Proyecto Módulo 5: Análisis estadístico sobre hábitos saludables en jóvenes universitarios.

Nombre: Luis León Barrera

Curso: Fundamentos de Ciencia de Datos

Fecha: 27/02/2026

Introducción

Dada la necesidad que tiene la universidad para identificar factores que influyen en los hábitos de sueño, alimentación y actividad física en sus estudiantes es que se realiza este análisis estadístico que busca interpretar ciertas variables como los minutos de actividad física, niveles de estrés, horas de sueño entre otros factores que pudieran impactar en la calidad de vida de los estudiantes.

1. Lección 1: Método científico y estadística

Esta sección se centra principalmente en la definición del problema a investigar con la formulación de hipótesis y variables.

1.1. Definición del problema a investigar

Se busca analizar factores que influyen en los hábitos saludables en jóvenes universitarios, con foco en la actividad física. Se quiere ***evaluar si el promedio de minutos de actividad física semanal es menor que la recomendación de 150 minutos establecida por guías de salud.***

1.2. Formulación de Hipótesis

Hipótesis Nula (H_0): Los jóvenes universitarios en hacen al menos 150 minutos de actividad física a la semana.

Hipótesis Alternativa (H_1): Los jóvenes hacen menos de 150 de actividad física a la semana.

En notación estadística:

- H_0: $\mu \geq 150$
- H_1: $\mu < 150$

1.3. Identificación de variables

Las variables en este estudio se dividen entre cuantitativas y cualitativas.

Variables Cuantitativas:

- **Actividad física semanal**: tiempo total en minutos dedicado al ejercicio por semana. Variable *continua*.
- **Horas de sueño**: promedio de horas dormidas durante la semana. Variable *continua*.

- **Edad del estudiante**: edad del estudiante al momento del estudio (en años). Variable *discreta*.
- **Consumo de agua**: litros de agua ingerida diariamente por el estudiante. Variable *continua*.

Variables Cualitativas:

- **Género**: Masculino, Femenino, Otro. *Nominal*
- **Carrera**: Área de estudio (ingenieria, salud, artes, etc.). *Nominal*
- **Nivel de estrés percibido**: Bajo, medio, alto. *Ordinal*
- **Frecuencia de consumo comida rápida**: Nunca, 1-2 veces por semana, +3 veces por semana. *Ordinal*
- **Calidad del sueño**: Buena, regular, mala. *Ordinal*

1.4. Enfoque metodológico

Para el desarrollo de este proyecto se trabajará con un enfoque cuantitativo con apoyo de variables cualitativas. Este enfoque permite medir numéricamente los hábitos saludables (por ejemplo, minutos de actividad física, horas de sueño) y, al mismo tiempo, relacionarlos con variables categóricas como género, carrera o nivel de estrés, facilitando el análisis estadístico posterior.

Se aplicará el método científico deductivo con un enfoque cuantitativo. Este proceso iniciará con la observación de la problemática de salud estudiantil, seguida por la formulación de hipótesis contrastables (H_0 y H_1). La fase de experimentación se sustituirá por una simulación de datos rigurosa y un análisis estadístico inferencial para validar o rechazar las premisas iniciales, asegurando la objetividad en las conclusiones.

1.5. Diseño preliminar del estudio

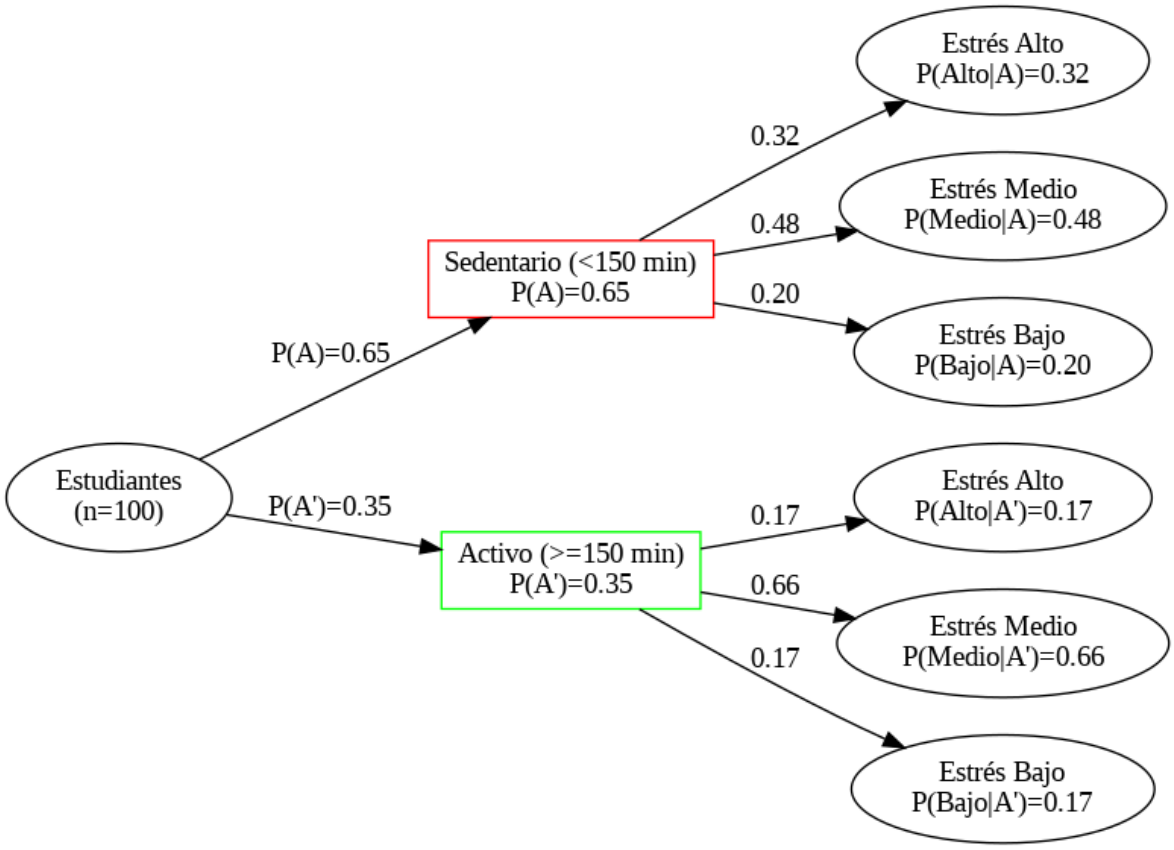
- **Población objetivo**: jóvenes universitarios de distintas carreras de una universidad.
- **Tipo de estudio**: observacional, transversal.
- **Técnica de muestreo**: se propone un muestreo aleatorio simple, con una muestra de al menos 100 estudiantes.
    - Para efectos de este proyecto se realizará una simulación de datos utilizando la librería Pandas en Python, respetando las características definidas para cada variable.

2. Lección 2: Probabilidad y estadística

En esta lección se definen eventos aleatorios y se simula una muestra de tamaño n = 100

2.1. Eventos aleatorios

- Evento A: Estudiante sedentario -> minutos_ejercicio < 150
- Evento B: Estrés alto -> nivel_estres = alto



Según lo que podemos ver en el árbol los estudiantes sedentarios (<150 minutos de ejercicio a la semana) tienen una mayor proporción de estrés alto (32%), comparado con los estudiantes activos que tienen una proporción de estrés alto de (17%).

Se buscará calcular la P(A), P(B), P(A ∩ B), P(A ∪ B)

P(A)= 0.65

P(B)= 0.27

P(A ∩ B)= 0.21

P(A ∪ B)= 0.71

2.2. Justificación del Muestreo

El método utilizado para este muestro es Muestreo Aleatorio Simple (MAS).

Se seleccionó este método para asegurar que cada individuo de la población universitaria tenga la misma probabilidad de ser incluido en la muestra. Dado que la simulación utiliza np.random, garantizamos la independencia de las observaciones, requisito fundamental para aplicar el Teorema del Límite Central en etapas posteriores.

2.3. Muestreo simulado:

Con el fin de generar los datos de estudio se utilizaron las librerías Numpy y Pandas para generar un dataset con tamaño n=100. A continuación, se puede ver el extracto del código utilizado.

```
np.random.seed(42)
n_estudiantes = 100

#Creación de dataset simulado
data = {
    'minutos_ejercicio': np.random.normal(142, 30, n_estudiantes).clip(0), #uso de clip para que el mínimo sea 0
    'horas_sueño': np.random.normal(7, 2, n_estudiantes).clip(0,24),
    'edad': np.random.randint(18,25, n_estudiantes),
    'consumo_agua': np.random.normal(3, 1, n_estudiantes).clip(0.5),
    'nivel_estres': np.random.choice(['Bajo', 'Medio', 'Alto'], n_estudiantes, p=[0.2, 0.5, 0.3]),
    'carrera': np.random.choice(['Ingenieria', 'Salud', 'Artes', 'Leyes'], n_estudiantes, p=[0.3, 0.3, 0.1, 0.3]),
    'genero': np.random.choice(['Masculino', 'Femenino', 'Otro'], n_estudiantes, p=[0.45, 0.50, 0.05]),
    'frecuencia_comida_rapida': np.random.choice(['Nunca', '1-2 veces por semana', '+3 veces por semana'], n_estudiantes, p=[0.3, 0.4, 0.3]),
    'calidad_sueño': np.random.choice(['Buena', 'Regular', 'Mala'], n_estudiantes, p=[0.3, 0.4, 0.3])
}

df = pd.DataFrame(data)
df.describe()
```

Se crearon las variables mencionadas en el punto anterior y se le asignaron distintas probabilidades para tener una muestra aleatoria y que cubra los distintos casos.

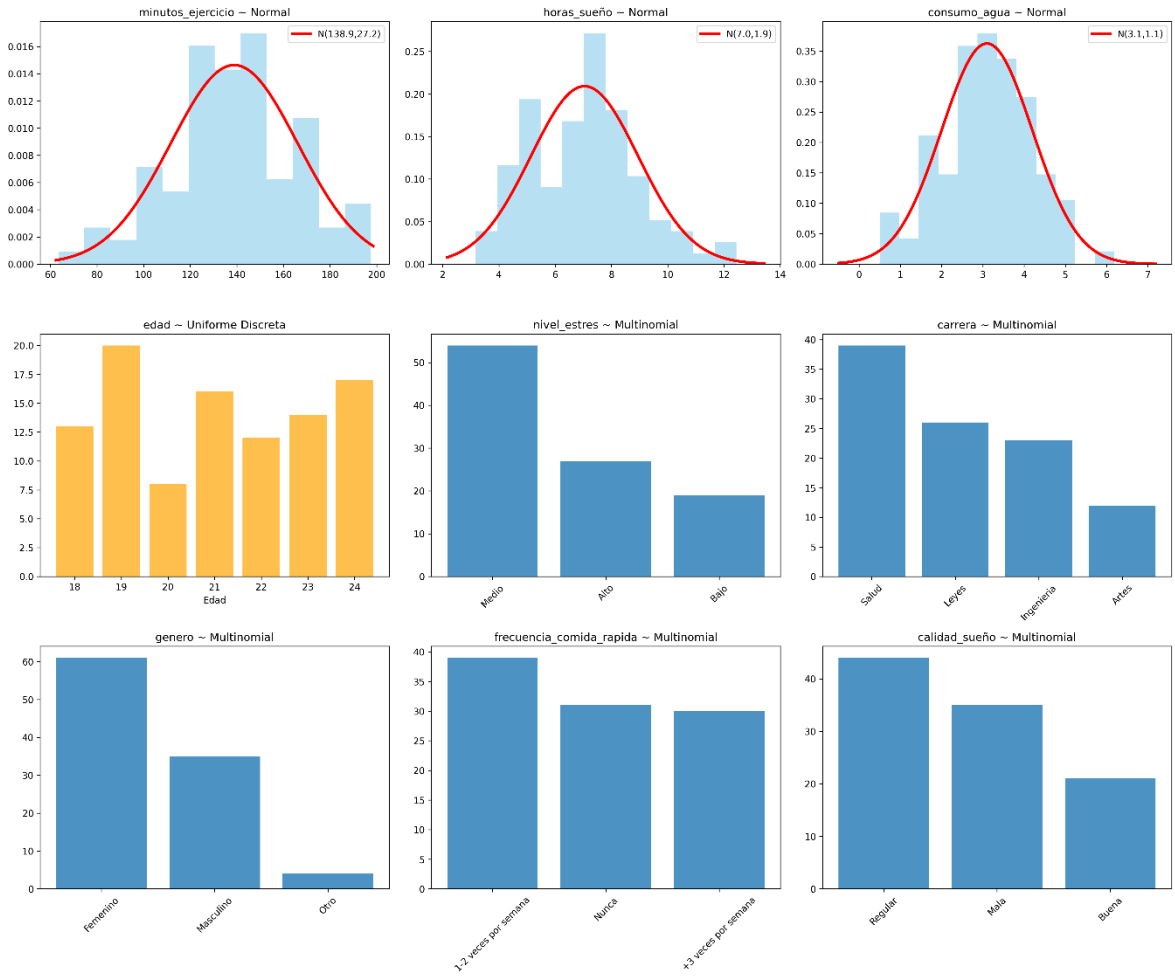
3. Lección 3: Distribución de probabilidad

Las distintas variables creadas se ajustan a distribuciones de tipo normal, uniforme y binomial según se lista a continuación:

- Distribución Normal: minutos_ejercicio, horas_sueño, consumo_agua. Se opta por este tipo de distribución debido a que para cada una de estas variables se pueden tomar distintos valores dentro de un rango, por ejemplo, una persona podría dormir desde 0 hasta 24 horas.
- Distribución Uniforme: edad. Esta distribución se justifica en que los estudiantes pueden tener solo números enteros entre los 18 y 25 años de edad.

- Distribución Binomial: nivel_estres, carrera, genero, frecuencia_comida_rapida, calidad_sueño. Para estas variables categóricas solo se puede elegir entre ciertas opciones y a cada una se le otorga un peso en probabilidad.

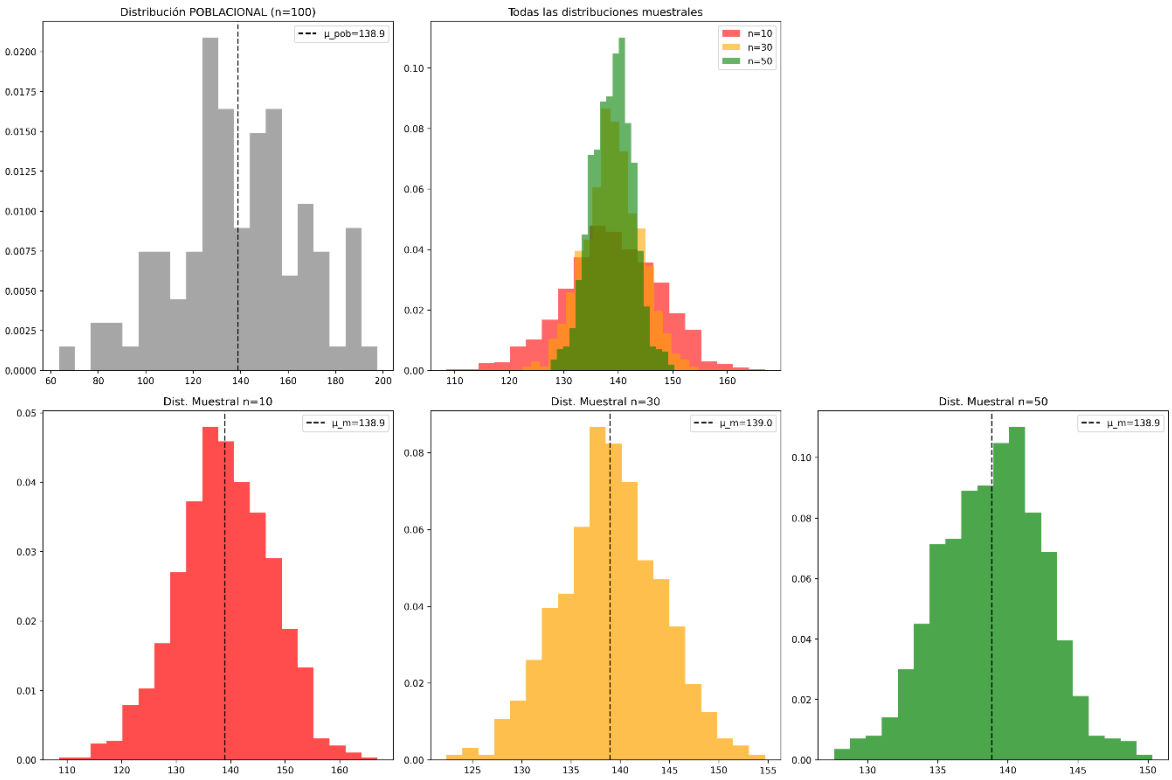
A continuación se puede observar las distribuciones de forma gráfica:



4. Lección 4: Distribución muestral y Teorema del Límite Central

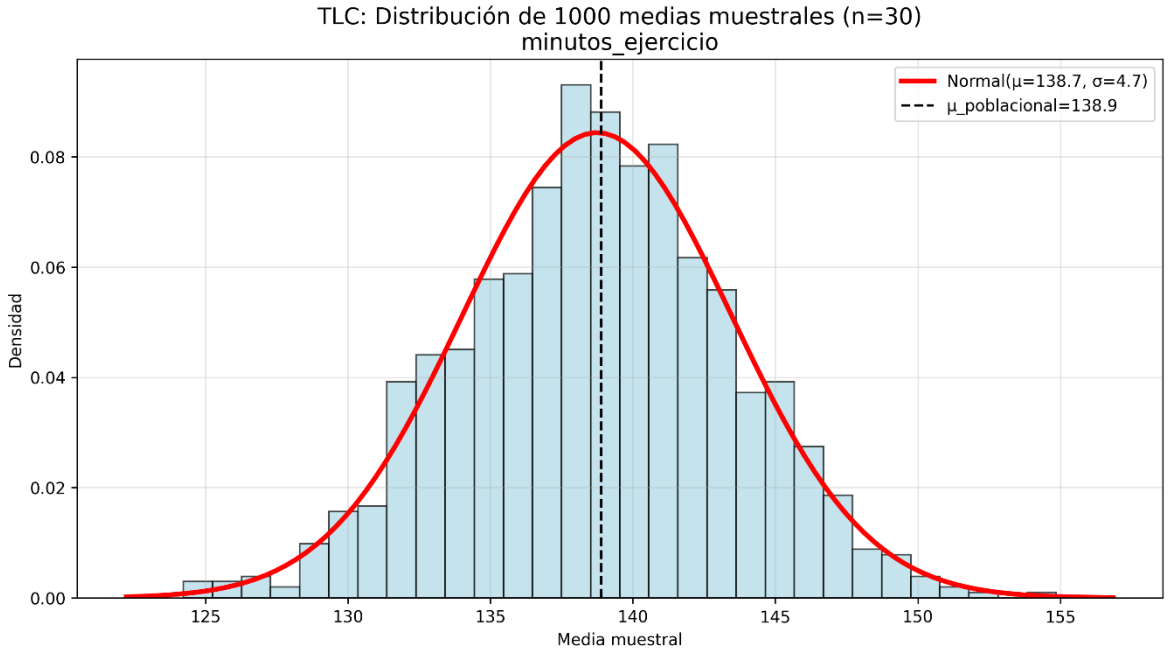
El TLC establece que, si se toman muestras lo suficientemente grandes de cualquier población (sin importar si su distribución es uniforme, sesgada o totalmente irregular), la distribución de las medias de esas muestras se aproximará a una distribución normal.

Para efectos de este proyecto se tomaron 1000 muestras de n = 10, 30 y 50, para poder visualizar el comportamiento de los datos.



Tal como se observa en la imagen las distintas medias se ajustan a una distribución normal.

También se probó solamente utilizando 1000 muestras de tamaño n=30

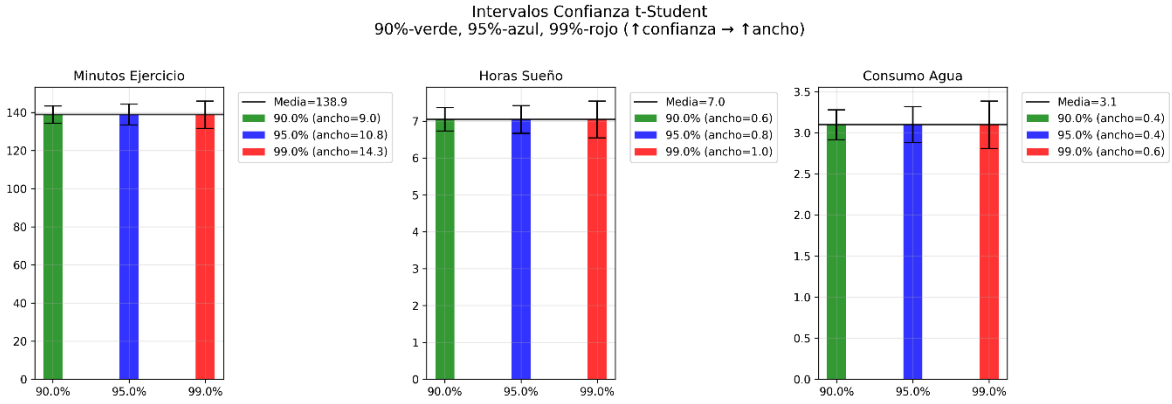


Se observa en ambos casos que la curva es más pareja, se hace más suave y se acerca a la media poblacional, junto con disminuir la desviación. La dispersión de las medias muestrales se hace más pequeña cuando n crece, tal como dice el Teorema del Límite Central.

5. Lección 5: Inferencia e intervalos de confianza para la media

En esta lección se calculan los intervalos de confianza para las variables principales de minutos_ejercicio, horas_sueño y consumo_agua.

Utilizando t-student se obtiene el IC de 90%, 95% y 99% con lo que se observa el aumento en el ancho con cada intervalo de confianza.



La real diferencia se puede ver al momento de cambiar el n, donde se observa la variación en el ancho del IC. Para un IC de 95%, a medida que varía el n y se hace más grande el IC tiene menor ancho, lo que lo hace más preciso

- n=30: [127.6, 148.4] (ancho=20.8)
- n=50: [128.1, 144.1] (ancho=16.0)
- n=100: [136.4, 146.0] (ancho=9.6)

Para un IC del 95% con un n=100 es un 37% más preciso de con n=30. Si se lleva esto a las variables en estudio se puede inferir que con un 95% de seguridad (n=100) los estudiantes harán entre 133.5 y 144.3 minutos de ejercicio a la semana, lo cual está por debajo de los 150 minutos recomendados por la OMS.

6. Lección 6: Test de significancia

En esta lección debemos recordar las hipótesis planteadas:

- Hipótesis Nula (H_0): Los jóvenes universitarios en hacen al menos 150 minutos de actividad física a la semana.
- Hipótesis Alternativa (H_1): Los jóvenes hacen menos de 150 de actividad física a la semana.

Con un valor de significancia α=0.05 se realizaron los análisis y se obtuvieron los siguientes resultados:

- Media observada: 138.88 min
- t_observado = -4.080
- t_crítico (α=0.05, 1 cola) = 1.660
- p-value = 0.00005

Con un p-value menor a 0.0001 tenemos evidencia altamente significativa para rechazar H_0, tenemos evidencia muy fuerte de que el promedio de ejercicio semanal es menor a 150 minutos.

En otras palabras, con 95% confianza, los universitarios hacen en promedio 138.9 < 150 min ejercicio/semana.

6.1. Errores Tipo I Y II

Los errores tipo I y II son riesgos inherentes a cualquier prueba estadística. Incluso con el mejor diseño del mundo, siempre existe una probabilidad (aunque sea mínima) de equivocarse. En el caso de este proyecto estaríamos ante este tipo de errores en estos casos:

- Error Tipo I (α = 0.05): Consistiría en concluir que los estudiantes no hacen suficiente ejercicio cuando en realidad sí cumplen el promedio de 150 min. El costo sería implementar programas de salud innecesarios.
- Error Tipo II (β): Consistiría en no detectar que los estudiantes están por debajo del nivel saludable. El costo sería la inacción ante un problema real de sedentarismo universitario.

7. Conclusiones y recomendaciones

Después del análisis realizado a este grupo de estudiantes universitarios se encontraron algunos hallazgos que permitirían a la universidad tomar acciones correctivas buscando una mejora en la salud de sus estudiantes.

Hallazgos estadísticos

Según el estudio la estadística indica que los estudiantes son generalmente sedentarios, con un 95% de seguridad los estudiantes realizan menos de 150 minutos de ejercicio a la semana con una media observada de 138.9 minutos.

Dado que el análisis probabilístico (Lección 2) sugiere una intersección importante entre el estrés alto y el bajo nivel de ejercicio, se recomienda que los programas deportivos se promocionen también como herramientas de salud mental.

Recomendaciones al Área Salud Universitaria

Para mejorar la salud de los estudiantes de la universidad se recomienda tomar acción inmediata en fomentar el tiempo que los estudiantes realizan ejercicio semanalmente para alcanzar como mínimo lo recomendado por la OMS (150 minutos). Para esto se pueden realizar las siguientes acciones:

- Ofrecer espacios adecuados para que los estudiantes se ejerciten al interior de la universidad.
- Establecer programas de ejercicio obligatorio, en los que deban cumplir con sesiones y tiempo semanal.
- Incluir dentro de los programas de estudio de las distintas carreras “cursos” o bloques obligatorios de ejercicio.

Este tipo de acciones apuntan a mejorar la salud y que también permitan a los estudiantes hacer actividades diferentes a las que realizan normalmente o a hacer algo distinto a solo estudiar. También ayudaría a reducir los niveles de estrés.

El proyecto se centro en los minutos de actividad física, pero sería interesante poder estudiar otras variables como el rendimiento académico, la deserción y como la actividad física influye en estas.