Inferencial:

**¿Alguna vez te preguntaste cómo se llegan a las conclusiones en estudios científicos?**

**Pongamos de ejemplo casos epidemiológicos.**

*¿Cómo se llega a la conclusión de que el ejercicio mejora la calidad de vida, fumar causa cáncer o que el uso de mascarillas reduce la probabilidad de contagio de algún virus?*

Quizá las respuestas son obvias. Pero cuando para llegar a conclusión debemos seguir un método científico.

Es imposible hacer un estudio en la población entera (en toda la humanidad). Por eso seleccionamos muestras significativas. A estas muestras se les aplica métodos INFERENCIALES para sacar una conclusión de toda la población bajo ciertos niveles de confianza.

En este curso vas a aprender lo que necesitas para adentrarte al mundo del diseño de experimentos, inferencia causal y muchas más cosas interesantes.

Diferencia con la descriptiva:

Busca entender el presente, entenderlo, limpiarlo, sacando estadísticos, índices.

1. Tendencia central de la variable.
2. Variabilidad.
3. Distribución.
4. Media, mediana, moda.

La inferencial:

Nos ayuda a determinar el futuro por medio de:

1. Muestreo
2. Intervalos de confianza
3. Validación de hipótesis
4. Evitar sesgos
5. Conclusiones que se obtienen a partir de los parámetros de la población de los datos
6. Estudio del grado de fiabilidad de los resultados obtenidos.

Paso a paso en el DS process:

1. Entender la distribución de los datos
2. Crear y validar la hipótesis
3. Hacer experimentos
4. Elegir los modelos predictivos adecuados según los datos

Estadísticos principales en la inferencia:

1. **Experimento:** Procedimiento que puede repetirse infinitamente y tiene un conjunto definido de resultaos posibles que corresponde al espacio muestral. Esto es, se tiene una variable aleatoria X a la cual se le hacen n número de experimentos que pueden ser de muchos tipos, pj Bernoulli.

Aleatorio: Si tiene más de un resultado posible. Una moneda o un dado al aire.

Determinista: Si sól tiene un resultado posible. Por ejemplo, ver si al lanzar una moneda va a caer a algún sitio, siempre va a ser P(determinista) = 1.

1. **Población y muestra:** La población es el total de todos los elementos posibles un denominador gigante, sobre el cua se pued hacer una extracción para obtener una muestra.   
   Condiciones:  
   - La muestra tiene que ser un número suficientemente grande de registros para ser estadísticamente significativo.

- Representación no sesgada de la información total.

Una población: Los habitantes de Sao Paulo. Una muestra: la población de entre 25 y 30 años.

1. **Evento:** Es cada uno de los posibles resultados dentro del espacio sigma el espacio muestral. Por ejemplo 1,2,3,4,5,6 para un lanzamiento de un dado.
2. **Variable:** cada una de las características que puede obtener diferentes valores. Hay de dos tipos:

* Cualitativas: atributos no numéricos, no medidas.
* Cuantitativas: números medibles que pueden ser discretos o continuos.

1. Probabilidad: la medida de qué tan posible es que ocurra un evento dentro del espacio muestral. EL análisis de estos eventos es lo que corresponde a la estadística.

**POBLACIONES**

La distribución simétrica, gaussiana de tipo normal de cmpana de Gauss. Moda = Mediana = Media. Simétrica, perfecta, con media en 1 y varianza 0.

**INTRODUCCIÓN AL MUESTREO Y AL TEOREMA CENTRAL DEL LÍMITE:**

Muestreo:

* Técnica para la selección de una muestra.
* Se obtiene a partir de una población estadística.
* La selección tiene que ser aleatoria y se espera que sus propiedades sean extrapolables a la población.
* No tiene que haber sesgos

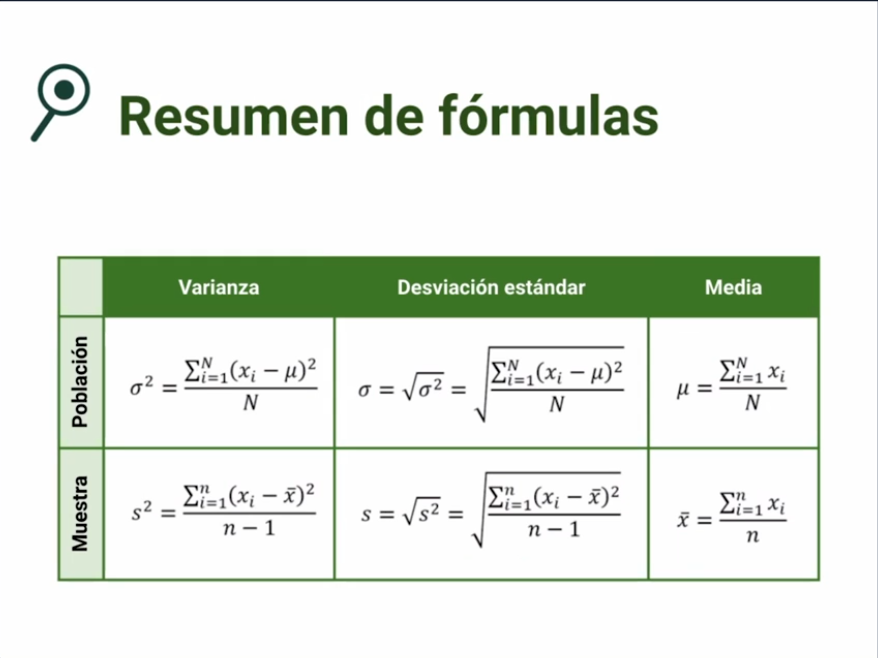
TEOREMA CENTRAL DEL LÍMITE:

Aleatorio simple: método de selección de ciertas unidades sacadas de una población de manera que cada una de las muestras tenga la misma posibilidad de ser elegida; por ejemplo: la lotería.

Sistemático: método de selección al azar y a continuación se eligen el resto siguiendo intervalos irregulares. Ejemplo: dar un premio cada 100 personas que llegan a una inscripción hasta llegar a un total de 1000 inscritos.

Estratificado: método de selección de ciertas unidad por segmentos exclusivos y homogéneos y a continuación se elige una muestra aleatoria simple de cada segmento. División por edades.

TLC: Todo tiende a ser normal siempre que la población sea lo suficientemente grande.



Intervalos de confianza\_

* Un par o pares de números entre los cuales se estima que estará cierto valor desconocido respecto de un parámetro poblacional con un determinado nivel de confianza.
* Son simétricos respecto a la media.

68%, 95%, 99.7%...

99% es muy estricto. Un 68% es un intervalo más abierto. En un estudio médico, sería más útil uno altísimo de certidumbre, pero si es un muestreo, podría ser del 68%

Nivel de significación:

* Es también llamado alfa, que es el límite para juzgar si un resultado es o no estadísticamente significativo.
* Si el valor de significación es menor que el nivel de significación, el resultado es que es estadísticamente significativo.

Cálculo del intervalo de confianza:

1. Conocer la media y la varianza.
2. No conocerlo: Entre qué valor está la población mínimo y máximo. Si es al 95%, los datos por fuera son el 2.5% al lado izquierdo y el 2.5 al lado derecho.

Siendo así, si s pone el intervalo de confianza más el desajuste más bajo estaríamos hablando del 97.5% de la tabla de values y probabilidades para la distribución estándar normal. La suma entre el 0.95 y el 0.027, siendo que en la tabla se buscaría el 0.975 que equivale a Z en 1.9 en la vertical y en Z 0.06 en la horizontal, o sea el índice 1.96. Entonces:

**IC al 95% = (-1.96,1.96)** nuestro valor está entre uno u otro, es decir, cualquier valor estará máximo entre -1.96 y 1.96, con media 0 y varianza 1.

Ejercicio. Hay una distribución N(28,4). ¿Cuál es el intervalo de confianza al 80%?

RTA: 1.2+0.09 = 1.28. Se ajusta y aproxima por debajo.

Si z = x-Mu/Desviación:

-1.28 = (x1 – 28)/4 ; 1.28 = (x2 – 28)/4

Siendo así X1 = 22.88, X2 = 33.12

IC 89% = (22.88; 33.12)

**Pruebas de hipótesis**

Son también llamadas pruebas de significancia. Ayudan a juzgar si existe una diferencia significativa entre el tamaño de la muestra y el parámetro general. Es decir, vamos a comprobar sobre una teoría si hay una diferencia entre los casos.  
Por ejemplo: H1 = Si en las ciudades donde hace más frío la gente vive más.  
La hipótesis nula sería: H0= Viven igual independientemente del clima.  
Rechazando la hipótesis: H2 = Si la gente vive más en los países calientes.

1. Establecer una hipótesis nula H0 y una hipótesis alternativa H1.
2. Seleccionar el nivel de significancia.
3. Seleccionar el estadístico de prueba.
4. Formular la regla de la decisión. Si al 68% no pasa nada, no es significativo, podríamos no tomarlo si lo decidiéramos. Si en cambio al 99% si pasa, habría que tomar una decisión si o si.
5. Interpretar los resultados y tomar una decisión.

**Tipos de pruebas de hipótesis:**

Distribución t-student: Se usa para estimar una ***media de población*** normalmente distribuida a partir de una muestra pequeña que sigue una distribución normal y de la que desconocemos la desviación estándar.

1. No conozco la desviación
2. Tengo la media muestral chirriquitica
3. Me interesa la media poblacional
4. Uso la t-student para “imputarlo”.

**Coeficiente de correlación Pearson:**

Se usa para medir la correlación lineal entre dos variables cuantitativas

**ANOVA: Análisis de la varianza:**

Se usa para comparar las varianzas entre las medias o el promedio de diferentes grupos.

**Errores frecuentes al hacer realizaciones de pruebas de hipótesis**

Contexto:

Las conclusiones a las que llegamos se basan en una muestra, así que podríamos estar equivocados.

Decisiones correctas:

1. Rechazar H0 cuando es falsa.
2. No rechazar H0 cuando es verdadera.

Decisiones incorrectas:

1. Rechazar H0 cuando es verdadera
2. No rechazar la H0 cuando es falsa.

Ejemplo: Hay dos medicamentos que tienen la misma eficacia (H0). H1-> no tienen la misma eficacia

Error tipo I (De omisión): Concluir que los medicamentos son muy diferentes, siendo que no lo son.

Error tipo II: concluir que no hay una diferencia significativa entre ambos, cuando si lo hay. Sería muy peligroso.

**TÉCNICA DEL BOOTSTRAPPING**

* Método de remuestreo de datos dentro de una muestra aleatoria.
* Se usa para hallar una aproximación a la distribución de la variable analizada.
* Muy útil en muestras pequeñas o en distribuciones muy sesgadas.

**VALIDACIÓN CRUZADA:**

Técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.

Se usa al FINAL de un análisis para demostrar que los datos de prueba son independientes de los datos de prueba.