Métodos Estadísticos y Simulación

Unidad 1: Conceptos generales y Análisis exploratorio de datos

Mg. J. Eduardo Gamboa U.

2025-03-30

Presentación

Datos del docente

Mg. Jesús Eduardo Gamboa Unsihuay
Correo: jgamboa@lamolina.edu.pe

Datos del curso

EP7192 Métodos Estadísticos y Simulación
Link de la sesión: Zoom
Repositorio: GitHub

Sumilla

El curso de Métodos Estadísticos y Simulación pertenece al área de formación específica y su naturaleza es teórico-práctica. Su propósito es desarrollar competencias en el estudiante para reconocer y comprender los conceptos utilizados en estadística. Comprende las siguientes unidades: Conceptos generales y Análisis Exploratorio de Datos, Probabilidades, Variables Aleatorias, Distribuciones de probabilidad y muestrales, Intervalos de Confianza, Pruebas de hipótesis, Pruebas No Paramétricas. Los resultados del análisis de los datos se obtendrán utilizando lenguajes de código abierto como R y/o Python.

Evaluaciones

- Evaluación 1 (individual):
  - Evalúa Unidades 1 y 2, incluye trabajos cortos 1 y 2
  - Semana 6 (domingo 04 de mayo)
  - Ponderación: 20%
- Evaluación 2 (individual):
  - Evalúa Unidades 1, 2, 3 y 4, incluye trabajos cortos 3 y 4
  - Semana 11 (domingo 08 de junio)
  - Ponderación: 20%
- Evaluación 3 (individual):
  - Evalúa todo el curso, incluye trabajo corto 5
  - Semana 17 (domingo 20 de julio)
  - Ponderación: 30%

- Trabajo de investigación (grupal, de 3 a 5 integrantes)
  - Avance 1: Semana 7 (viernes 09 de mayo) - 5%
  - Avance 2: Semana 14 (viernes 27 de junio) - 5%
  - Entrega final: Semana 17 (domingo 20 de julio) - 10%
- Asistencia a lo largo del curso: 10%

Unidades del curso

1. Conceptos generales y análisis exploratorio de datos
2. Probabilidades
3. Variable aleatoria
4. Distribuciones de probabilidad y muestrales
5. Intervalos de confianza
6. Pruebas de hipótesis
7. Pruebas no paramétricas

Conceptos generales

Estadística

La Estadística es la ciencia del aprendizaje a partir de los datos, basada en la teoría de probabilidades, que aplica principios matemáticos para la recolección, organización, análisis, modelado y visualización de datos, con el objetivo de extraer conocimiento confiable, representar fenómenos sujetos a variabilidad e incertidumbre y realizar inferencias sobre una población a partir de una muestra.

División de la estadística

Estadística Descriptiva: Se ocupa de la clasificación, descripción, simplificación y presentación de los datos. Comprende la elaboración de tablas de frecuencias, gráficos y el cálculo de medidas estadísticas.

Estadística Inferencial: Se ocupa de la estimación y prueba de hipótesis de los parámetros poblacionales, a partir de una o más muestras aleatorias extraídas de la(s) población(es) de interés

Población

Es el conjunto de unidades elementales con características similares. El estudio de toda la población constituye un censo.

Ejemplos:

1. Todos los pacientes asmáticos del Perú que reciben atención en hospitales públicos.
2. Todos los estudiantes de primaria de colegios privados del país.
3. Toda la producción mensual de botellas de gasesosa de una marca.
4. Todas las parcelas de maíz sembradas en la región Ayacucho.

Muestra

Es un subconjunto de unidades que forman parte de la población. El proceso de selección de una muestra se le llama "muestreo". Una muestra representativa cumple con las siguientes condiciones:

- Debe haber sido obtenida al azar.
- Su tamaño y sus elementos deben haber sido seleccionados aplicando un método de muestreo.

Ejemplos:

1. 320 pacientes asmáticos del Perú que reciben atención en hospitales públicos
2. 200 estudiantes de primaria de colegios privados del país.
3. 100 botellas de gaseosa de una marca producidas en un mes.
4. 30 parcelas de maíz sembradas en la región Ayacucho.

Unidad elemental

Es un elemento particular que forma parte de la población o muestra.

Ejemplos:

1. Un paciente asmático del Perú que recibe atención en un hospital público.
2. Un estudiante de primaria de un colegio privado del país.
3. Una botella de gaseosa de una marca producida en un mes.
4. Una parcela de maíz sembrada en la región Ayacucho.

Variables

Son las características que toman diferentes valores cuando son evaluadas las unidades elementales de una población o muestra.

Se le representa con las últimas letras mayúsculas del alfabeto: X, Y, Z, W, P, T, X1, X2, Y1, etc.

Pueden ser de dos tipos: Cualitativas y Cuantitativas

Variables Cualitativas.

Poseen determinadas características o atributos, y se definen mediante categorías, orden o clase. Pueden ser: Nominales, y Ordinales o Jerárquicas.

Variable Cualitativa Nominal.

Sólo representan atributos y cuyos valores no se pueden clasificar a partir de un criterio de orden o jerarquía.

Ejemplos:

1. Sexo de un estudiante (Masculino o Femenino)
2. Estado civil del solicitante (Soltero, Casado, Viudo o Conviviente)
3. Ubicación de la sucursal de un banco (Sur, Centro o Norte)
4. Estado del agua (Sólido, Líquido o Gaseoso)

Variable Cualitativa Jerárquica u Ordinal

Es aquella donde sí se puede establecer un criterio de orden o jerarquía a partir de los valores cualitativos que asume.

Ejemplos:

1. Nivel de satisfacción con el comedor universitario (Bueno, Regular o Malo)
2. Nivel educativo alcanzado (Sin instrucción, Primaria, Secundaria o Superior)
3. Clasificación socioeconómica según SISFOH (No pobre, pobre o pobre extremo)
4. Rango de ingreso familiar (500 - 2500, 2501 – 5000, 5001 – 7500, Más de 7500)

Variables Cuantitativas.

Son aquellas que se representan mediante un valor numérico y por lo tanto pueden realizarse operaciones matemáticas. Pueden ser: Continuas o Discretas.

Variable Cuantitativa Discreta.

Toma valores que pertenecen a los números enteros y cumplen con la condición de que entre un valor entero y su valor consecutivo no existen valores intermedios. Las observaciones cuantitativas discretas se registran por conteo.

Ejemplos:

1. Número de pacientes atendidos por intervalo de 15 minutos en el área de emergencias de un hospital de Essalud
2. Número de predios inscritos a nombre de un agricultor del valle del Mantaro.
3. Cantidad diaria de reclamos presentados por los clientes de una empresa de telefonía.
4. Número de pasajeros que abordan una línea de bus en un paradero de la Avenida Javier Prado entre las 6:30 a.m. y 7:00 a.m.

Variable Cuantitativa Continua.

Pueden tomar cualquier valor numérico dentro de un intervalo continuo en el conjunto de números reales. Generalmente se utiliza un instrumento de medición (balanza, termómetro, test, escala, cronómetro, wincha, etc.) para generar sus valores y poseen una magnitud (peso, altura, distancia, etc.) y unidades de medida (kg, m, km, etc.).

Ejemplos:

1. Costo mensual de inventario en almacén de medicamentos (soles)
2. Longitud corporal del langostino, medido desde el rostro hasta el extremo de la cola (cm)
3. Tiempo transcurrido desde que se toma el pedido hasta que se entrega al cliente de una pizzería (minutos)
4. Consumo mensual de electricidad en kilovatios-hora (kWh).

Medidas estadísticas

Se calculan a partir del conjunto de datos de una variable, tienen como finalidad describir el comportamiento de una población o de una muestra. Se clasifican en:

Parámetro

- Describe el comportamiento de una variable en la población.
- Es calculado con los datos de toda la población.
- Es un valor constante.
- Se representan con letras griegas.

Ejemplos

1. Suponga que la estatura promedio de todos los estudiantes de la UNALM es 168.4 cm ($\mu = 168.4$).
2. El Censo Nacional 2017: XII de Población y VII de Vivienda y y III de Comunidades Indígenas indica que el 5.8% de la población de 15 años o más no sabe leer ni escribir ($\pi = 0.058$).
3. Suponga que la verdadera proporción de personas mayores de 15 años que sufren de hipertensión arterial en Perú es $\pi = 0.25$.

Estimador, Estadístico o Estadígrafo

- Describe el comportamiento de una variable en la muestra.
- Es calculada con los datos obtenidos de una muestra.
- Los estadísticos sirven para estimar a los parámetros.
- Se representan con letras latinas.

Ejemplos

1. En una muestra de 300 estudiantes de la UNALM, la estatura promedio es 166.8 cm, es decir $\overline{x} = 166.8$
2. En la Encuesta Nacional de Hogares (ENAHO) 2023, se señala que 4.8% de las personas de 15 años o más no sabe leer y ni escribir, es decir $p = 0.048$.
3. En el Perú, existen 5.5 millones de personas mayores de 15 años que sufren de hipertensión arterial ($p = 0.221$), según medición de la Encuesta Demográfica y de Salud Familiar (ENDES).

Medidas de tendencia central

Permiten resumir y representar los datos mediante un valor localizado en la parte central de la distribución de los datos. Son medidas de tendencia central:

1. Media
2. Mediana
3. Moda

Media

- Es la suma de todos los valores dividida entre el número total de observaciones.
- Amplio uso descriptivo e inferencial.
- Es más representativa cuando los datos son simétricos.
- Sensible a los valores extremos (outliers).
- Se simboliza como μ (parámetro) o $\overline{x}$ (estimador).

Mediana

- Es el valor central cuando los datos están ordenados.
- Es robusta ante valores extremos.
- También se recomienda su uso cuando hay datos asimétricos.
- Se simboliza como Me (parámetro) o me (estimador).

Moda

- Es el valor que más se repite en el conjunto de datos.
- Es útil tanto para datos cualitativos como cuantitativos.
- No se ve afectada por valores extremos.
- Puede haber más de una moda (bimodal o multimodal), o podría no haber.
- Se simboliza como Mo (parámetro) o mo (estimador).

Medidas de posición

- Dividen un conjunto de datos previamente ordenado.
- Conservan las mismas unidades que la variable analizada.
- El percentil P_q deja $q\%$ de datos por debajo y $(100-q)\%$ por encima.
- Dependiendo del número de divisiones, reciben diferentes nombres:
  - Cuartiles: Dividen en 4 partes (Q_1,Q_2,Q_3)
  - Quintiles: Dividen en 5 partes
  - Deciles: Dividen en 10 partes
  - Percentiles: Dividen en 100 partes (P_1,...,P_{100})
- Como no se basan en todos los datos directamente (como la media), no se ven **tan** afectadas por outliers.
- En muestras pequeñas, los percentiles y deciles pueden no ser muy estables o informativos.
- Son útiles descriptivamente, pero no suelen emplearse en inferencia estadística.

Medidas de dispersión

Las medidas de dispersión indican qué tan dispersos o concentrados están los datos respecto a una medida central (como la media o la mediana). Es decir, muestran el grado de variabilidad o heterogeneidad en un conjunto de datos. Son medidas de dispersión:

- Rango
- Rango intercuartil
- Varianza
- Desviación estándar
- Coeficiente de variación

Rango

- Diferencia entre el valor máximo y el mínimo.
- Muy sencillo de calcular.
- Sensible a valores extremos.
- No evalúa la dispersión interna de los datos.
- Se simboliza como R (parámetro) o r (estimador).

Rango intercuartil

- Diferencia entre el tercer y primer cuartil.
- Mide la dispersión del 50% central de los datos.
- Robusto ante outliers.
- No capta la variabilidad en los extremos del conjunto de datos.
- Puede ocultar información importante sobre la dispersión total.
- Se simboliza como RIC (parámetro) o ric (estimador).

Varianza

- Promedio de las desviaciones cuadráticas respecto a la media.
- Varianza poblacional:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

- Varianza muestral:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

- Ampliamente usada en inferencia estadística (por ejemplo: ANOVA).
- Cada dato aporta a la magnitud de dispersión.
- Expresada en unidades al cuadrado, por eso no es tan interpretable directamente.
- Está afectada por outliers.

Desviación estándar

- Raíz cuadrada de la varianza.
- Mide cuánto se alejan, en promedio, los datos de la media.
- Desviación estándar poblacional:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

- Desviación estándar muestral:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

- Cada dato aporta a la magnitud de dispersión.
- Se expresa en las mismas unidades que la variable original, lo que la hace más interpretable que la varianza en contextos descriptivos.
- Está afectada por outliers.

Coeficiente de variación

- Es una medida de dispersión relativa (no tiene unidades)
- Es la razón entre la desviación estándar y la media aritmética de un conjunto de observaciones.
- Coeficiente de variabilidad poblacional:

$$CV = \frac{\sigma}{|\mu|} \times 100\%$$

- Coeficiente de variabilidad muestral:

$$CV = \frac{s}{|\overline{x}|} \times 100\%$$

- Al dividir entre una media muy pequeña, el CV puede volverse exageradamente grande o inestable.

Medidas de asimetría

Las medidas de asimetría indican si la distribución de los datos es simétrica o presenta sesgo.

- Distribución simétrica: La curva es equilibrada respecto a la media
- Distribución asimétrica positiva. La distribución de los datos presenta una curva con cola a la derecha. La mayor cantidad de los datos son de menores valores.
- Distribución asimétrica negativa. La distribución de los datos presenta una curva con cola a la izquierda. La mayor cantidad de los datos son de mayores valores.

Coeficiente de Fisher-Pearson

- Es la medida clásica de asimetría
- Es un estimador basado en momentos:

$$as_{FP} = \frac{1}{n} \times \frac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{s^3}$$

- Utiliza toda la información de los datos
- Sensible a outliers

Coeficiente de asimetría de Bowley

- Es una medida de asimetría del 50% central de los datos.
- Es un estimador basado en cuartiles:

$$as_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

- Robusto frente a valores extremos
- Fácil interpretación visual con boxplot

Tablas de frecuencia

Una tabla de frecuencia es una herramienta estadística que organiza y resume un conjunto de datos, mostrando cuántas veces (frecuencia) ocurre cada valor o grupo de valores en una variable.

Está compuesto de:

- Frecuencia absoluta (f_i): Número de veces que aparece un valor o clase.
- Frecuencia absoluta acumulada (F_i): Suma acumulada de frecuencias hasta una clase determinada.
- Frecuencia relativa (fr_i): Proporción respecto al total de datos. Solo para valores cualitativos jerárquicos o cuantitativos.
- Frecuencia relativa acumulada (Fr_i): Suma acumulada de las frecuencias relativas. Solo para valores cualitativos jerárquicos o cuantitativos.

Gráficos

Los gráficos estadísticos son representaciones visuales de datos que permiten identificar tendencias, patrones, comparaciones y distribuciones de forma rápida y comprensible.

Tipo de variable	Gráfico recomendado	Uso principal
Cualitativa nominal	Gráfico de barras, gráfico de sectores (pastel)	Comparar frecuencias por categoría
Cualitativa ordinal	Gráfico de barras ordenadas	Comparar y respetar el orden lógico de las categorías
Cuantitativa discreta	Gráfico de barras	Mostrar frecuencia de valores individuales
Cuantitativa continua	Histograma, polígonos de frecuencia, boxplot	Visualizar la distribución agrupada en intervalos
Variable temporal (serie de tiempo)	Gráfico de líneas	Observar la evolución en el tiempo