Curso de

**Estadística Inferencial para Data Science e Inteligencia Artificial**

Sílvia Ariza Sentís

**¿Para quién es este curso?**

Estudiantes que saben de análisis descriptivo en Python y quieren explorar el ámbito inferencial. 



**Estadística descriptiva vs inferencial**

**Estadística descriptiva vs inferencial**

**DESCRIPTIVA**

Parte de la estadística que arregla los datos de forma que puedan ser analizados e interpretados.

**INFERENCIAL**

Parte de la estadística que busca predecir o deducir características o resultados esperados de una población, basados en los datos obtenidos de una muestra de esa población.

**Estadística descriptiva**

Nos ayuda a determinar:

● La **tendencia** central de una

variable. 

● La **variabilidad** de una variable.

**Media**

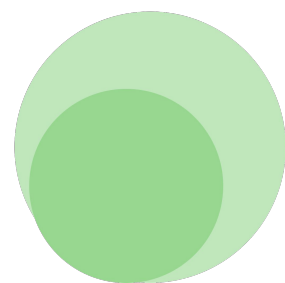
**Mediana**

● La **distribución** de una variable.

**Moda**

**Unimodal Bimodal Multimodal**

**Estadística inferencial**

Nos ayuda a determinar: 

● Muestreo

**Población**

● Intervalos de

confianza

● Validación de hipótesis

● Evitar sesgos

**Muestra**

**Inferencia estadística**

● Conclusiones que se obtienen sobre los parámetros de la población de datos.

● Estudio del grado de fiabilidad de los resultados extraídos del estudio.

**Uso en data science y machine learning**

Tanto en un análisis como en un modelo predictivo, la estadística inferencial servirá para:

● Entender la **distribución** de nuestros datos. ● Crear y validar **hipótesis**.

● Hacer **experimentos**.

● Elegir los **modelos predictivos** adecuados según los datos.

**Estadísticos principales**

**Experimento**

Procedimiento que puede repetirse infinitamente y tiene un conjunto bien definido de resultados posibles, conocido como espacio muestral.

● **Aleatorio**: si tiene más de 

un resultado posible.

● **Determinista**: si solo tiene

un resultado posible.

**Población y muestra Muestra:** subconjunto de datos perteneciente a una población.

Condiciones:

● **Número** suficiente de registros para ser

estadísticamente

significativo.

● Representación **no sesgada** de la información total.

**Población**

**Muestra**

**Evento**

Cada uno de los 

posibles resultados de un experimento.

**Variable**

Es una característica que puede obtener diferentes valores.Tipos:

● Cualitativas: atributos (no medibles).

● Cuantitativas: números (medibles).

○ Discretas

○ Continuas

**Color de pelo =** 

**castaño**

**Estatura =**

**1.68 m**

**Peso =**

**61 kg**

**Probabilidad**

Mide **qué tan posible** es que ocurra un evento determinado.

El análisis de los eventos probabilísticos se denomina **estadística**.

**Probabilidad condicionada**

Posibilidad de que ocurra un evento como consecuencia de que **otro evento** haya sucedido.

**A B** 

**A∩B**

**Poblaciones normales**

**Distribución normal**

● Distribución normal = Distribución de Gauss. ● Su moda = su media = su mediana. ● Es simétrica.

● Tiene forma de campana.

**y** 

**0 x**

**Ejemplos de** 

**población normal**

● Calorías ingeridas y peso.

● Presión sanguínea.

● Tamaño de los coches producidos por una máquina.

**Introducción al**

**muestreo y teorema del límite central**

**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.

**Tipos de muestreo**

**Aleatorio simple**

Método de selección de ciertas unidades sacadas de una población de manera que cada una de las muestras tiene la **misma** 

**probabilidad** de ser elegida.

**Ejemplo:** lotería.

**Sistemático**

Método de selección de ciertas unidades al **azar** y, a continuación, se eligen el resto siguiendo **intervalos regulares**. 

**Ejemplo:** dar un premio cada cien personas que hagan una

inscripción hasta llegar a un total de mil inscritos.

**Estratificado**

Método de selección de ciertas unidades por segmentos exclusivos y homogéneos y, a continuación, se elige una muestra aleatoria simple de cada segmento. 

**Ejemplo:** división por edades.

**Teorema del límite central**

**Teorema del límite central**

Teoría estadística que

establece que, dada una

muestra aleatoria 

suficientemente grande de la

población, la distribución de

las medias seguirá una

distribución normal.**1 - α**

**α/2 α/2**

**Funciones de**

**muestreo en Python**

****

**Muestreo estratificado en Python**

****

**La media muestral**

**Media, moda y mediana**

**Media**

Suma de los datos dividida entre la cantidad de datos. 

**media**

**Moda**

El dato que más se repite.

**moda**

**Mediana**

Es el dato que está en el centro de todos.

**50% 50%mediana**

**Media muestral**

**Media muestral**

● Media aritmética = promedio = media.

● Valor que se obtiene de sumar un conjunto de valores cuantitativos y dividirlo por el número total de sumados.



**Media muestral**

X

● Media muestral ≠ media poblacional

● **Ejemplo:** estimación puntual de la edad promedio de una población.



**Cálculo** 

Calcula la edad promedio de una clase donde los estudiantes tienen las siguientes edades: 28, 24, 25, 23, 38, 52.

X = 28+24+25+23+38+52 6

= 31.7 años

**Varianza y desviación estándar muestral**

**Varianza y** 

**desviación estándar**

● Indica qué tan dispersos están los datos respecto a la media. **● La desviación estándar es la raíz cuadrada de la varianza.** ● Ejemplo: edades de la población de una ciudad.

**Desviación baja Desviación alta**

****

**Cálculo de la varianza y desviación estándar** 

**Muestral Poblacional**

Desviación estándar de una clase donde los estudiantes tienen las siguientes edades: 28, 24, 25, 23, 38, 52.

Sabemos que la edad promedio es 31.7 años.

**Cálculo de la varianza y desviación estándar**

Varianza muestral =

(28-31.7)^2+(24-31.7)^2+(28-31.7)^2+(24-31.7)^2+(28-31.7)^2+(24-31.7)^2 5

Desviación estándar muestral = 43.8 = 6.62

= 43.8

**Resumen de fórmulas**

**Varianza Desviación estándar Media**

**n**

**ó**

**i**

**c**

**a**

**l**

**b**

**o**

**P**

**a**

**r**

**t**

**s**

**e**

**u**

**M**

**Varianza y desviación estándar muestral en Python**

****

**Intervalos de confianza**

**Intervalos de confianza**

● Un par o varios 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%**

**Nivel de significación**

● El nivel de significación o alfa es el nivel límite para juzgar si un resultado es o no es estadísticamente significativo.

● Si el valor de significación es menor que el nivel de significación, el resultado es estadísticamente significativo.

**No rechazo Ho 95%** 

**Rechazo 5%**

**Interpretación** 

**de un resultado**

**Intervalo de confianza del 95%:**

Sabemos que con un 95% de certeza las edades de las personas que esquían están entre dos valores.

**Cálculo de intervalos de confianza**

**Ejercicio I:** 

**nivel de significación** El valor de alfa es del 5%.

2.5% 2.5% 

**95% de**

**confianza**

**-Zα/2 Zα/2**

**Consideraciones al buscar en la tabla** 

α/2 

**% Nivel de**

**confianza**

Zα/2

Confianza + α/2

2.5%

**Ejemplo al 95% de confianza**

2.5% 

**95%**

0.975 Zα/2

**95%**

Zα/2

95% + 2.5% 97.5%

**Búsqueda en la tabla**

**Resultado**

**95%** -1.96 0 1.96

IC95% = (-1.96 ; 1.96)

**Ejercicio II**

La duración en días de un cepillo de dientes se ajusta a la distribución normal (28, 4). ¿Cuál es el intervalo de confianza al 80%? μ = 28

σ = 4

**80%** 

**de confianza**

Zα/2 -Zα/2

**Nivel de significación** El valor de alfa es del 20%.

**10% 10%**

**80%**

**de confianza**

Zα/2 -Zα/2



**10%** 

**80%**

**Consideraciones al buscar en la tabla**

**10%**

**80%**

Zα/2

0.90

Zα/2

80% + 10% 90%

**Búsqueda en la tabla**

**Resultado 80%**

-1.28 0 1.28 **Z N(0,1)**

Este es el caso de media 0, ahora tenemos que convertirlo a la media 28 del ejercicio con la fórmula: 

**Conversión** -1.28 = (x1-28)/4 x1 = **22.88** 

**80%**

1.28 = (x2-28)/4 x2 = **33.12**

**22.88 28 33.12 IC80% = (22.88; 33.12)**

**Cálculo de intervalo de confianza**

**en Python**

****

**Pruebas de hipótesis**

**Prueba de hipótesis**

La prueba de hipótesis o prueba de significación ayuda a juzgar si existe una diferencia significativa entre el tamaño de la muestra y el parámetro general.

Área de rechazo de H0 Área de rechazo de H0 

**Área de NO rechazo**

**de H0**

**Pasos a seguir**

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 decisión.

5) Interpretar los resultados y tomar una decisión.

**Tipos de pruebas de hipótesis**

**Distribución t de 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.

**k=1** 

**k=2**

**k=3** 

**k=1o**

**k=infini**

**Coeficiente de Pearson**

Se usa para medir la dependencia lineal (correlación) entre dos variables aleatorias cuantitativas.

Tabla de interpretación de resultados:



**Análisis de la varianza (ANOVA)** 

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

**A B C D**

Observaciones

Variable independiente

**Tipos de errores**

**Contexto**

Las conclusiones a las que llegamos se basan en una muestra, por lo que podemos equivocarnos.

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 H0 cuando es falsa.

**Tipos de errores**

**H0 verdadera H0 falsa**

**Rechazamos H0Error tipo I**

P(Error tipo I) = α **Decisión correcta**

**No rechazamos H0 Decisión correcta Error tipo II** P(Error tipo II) = β

**Ejemplo**

**HIPÓTESIS**:

● Hipótesis nula (H0): μ1= μ2

Los dos medicamentos tienen la misma eficacia.

● Hipótesis alternativa (H1): μ1≠ μ2

Los dos medicamentos no tienen la misma eficacia.

**ERROR TIPO I**: concluir que los dos medicamentos son muy diferentes cuando no lo son.

**ERROR TIPO II**: concluir que no hay una diferencia significativa entre ambos medicamentos. Muy peligroso.

**Pruebas de hipótesis en Python**

****

**Bootstrapping**

**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. 

**Bootstrapping en Python**