Especialización en Ciencia de Datos

Actividades

Estudiante: Yoseph Barrera
Profesor(a): Silvia Salinas Ayaviri
Fecha: 20 de febrero de 2026

Proyecto Módulo 4: Inferencia Estadística (ENCAVI 2023–2024)

Paso 0: Obtención y preparación de datos

Se utiliza la Encuesta Nacional de Calidad de Vida y Salud (ENCAVI) 2023–2024 (Chile).[1] La base original se descargó desde el portal de datos abiertos y se trabajó con el archivo en formato `.dta`. Posteriormente, se filtró la muestra a jóvenes entre 18 y 29 años, obteniendo $n = 2054$ observaciones. Para facilitar el análisis se construyó una base reducida con variables esenciales vinculadas a actividad física y hábitos alimentarios.

Las variables seleccionadas en la base final son:

- `folio_encuesta`: identificador de encuesta.
- `nom_region`: región.
- `area`: zona (urbano/rural).
- `sexo`: sexo.
- `edad`: edad.
- `sum_comidas`: número de comidas diarias.
- `sum_comidas_cat`: categoría de comidas diarias (por ejemplo: "tres comidas", "cuatro comidas").
- `valor_ipaq`: nivel de actividad física (por ejemplo: "Inactivo", "Minimamente activo").
- `audit_cat`: clasificación asociada a consumo de alcohol (puede contener valores faltantes).

Paso 1: Planteamiento del problema e hipótesis

Pregunta de investigación. En jóvenes de 18 a 29 años, **¿existe asociación entre el nivel de actividad física** medido por `valor_ipaq` y **la cantidad de comidas diarias** medida por `sum_comidas_cat`?

Hipótesis. Sea A la variable categórica `valor_ipaq` (nivel de actividad física) y sea C la variable categórica `sum_comidas_cat` (categoría de número de comidas al día).

- H_0: A y C son independientes (no existe asociación entre actividad física y número de comidas).
- H_1: A y C no son independientes (existe asociación entre actividad física y número de comidas).

Estrategia de análisis. Se utilizarán tablas de contingencia y proporciones para describir la relación entre categorías. Posteriormente, la hipótesis se evaluará con una prueba de independencia (χ^2) con un nivel de significancia $\alpha = 0{,}05$. Como apoyo descriptivo, se considerará `sum_comidas` como medida cuantitativa complementaria.

[1] Fuente: ENCAVI 2023–2024, datos.gob.cl.

Paso 2: Probabilidad (eventos y cálculos con ENCAVI)

Para este paso se consideran las observaciones con información no faltante en `valor_ipaq` y `sum_comidas_cat`, manteniendo $n = 2054$.

Definición de eventos.

- Evento A: la persona es *inactiva* según `valor_ipaq` (categoría "Inactivo").
- Evento B: la persona reporta *cuatro comidas* al día según `sum_comidas_cat` (categoría "cuatro comidas").

Estructura (reglas básicas).

$$P(A \cap B) = P(B)\,P(A \mid B), \qquad P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Resultados empíricos. Se obtienen los conteos $n_A = 974$, $n_B = 878$ y $n_{A\cap B} = 415$. Por lo tanto:

$$\hat{P}(A) = \frac{974}{2054} = 0{,}4742, \qquad \hat{P}(B) = \frac{878}{2054} = 0{,}4275, \qquad \hat{P}(A \cap B) = \frac{415}{2054} = 0{,}2020.$$

La probabilidad de la unión es:

$$\hat{P}(A \cup B) = 0{,}4742 + 0{,}4275 - 0{,}2020 = 0{,}6996.$$

Las probabilidades condicionales:

$$\hat{P}(A \mid B) = \frac{415}{878} = 0{,}4727, \qquad \hat{P}(B \mid A) = \frac{415}{974} = 0{,}4261.$$

Y los complementos:

$$\hat{P}(A^c) = 1 - \hat{P}(A) = 0{,}5258, \qquad \hat{P}(B^c) = 1 - \hat{P}(B) = 0{,}5725.$$

Interpretación breve. Aproximadamente el 47.4 % de los jóvenes es inactivo y el 42.8 % reporta cuatro comidas al día. Además, $\hat{P}(A \mid B) = 0{,}4727$ es muy similar a $\hat{P}(A) = 0{,}4742$, lo que sugiere preliminarmente una asociación débil entre ambas variables; esto se evaluará formalmente en los pasos de inferencia.

Paso 3: Distribuciones de probabilidad

En esta etapa se identifican variables aleatorias discretas construidas a partir de las variables observadas, con el fin de modelar probabilidades y conectar con la inferencia posterior.

Variable Bernoulli para inactividad. Definimos la variable aleatoria

$$X = \mathbf{1}\{\texttt{valor\_ipaq} = \text{"Inactivo"}\},$$

donde $X = 1$ si la persona es inactiva y $X = 0$ en caso contrario. Por construcción, X sigue una distribución Bernoulli con parámetro $p = P(X = 1)$. A partir del Paso 2, una estimación natural es $\hat{p} = 0{,}4742$.

Distribución Binomial para conteos en una muestra. Si se seleccionan n personas de manera aleatoria e independiente (conceptualmente) desde la población objetivo, y se define $S = \sum_{i=1}^{n} X_i$ como el número de personas inactivas en esa muestra, entonces:

$$S \sim \text{Binomial}(n, p), \qquad P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Esta formulación permite calcular probabilidades sobre el número de inactivos en muestras de tamaño n, y será útil para construir intervalos de confianza y pruebas sobre proporciones en etapas posteriores.

Variable Bernoulli para cuatro comidas. De forma análoga, definimos

$$Y = \mathbf{1}\{\texttt{sum\_comidas\_cat} = \text{“cuatro comidas”}\},$$

donde $Y = 1$ si la persona reporta cuatro comidas diarias y $Y = 0$ en caso contrario. Entonces Y es Bernoulli con parámetro $q = P(Y = 1)$, estimado empíricamente como $\hat{q} = 0{,}4275$.

Comentario sobre el enfoque. En lugar de imponer una distribución paramétrica para las variables categóricas originales (`valor_ipaq` y `sum_comidas_cat`), se trabaja con indicadores Bernoulli que permiten una modelación clara y directa de eventos de interés.

Paso 4: Distribución muestral y Teorema del Límite Central (TLC)

Para evaluar el Teorema del Límite Central se utiliza la variable cuantitativa `sum_comidas` (número de comidas diarias). Se construyó empíricamente la distribución muestral de la media $\bar{X}_n$ mediante remuestreo con reemplazo (bootstrap), repitiendo $B = 4000$ veces para tamaños muestrales $n \in \{30, 50, 100\}$. De acuerdo con el TLC, al aumentar n la distribución de $\bar{X}_n$ tiende a aproximarse a una Normal y su dispersión disminuye, consistente con una desviación estándar proporcional a $1/\sqrt{n}$.

En los resultados se observa que la media de `sum_comidas` en la muestra es $\bar{x} = 3{,}3301$ con desviación estándar $s = 0{,}6585$. Además, la desviación estándar de las medias muestrales disminuye al aumentar n: para $n = 30$ es 0,1207, para $n = 50$ es 0,0927 y para $n = 100$ es 0,0654, lo que coincide con la predicción del TLC.

Paso 5: Intervalos de confianza (IC) para la media

Se estima la media poblacional de `sum_comidas` usando la media muestral $\bar{x}$ y su desviación estándar muestral s, con n observaciones válidas. Un intervalo de confianza del $100(1-\alpha)\,\%$ para μ se construye como:

$$IC_{1-\alpha}(\mu) = \bar{x} \pm t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}}.$$

Con los datos se obtuvo $n = 2054$, $\bar{x} = 3{,}3301$ y $s = 0{,}6585$. Los intervalos calculados fueron:

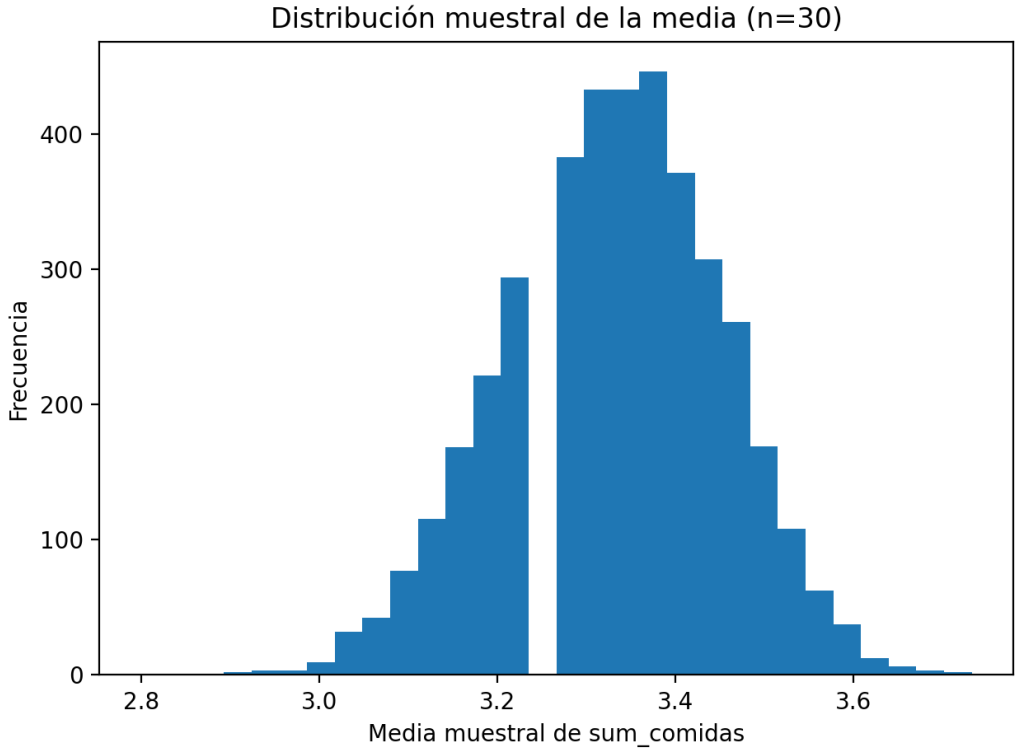

Figura 1: Distribución muestral de la media de `sum_comidas` (bootstrap, $n = 30$).

- IC 90 %: [3,3062, 3,3540]
- IC 95 %: [3,3016, 3,3586]
- IC 99 %: [3,2926, 3,3675]

Como es esperable, a mayor nivel de confianza el intervalo es más amplio (por ejemplo, el IC 99 % es más ancho que el IC 95 %).

Paso 6: Prueba de significancia (independencia χ^2)

Para evaluar la pregunta del Paso 1 se contrasta si `valor_ipaq` (actividad física) y `sum_comidas_cat` (categoría de comidas diarias) son independientes. Se utiliza una prueba χ^2 de independencia basada en una tabla de contingencia entre ambas variables.

- H_0: `valor_ipaq` y `sum_comidas_cat` son independientes.
- H_1: no son independientes.

El estadístico de prueba es:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad E_{ij} = \frac{(\text{total fila } i)(\text{total columna } j)}{n}.$$

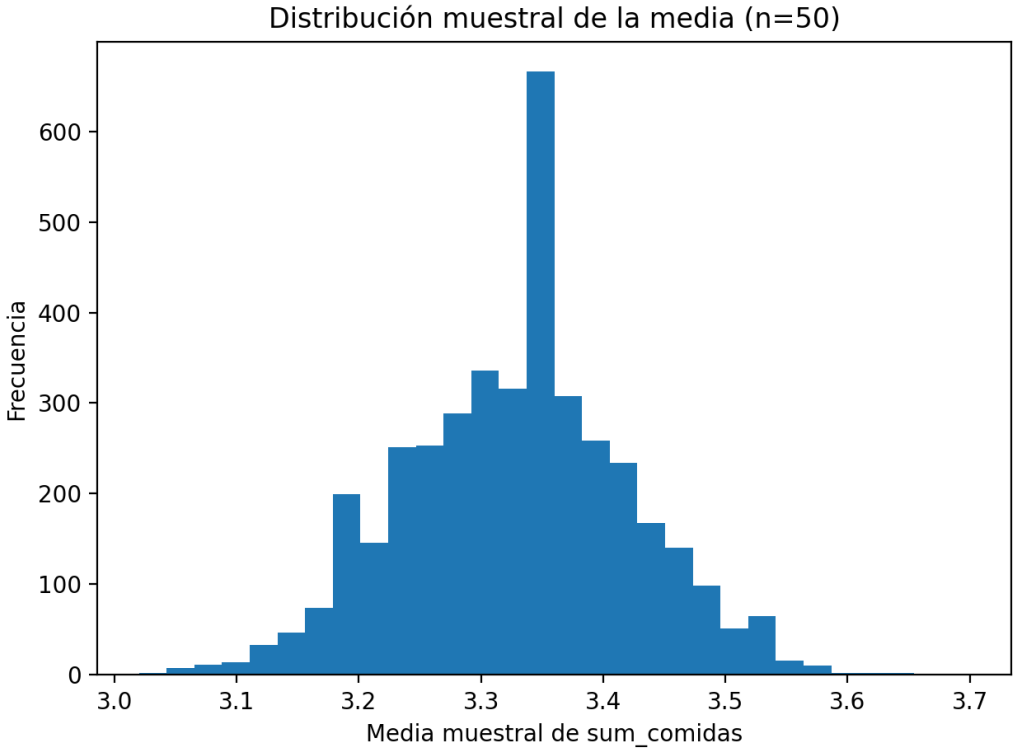

Figura 2: Distribución muestral de la media de `sum_comidas` (bootstrap, $n = 50$).

Con $n = 2054$ observaciones (sin faltantes en ambas variables), se obtuvo:

$$\chi^2 = 6{,}0870, \quad gl = 4, \quad p\text{-valor} = 0{,}192744.$$

Usando $\alpha = 0{,}05$, no se rechaza H_0 ya que el p-valor es mayor que α. En consecuencia, no se encuentra evidencia estadísticamente significativa de asociación entre el nivel de actividad física (`valor_ipaq`) y la categoría de comidas diarias (`sum_comidas_cat`) en esta muestra de jóvenes.

Como referencia descriptiva, la tabla de contingencia utilizada (resumen) fue:

`valor_ipaq`	a lo más dos comidas	cuatro comidas	tres comidas
Activo HEPA	30	202	244
Inactivo	96	415	463
Minimamente activo	58	261	285

Conclusión general

Con datos de ENCAVI 2023–2024 para jóvenes de 18 a 29 años ($n = 2054$), se estimó una media de 3,33 comidas diarias y se construyeron intervalos de confianza consistentes para este promedio. El análisis del TLC mostró que la distribución muestral de la media de `sum_comidas` se aproxima a una forma aproximadamente normal y su variabilidad disminuye al aumentar el tamaño muestral. Finalmente, la prueba χ^2 de independencia no entregó evidencia estadísticamente significativa de asociación entre el

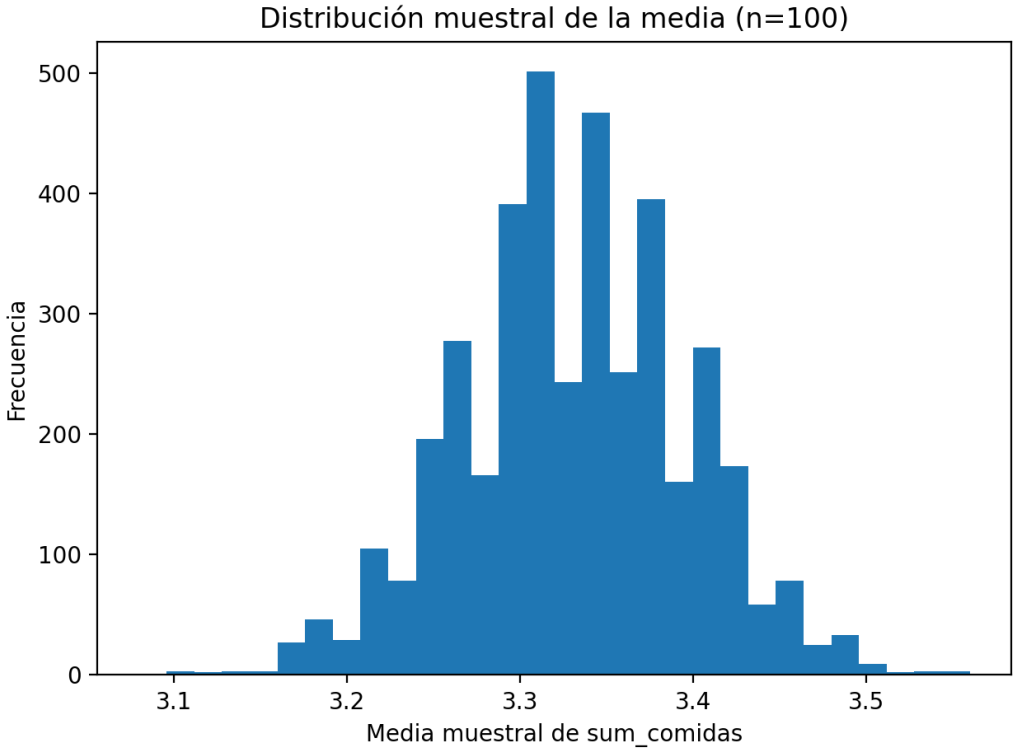

Figura 3: Distribución muestral de la media de `sum_comidas` (bootstrap, $n = 100$).

nivel de actividad física (`valor_ipaq`) y la categoría de comidas diarias (`sum_comidas_cat`) en esta muestra (p-valor = 0.1927).

Dado que el estudio es observacional, estos resultados describen asociaciones (o su ausencia) y no permiten establecer relaciones causales.

Diccionario de variables (base reducida)

Variable	Tipo	Valores/Categorías	Descripción
`folio_encuesta`	ID	numérico	Identificador de encuesta
`nom_region`	Categórica	texto	Región
`area`	Categórica	Urbano / Rural	Zona de residencia
`sexo`	Categórica	Mujer / Hombre	Sexo
`edad`	Numérica	enteros (años)	Edad
`sum_comidas`	Numérica	conteo	Número de comidas diarias
`sum_comidas_cat`	Categórica	a lo más dos / tres / cuatro	Categoría de comidas diarias
`valor_ipaq`	Categórica	Inactivo / Minimamente activo / Activo HEPA	Nivel de actividad física
`audit_cat`	Categórica	No / consumo riesgoso / NA	Clasificación asociada a alcohol