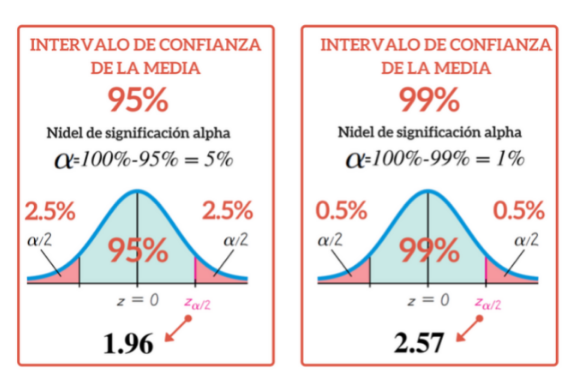
* Con el **Error Estándar** podemos **medir la dispersión** respecto a la **media muestral,** ajustando por el tamaño muestral. Esto nos **permite comparar estimadores** de la media **con diferentes tamaños muestrales**.
* El resultado es **asintótico (**límite con **n tendiendo a infinito)**; pero en la práctica con muestras específicas, n siempre va a ser finito. **El tamaño de la muestra necesario va a depender de la distribución original de la que partamos**.

**Algunos Estimadores Puntuales:**

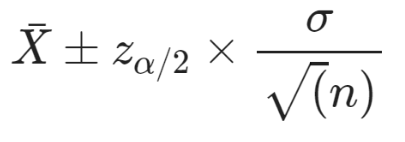
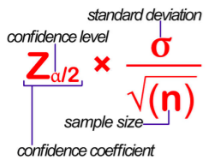


**Intervalos de Confianza:**

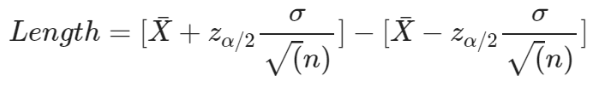
En la **estimación puntual**, la **media de la muestra** se calcula como **el valor estadístico** de la misma; pero en una **estimación por intervalos**, buscamos tener un **rango de valores posibles** para el parámetro **con un valor de confianza asociado**. Para esto es necesario obtener un **estadístico** con **distribución conocida que no dependa de parámetros desconocidos**.



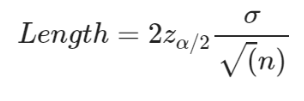
El **intervalo de confianza** nos indica que cada vez que tomemos una muestra y le calculemos el IC (**Intervalo de Confianza)**, la **probabilidad** de que el **estimador contenga el parámetro poblacional** será **(1-α)%**.

Un intervalo de confianza matemáticamente se define como: ; siendo el margen de error  cuánto por arriba o por debajo de la media muestral puede estar mi parámetro poblacional.

El rango del intervalo de confianza está definido por la resta entre el nivel superior e inferior del mismo; es decir:



Las se anulan y nos queda:

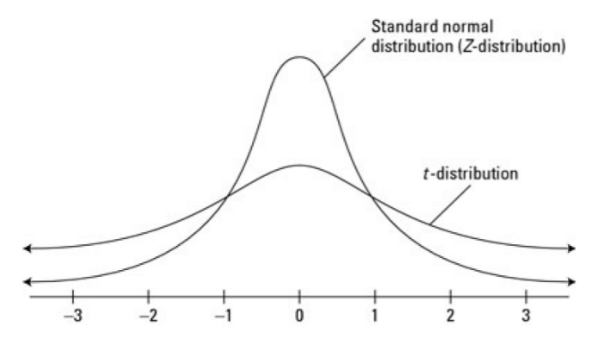


El **largo del intervalo de confianza** entonces **depende de**:

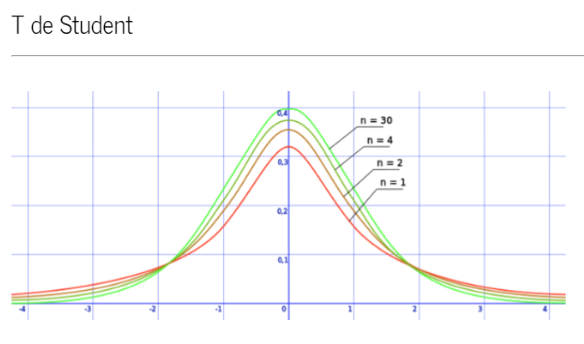
1. La **Varianza Poblacional** (σ).
2. El **Tamaño de la Muestra** (n).
3. El **Nivel de Confianza** deseado (α).

**Cuanto más estrecho** sea un **intervalo de confianza**, **más nos sirve** como resultado de un análisis.

Si se desconoce el desvío estándar poblacional, si bien con un n “grande” se puede estimar que S = σ; es **más conservador aproximar usando una distribución T de Student**.



Los grados de libertad que se usan en una distribución t son **v = n-1**.



A medida que n aumenta, t se va aproximando cada vez más a una distribución normal.

El **intervalo de confianza** **depende** de la **muestra**. Por ende, los límites del intervalo de confianza también son variables aleatorias. Como el intervalo de confianza no puede depender de parámetros desconocidos, se usa un estimador del desvío poblacional, o bien se asume como conocido el desvío poblacional.

**En Python**: Podemos calcular el valor Z para un alfa dado para una distribución normal con stats.norm.ppf de la biblioteca scipy.

IE:

alpha = 0.1

z\_alfa\_sobre\_2\_norm = stats.norm.ppf(alpha/2)

z\_alfa\_sobre\_2\_norm

Podemos calcular el valor Z para un alfa y unos grados de libertad v = n-1 dados para una distribución T de Student con stats.t.ppf de la biblioteca scipy.

IE:

n = 20

alpha = 0.1

grados\_libertad = n - 1

z\_alfa\_sobre\_2\_t = stats.t.ppf(alpha/2, grados\_libertad)

z\_alfa\_sobre\_2\_t