Elementos de Probabilidad y Estadística

Diplomatura en Ciencias de Datos

2024

Dr. Matías Hisgen – Lic. Celine Cabás – Lic. Fernando Álvarez

FACENA - UNNE

Fenómenos Aleatorios e Incertidumbre

- Un ***fenómeno aleatorio*** (FA) es un proceso de la realidad que produce ***resultados inciertos***, es decir, que no son ***predecibles*** con exactitud.

- Las ciencias en general modelan estos fenómenos o procesos de la realidad, tratando de ***explicar*** sus causas y/o ***predecir***.

- La Teoría de la Probabilidad aporta modelos probabilísticos útiles para modelar tales FA.
- Dichos modelos se conocen en la jerga como “Procesos Generadores de Datos” (en inglés, “Data Generating Process” o DGP),

Datos y Estadística

- Los FA producen resultados, los que son medidos y registrados mediante métricas o codificaciones alfa-numéricas, las que son almacenadas en ***Bases*** (o conjunto) ***de Datos***.

- En términos prácticos, la Estadística es la ciencia de recolectar, organizar, describir, comparar e interpretar **conjuntos de datos**.

- En ***Ciencias de Datos***, la Estadística aporta metodologías de "Aprendizaje Estadístico", extrayendo información de conjuntos de datos, con el propósito de ayudar a una **toma de decisiones** más efectiva.

Tipos de análisis estadístico

- **Análisis descriptivo (estadística descriptiva):** métodos para organizar, resumir y presentar datos de manera informativa.
- *Ejemplo: un sondeo de opinión encontró que 80% de los clientes se encuentran satisfechos con la marca de celular que vende la empresa. La estadística "80" describe el número de personas satisfechas por cada 100 clientes.*

- **Análisis de inferencia estadística**: proceso que conduce a una estimación, predicción o generalización sobre una característica (no observada) de la población, con base en una muestra (observada).
- Ejemplo: las plataformas de Streamming monitorean la popularidad de sus canales de video para (predecir/estimar) las preferencias de sus usuarios.

Tipos de datos estadísticos según su Fuente

- **Datos de registros administrativos:** datos que surgen de registros dentro de organizaciones tanto públicas (ANSES, AFIP) como privadas (empresas, instituciones).
- *Ejemplo: Registro de ventas diarias en una empresa, información de remuneración bruta de trabajadores asalariados (ANSES-SIPA).*

- **Datos recolectados mediante muestras:** *Ejemplo: Encuesta Permanente de Hogares para estimar tasas de Desempleo y Pobreza.*

- **Datos Censales:** Se releva exhaustivamente la información de interés de toda la Población censada.

Tipos de variables aleatorias (v.a.)

- **Variable Cuantitativa:** suceptible de medición vía unidades de medida numéricas, por lo que sus registros son comparables numéricamente.
- *Ejemplos:* temperatura ambiente, presión arterial, n° de integrantes de un hogar, monto anual de ventas, cotización de una acción en la bolsa.

- **Variable Cualitativa:** la característica o variable que se estudia no es susceptible de medición vía unidades de medida comparables.
- *Ejemplos:* tipo de personería jurídica, afiliación sindical, lugar de nacimiento, tipo de automóvil (deportivo, utilitário, familiar) .

Tipos de v.a. cuantitativas

- **Variable Cuantitativa Continua:** pueden tomar cualquier valor dentro de un intervalo específico (métricas altamente divisibles).
- *Ejemplos:* precio de un producto, el tiempo que tarda un envío de un producto, el saldo en cuenta corriente.

- **Variable Cuantitativa Discreta:** contienen un número acotado de valores (generalmente números enteros).
- *Ejemplos:* el número de habitaciones de una casa, número de menores en el hogar, cantidad de días lluviosos en un mes.

Tipos de v.a. cualitativas

- **Variable Cualitativa Nominal:** los datos sólo se puede clasificar en categorías, las cuales no se pueden ordenar bajo ningún criterio (relevante a nivel práctico).
- *Ejemplos:* Género de los clientes, afiliación sindical de los trabajadores.

- **Variable Cualitativa Ordinal:** involucra datos que se pueden ordenar bajo algún criterio, pero los valores de los datos no tienen significado métrico.
- *Ejemplos:* categorías de Riesgo de Bonos Soberanos (Riesgo Bajo, Medio y Alto), preferencia por un tipo de producto (me gusta, me da igual, no me gusta).

Descripción de los datos: distribuciones de frecuencias

- **Distribución de frecuencias:** ***agrupamiento*** de datos en ***categorías*** que muestran el ***número de observaciones*** en cada categoría mutuamente excluyente.

- **Variable es discreta:** cada valor que toma puede ser considerado como una categoría.

- **Variable Continua:** cada categoría comprende un intervalo continuo de valores (*Intervalo de Clase*).

Ejemplo: variable discreta

- De una muestra de 105 casas vendidas se presenta tabla de frecuencias para el n° de habitaciones (variable *cuantitativa discreta*)

Nº de Habitaciones	Frecuencia	Porcentaje	Frec. Acum.	Frec. Relativa
2	24	22.86	22.8	0.2286
3	26	24.76	47.62	0.2476
4	26	24.76	72.38	0.2476
5	11	10.48	82.86	0.1048
6	14	13.33	96.19	0.1333
7	2	1.90	98.10	0.019
8	2	1.90	100	0.019
Total	105	100		1

Ejemplo: variable continua

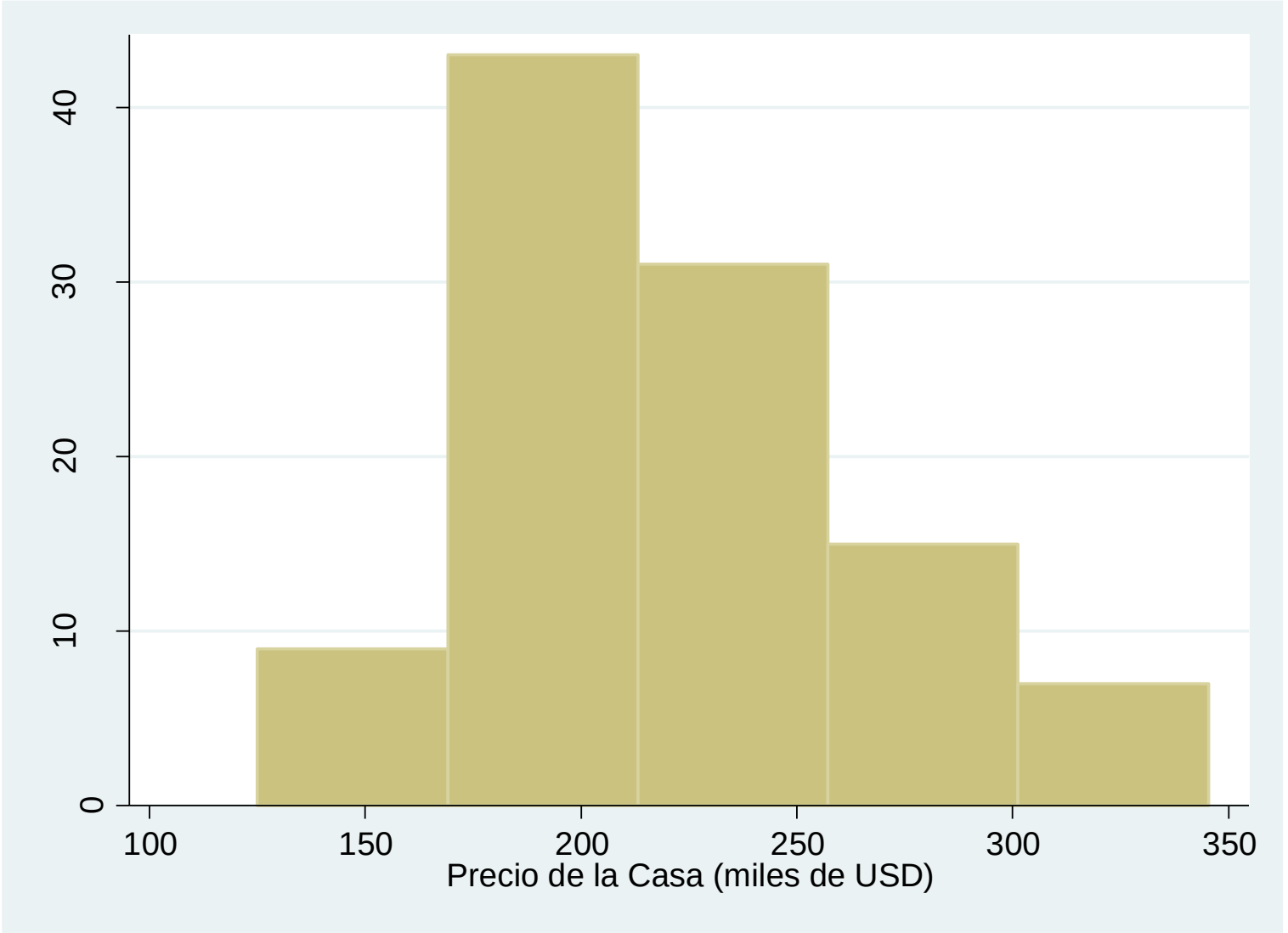
- Muestra de 105 casas vendidas: tabla de frecuencias para el Precio de venta (variable *cuantitativa continua*)

Precio	Frecuencia	Porcentaje	Frec. Acum.	Frec. Relativa
125-174	16	15.24	15.24	0.1524
175-224	46	43.81	59.05	0.4381
225-274	29	27.62	86.67	0.2762
275-324	11	10.48	97.14	0.1048
325-374	3	2.86	100.00	0.0286
Total	105	100		1

Presentación gráfica de una distribución de frecuencias: Histograma

- Las forma de gráfico más usada es el **histograma**, el cual grafica la distribución de frecuencias (absoluta o relativa) para cada intervalo de clase.
- **Histograma:** gráfica donde las clases se marcan en el eje horizontal y las frecuencias de clase en el eje vertical. Las frecuencias de clase se representan por las alturas de las barras y éstas se trazan adyacentes entre sí. También puede ser graficarse la frecuencia relativa, sea como porcentaje o proporción (“densidad”).

Histograma para el precio de las casas



Parámetros de Tendencia Central: Media

- **Parámetro**: característica descriptiva (o *resumen*) de una población, que a veces es desconocida y que queremos conocer.
- **Media de la población**: es la suma de todos los valores en la población, dividida entre el total de valores en dicha población:

$$\mu = \sum_{i=1}^{N} X_i \Big/ N$$

- N es el número total de elementos en la población.
- X_i representa el i-ésimo valor en la población.
- Σ indica la operación de sumar.

Media o promedio muestral

- **Estadístico**: característica descriptiva (o resumen) de una muestra.

- La media de una muestra es la suma de todos los valores muestrales divididos entre el número total de observaciones muestrales:

$$\overline{X} = \sum_{i=1}^{n} X_i \Big/ n$$

- Donde *n* es el número total de valores (observaciones) en la muestra.

Características de la media aritmética

- Todo conjunto de datos expresados numéricamente (mediante números) tiene un valor medio.
- Al evaluar la media se incluyen **todos** los valores del conjunto (sean poblacionales o muestrales).
- Un conjunto de valores sólo tiene una **única** media.
- La media es la única medida de ubicación (o tendencia central) donde la suma de las desviaciones de cada valor con respecto a la media, siempre es cero.

Medidas de Dispersión: Varianza de la población

- La varianza de la población es la media aritmética de las desviaciones cuadráticas respecto a la media de la población:

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

- Es una medida que resume la variabilidad total de un atributo alrededor de su media.
- Se toma potencia 2 para que los desvíos respecto a la media sean todos positivos y no se cancelen.

Varianza muestral

- La **varianza muestral** estima la varianza de la población mediante las realizaciones muestrales X_i de la variable X y la media muestral

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

- Notar que la varianza no se representa (o se mide) en las mismas unidades de medida que X, ya que se toma la potencia 2.

Desviación estándar poblacional

- La **desviación estándar poblacional** es la raíz cuadrada de la variancia de la población.

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

- Es una transformación de la varianza que permite medir/resumir la variabilidad en las mismas unidades de medida que posee *X*.

Desviación estándar muestral

- La desviación estándar muestral es la raíz cuadrada de la varianza muestral.

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$$

Dispersión relativa

- El **coeficiente de variación** es la razón de la desviación estándar a la media aritmética, expresada como porcentaje:

$$CV = \frac{S}{\overline{X}}.100$$

- La idea es medir “cuántas medias” equivalen a un desvío estándar o “*cuantas medias se desvía la variable en promedio*”

Probabilidad Frecuentista: v.a. discreta

- Desde une interpretación **Frecuentista**, podemos definir la probabilidad de que una casa de la muestra (tomada al azar) tenga 2 habitaciones: P(*Habitaciones*=2) = 0.2286, así P(*Habitaciones* =6)=0.1333

Nº de Habitaciones	Frecuencia	Porcentaje	Frec. Acum.	Frec. Relativa
2	24	22.86	22.8	0.2286
3	26	24.76	47.62	0.2476
4	26	24.76	72.38	0.2476
5	11	10.48	82.86	0.1048
6	14	13.33	96.19	0.1333
7	2	1.90	98.10	0.019
8	2	1.90	100	0.019
Total	105	100		1

Probabilidad Frecuentista: v.a. discreta

- Definiendo al **evento** *A*=“Casa con 2 Habitaciones”, podemos escribir

p(*A*) como la Probabilidad de ocurrencia del evento “*A*”.

- La interpretación frecuentista se basa en computar la probabilidad en base a la **frecuencia relativa** de un evento en un gran número de repeticiones (infinito o igual al tamaño de toda la población “*N*”:

$$p(A) = \lim_{N \to \infty} \frac{N_A}{N} = \frac{n_A}{n}$$

Función de Probabilidad: v.a. discreta

- Con los eventos A="Casa con 2 Habitac.", B="Casa con 3 Habitac.", C="Casa con 4 Habitac." y D="Casa con más de 4 Habit.", se tendrá:

$$p(A)+p(B)+p(C)+p(D)+p(E) = 1$$

- Definiendo la v.a. *X* = "N° de Habitaciones", su ***función de probabilidad*** es $p(X=x) = f(x)$ siendo "x" un número de habitaciones en concreto.
- Y la ***Probabilidad Acumulada*** es: $p(X<=x)$.

Función de Probabilidad: v.a. continua

- La probabilidad de que una casa de la muestra (tomada al azar) tenga un precio entre 225 y 274 es igual a $p(225=<Precio>=274)=0.2762$.
- Así, $p(Precio<=274)=0.8667$ es la probabilidad acumulada.

Precio	Frecuencia	Porcentaje	Frec. Acum.	Frec. Relativa
125-174	16	15.24	15.24	0.1524
175-224	46	43.81	59.05	0.4381
225-274	29	27.62	86.67	0.2762
275-324	11	10.48	97.14	0.1048
325-374	3	2.86	100	0.0286
Total	105	100		1

Probabilidad Frecuentista: variable continua

- Siendo *X* = “Precio de la casa”, $p(X<=x) = f(x)$ es la ***Función de Densidad de Probabilidad*** y $p(X<=x)$ es la **Probabilidad Acumulada**.
- Las ***Funciones de Distribución Acumuladas (FDA)*** para v.a. Discretas y Contínuas son:

$$F_{\tilde{x}}(x) = p(\tilde{x} \leq x) = \sum_{k=-\infty}^{x} f(k)$$

$$F_{\tilde{x}}(x) = p(\tilde{x} \leq x) = \int_{-\infty}^{x} f(t)dt$$

- La FDA caracteriza o especifica completamente a la Distribución de Probabilidad de una variable aleatoria.