David Mestres

Actividad 1-

A-La figura muestra la prevalencia de distintas enfermedades en Estados Unidos antes y después de la introducción de las vacunas. Aunque no se hagan cálculos concretos, la tendencia es bastante clara: antes de la vacuna, los niveles de prevalencia eran altos y relativamente estables, mientras que después de su aplicación la prevalencia cae de manera muy marcada.

Esto me lleva a pensar que existe una relación directa entre la administración de la vacuna y la reducción de la enfermedad. Es decir, cuanto más extendida estuvo la vacunación, menor fue la proporción de población afectada. No parece una simple coincidencia, ya que la caída es fuerte y sostenida en el tiempo.

Desde el punto de vista estadístico, lo que observamos es una asociación muy clara entre las dos variables (prevalencia y vacunación). Sin embargo, también es importante ser prudentes: la gráfica por sí sola no demuestra causalidad de forma estricta. En principio, podrían existir otros factores (mejores condiciones sanitarias, cambios en los hábitos de vida, etc.). Aun así, con la evidencia epidemiológica que se conoce, la interpretación más razonable es que la vacunación fue el factor clave que explica la reducción tan drástica de los casos.

Conclusión: los resultados apoyan con fuerza la idea de que la introducción de las vacunas tuvo un impacto directo en la disminución de la prevalencia de las enfermedades en Estados Unidos.

B-El debate descrito refleja muy bien las diferencias entre asociación y causalidad. El estudio de Doll y Hill mostró que fumar estaba asociado con un mayor riesgo de cáncer de pulmón, es decir, que ambas variables aparecían relacionadas en los datos. Sin embargo, como señalaron críticos como Fisher, esto no significa automáticamente que el tabaco fuera la causa del cáncer.

Fisher plantea varias alternativas que ejemplifican posibles factores de confusión:

-Que el cáncer (o lesiones precancerosas) provocara irritación en los pulmones y eso llevara a fumar más.
-Que existiera una predisposición genética que explicara tanto la propensión a fumar como el riesgo de cáncer.
-Que ciertas características personales (resistencia psicológica, capacidad de autocontrol) influyeran tanto en evitar fumar como en resistir al cáncer.

Estos argumentos muestran cómo una asociación estadística puede tener distintas explicaciones y cómo los factores de confusión pueden distorsionar la interpretación de los resultados.

Con los datos de la época, lo más que se podía afirmar era que había una fuerte asociación entre fumar y cáncer de pulmón. Pero sin estudios longitudinales, experimentales u otras evidencias adicionales, no se podía demostrar de forma concluyente la causalidad.

Conclusión: el ejemplo hace referencia a la importancia de diferenciar asociación de causalidad y de tener en cuenta posibles factores de confusión antes de llegar a conclusiones definitivas.

Actividad 2

A-Spotify utiliza variables como tempo, valence, energy, acousticness, danceability, instrumentalness o speechiness, que son medidas numéricas extraídas directamente del archivo de audio. A partir de esas características, el objetivo es clasificar cada canción en un género musical (rock, pop, rap, etc.).

Este tipo de problema corresponde a una tarea de clasificación supervisada, ya que lo que se busca es asignar una categoría (género) a partir de un conjunto de variables explicativas. La técnica más probable que utilice Spotify es alguna forma de análisis discriminante o, más modernamente, modelos de clasificación supervisada como regresión logística multinomial, árboles de decisión o algoritmos de machine learning (random forest o redes neuronales por ejemplo).

En un marco puramente estadístico, la técnica clásica sería el análisis discriminante , porque justamente sirve para separar grupos (géneros) en función de variables cuantitativas. Sin embargo, dada la complejidad de los datos de audio, en la práctica Spotify combina modelos más avanzados de clasificación.

Conclusión: Spotify probablemente utiliza una técnica supervisada de clasificación, siendo el análisis discriminante la más cercana dentro de las técnicas estadísticas tradicionales.

B-En este caso, lo que Spotify quiere predecir es si una canción llegará o no a las 1000 reproducciones en un año. Eso es básicamente una variable binaria: sí o no. La técnica más adecuada para este tipo de problema es la regresión logística, porque justamente sirve para estimar probabilidades de que ocurra un evento a partir de varias características.

Si pensamos en los datos que Spotify podría tener de los últimos 3 años, seguramente incluirían:

-Las características musicales de cada canción (tempo, energy, acousticness, danceability, etc.).
-Datos de contexto como si la canción entró en playlists importantes o en qué año fue publicada.
-El número de reproducciones en su primer año, que es lo que se usaría para marcar si pasó o no el umbral de 1000.

Con este histórico, el modelo aprendería qué tipo de canciones tienden a superar el mínimo y cuáles no. Por ejemplo, podría detectar que las canciones con alta “danceability” y que aparecen en playlists grandes tienen muchas más probabilidades de generar ingresos.

Conclusión: Spotify seguramente utiliza una regresión logística (o algún modelo similar de clasificación supervisada) para predecir, basándose en esos datos de los últimos 3 años, si una canción llegará al mínimo de reproducciones.

Actividad 3

A-Por el tipo de conclusión, parece una investigación observacional.
Probablemente se siguió a un grupo de chicas desde que practicaban deporte en la etapa escolar hasta su vida adulta, y luego se comparó su situación laboral con la de chicas que no practicaban deporte. No es un experimento controlado (no se puede obligar a unas chicas a hacer deporte y a otras no)

B-Las conclusiones son interesantes, pero hay que tomarlas con precaución. El hecho de encontrar que las chicas que hacían deporte tienen un 50% más de probabilidades de conseguir puestos directivos muestra una asociación, pero no necesariamente una causalidad directa. Puede haber otros factores detrás, como el nivel socioeconómico, el apoyo familiar, la educación recibida o incluso la personalidad (motivación, liderazgo, perseverancia), que también influyen tanto en practicar deporte como en tener éxito profesional.

C-Para sacar una conclusión como esta seguramente no se limitaron a mirar una tabla con porcentajes, sino que habrán usado alguna técnica que permita analizar varias variables a la vez. Lo más probable es que aplicaran modelos de regresión logística, porque con ellos se puede estimar la probabilidad de acabar en un puesto directivo teniendo en cuenta la práctica deportiva y, al mismo tiempo, controlar otros factores como el nivel de estudios o la edad.

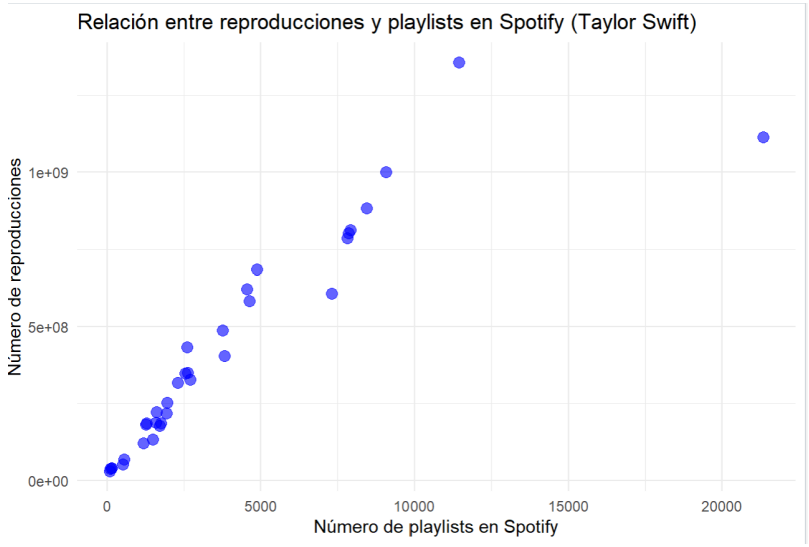
También podrían haber probado con otros modelos de regresión múltiple si trabajaban con variables más continuas (por ejemplo, ingresos o años de experiencia), pero la idea básica es la misma: ver qué peso tiene la práctica deportiva cuando se ponen todas las variables en conjunto.

Concluyendo, es muy probable que hayan usado algún tipo de análisis multivariante de regresión, porque es la manera más clara de separar lo que realmente aporta el deporte de lo que se debe a otras características de las chicas del estudio.

Actividad 4

A-
```
# Gráfico de dispersión
ggplot(taylor, aes(x = in_spotify_playlists, y = streams)) +
  geom_point(color = "blue", alpha = 0.6, size = 3) +
  labs(title = "Relación entre reproducciones y playlists en Spotify (Taylor Swift)",
       x = "Número de playlists en Spotify",
```

y = "Número de reproducciones") +
theme_minimal()



Se ve claro que cuanto más aparece una canción de Taylor Swift en playlists de Spotify, más reproducciones consigue. Hay una relación positiva muy marcada: las canciones con poca presencia en playlists apenas suman streams, mientras que las que están en miles de playlists superan con facilidad los cientos de millones. Eso muestra lo importante que es la visibilidad en playlists para el éxito de una canción

B-La comparación entre total y media permite ver que Taylor Swift tiene tanto años con canciones individuales muy potentes (2014) como años con grandes volúmenes de lanzamientos que acumulan muchísimos streams en conjunto (2022)

```
## # A tibble: 9 × 4
##   released_year total_streams mean_streams n_songs
##           <int>         <dbl>        <dbl>   <int>
## 1          2010     621660989   621660989        1
## 2          2012     882831184   882831184        1
## 3          2014    3255979784  1085326595.       3
## 4          2017     685032533   685032533        1
## 5          2019     986081433   493040716.       2
## 6          2020    1419143333   709571666.       2
## 7          2021     934068522   467034261        2
## 8          2022    5002045875   312627867.      16
## 9          2023     266814647    44469108.       6
```

C-Si la variable dependiente es continua, la técnica estadística más adecuada es la regresión lineal múltiple. Este modelo permite analizar cómo influyen tanto el número de playlists como el año de lanzamiento en la cantidad de streams, y además puede identificar si una variable explica mejor que la otra el éxito de las canciones.

D-Variable dependiente (lo que queremos explicar o predecir):*Número de reproducciones* (streams).

Variables independientes (los factores que usamos para explicar la dependiente):*Número de playlists de Spotify* (in_spotify_playlists) y*año de lanzamiento* (released_year).