UNIDAD N° 4

: INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

Análisis de relaciones entre variables estadísticas

Análisis de relaciones entre variables estadísticas

1. Introducción a distribuciones bidimensionales

En Estadística, muchas veces nos interesa estudiar **dos variables al mismo tiempo**, para entender si existe alguna relación entre ellas. A esto se lo conoce como una **distribución bidimensional**, y constituye la base del análisis bivariado.

Sea una población o muestra donde a cada individuo se le han medido **dos variables** X e Y, los datos pueden representarse como pares ordenados:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Estos datos pueden organizarse en una **tabla de doble entrada** si se trata de variables cualitativas o discretas, o graficarse en un **diagrama de dispersión** si son cuantitativas.

---

2. Tablas de doble entrada

Una **tabla de doble entrada** o **tabla de contingencia** muestra cómo se distribuyen conjuntamente dos variables. Por ejemplo, si se clasifican estudiantes según el turno (mañana/tarde) y si aprueban o no una materia, se construye una tabla como esta:

	Aprobó	**No aprobó**	**Total**
Mañana	20	10	30
Tarde	25	15	40
Total	45	25	70

---

3. Distribuciones marginales y condicionadas

- **Distribuciones marginales**: se obtienen sumando los valores por fila o por columna, sin considerar el cruce con la otra variable. En el ejemplo anterior, la distribución marginal por turno es:

$$P(\text{Turno Mañana}) = \frac{30}{70}, \quad P(\text{Turno Tarde}) = \frac{40}{70}$$

- **Distribuciones condicionadas**: muestran cómo se distribuye una variable **dado un valor de la otra**. Por ejemplo, la distribución condicional de aprobación dado que el estudiante asiste a la tarde es:

$$P(\text{Aprobó} \mid \text{Tarde}) = \frac{25}{40}, \quad P(\text{No aprobó} \mid \text{Tarde}) = \frac{15}{40}$$

---

4. Covarianza

La **covarianza** mide la variación conjunta de dos variables cuantitativas. Se define como:

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Si la covarianza es **positiva**, indica que cuando una variable crece, la otra también tiende a crecer.
- Si es **negativa**, al aumentar una, la otra tiende a disminuir.
- Una covarianza **cercana a 0** indica poca o nula relación lineal.

Observación: La covarianza tiene la unidad del producto de las unidades de X e Y, por lo que **no es adimensional** ni comparable entre distintos contextos.

---

5. Correlación lineal de Pearson

El **coeficiente de correlación de Pearson** es una versión normalizada de la covarianza, que **mide la intensidad y dirección de una relación lineal** entre dos variables cuantitativas. Se calcula como:

$$r_{xy} = \frac{\text{Cov}(X, Y)}{s_X s_Y}$$

donde:

- s_X y s_Y son los desvíos estándar muestrales de X e Y

El valor de r_{xy} siempre está entre -1 y 1:

- $r_{xy} = 1$: correlación lineal positiva perfecta
- $r_{xy} = -1$: correlación lineal negativa perfecta
- $r_{xy} = 0$: no hay correlación lineal

Cuidado: correlación **no implica causalidad**.

Valor de r	Interpretación
0.9 – 1.0	Muy fuerte positiva
0.7 – 0.89	Fuerte positiva
0.4 – 0.69	Moderada positiva
0.1 – 0.39	Débil positiva
0.0	Nula
–0.1 a –0.39	Débil negativa
–0.4 a –0.69	Moderada negativa
–0.7 a –0.89	Fuerte negativa
–0.9 a –1.0	Muy fuerte negativa

---

6. Correlación de Spearman

Cuando los datos no cumplen con supuestos de linealidad, o se trata de variables **ordinales**, se utiliza la **correlación de Spearman**, basada en **rangos**:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

donde:

- d_i es la diferencia entre los rangos de cada observación
- n es el número de pares de datos

Este coeficiente también oscila entre -1 y 1, con la misma interpretación que Pearson, pero es **más robusto ante valores atípicos y relaciones no lineales**.

---

7. Resumen comparativo

Característica	Pearson r	Spearman r_s
Tipo de variable	Cuantitativa continua	Cuantitativa u ordinal
Supone relación lineal	Sí	No necesariamente
Sensible a outliers	Sí	Menos sensible
Basado en	Valores originales	Rangos

---

8. Independencia, dependencia funcional y estadística

- **Independencia estadística**: conocer una variable no proporciona información sobre la otra. En este caso, la covarianza y la correlación son cercanas a cero.
- **Dependencia funcional**: existe una relación exacta, por ejemplo: y=2x+3
- **Dependencia estadística**: hay relación, pero con variabilidad, como sucede en la mayoría de los casos reales.

---

9. Cierre: Aplicaciones en programación

Comprender la relación entre variables permite, por ejemplo:

- Optimizar procesos basados en múltiples factores (como tiempo de ejecución y consumo de memoria).
- Analizar resultados de pruebas A/B.
- Estimar el comportamiento de usuarios en base a variables de entrada.