# Modelo probabilístico de herencia de color naranja en gatos

Liz de María Salazar Amaya

Universidad de Costa Rica

CA0204 – Herramientas para Ciencia de Datos I

Profesor: Luis Juárez Potoy

2 de diciembre del 2025

# 1 Introducción

El presente proyecto desarrolla un modelo probabilístico en R para estudiar la herencia del color naranja en gatos, un rasgo determinado por el gen (O), ubicado en el cromosoma X. Este gen presenta dos alelos principales:

- (O): color naranja
- (o): no naranja

Debido al ligamiento al cromosoma X, la expresión de este rasgo presenta diferencias entre machos (XY) y hembras (XX). Los machos tienen un solo cromosoma X, por lo que solo pueden portar un alelo del gen (O), mientras que las hembras pueden ser homocigotas u heterocigotas, donde pueden presentar características de gatos calicó.

El objetivo del proyecto es comparar un modelo teórico que se base en el equilibrio de Hardy-Weirberg y un modelo empírico que se enfoque en los datos reales recolectados de la plataforma de adopción PetFinder. En ambos modelos se simulan cruzamientos entre gatos para evaluar la distribución esperada de genotipos y determinar en qué medida la población real se ajusta o se desvía de la teoría.

# 2 Datos utilizados

Se recopiló información de gatos mayores o iguales a un año, obtenida de refugios ubicados en Nueva York mediante la plataforma PetFinder. Las variables más relevantes utilizadas en el análisis fueron:

- **Sexo**: Macho / Hembra
- **Naranja**: indica si el gato presenta color naranja en el pelaje (Sí / No)
- **Color_Base**: descripción del patrón de color (por ejemplo, Black, Calico, Dilute Calico, Tabby, etc.)

En la limpieza de datos se identificaron problemas típicos de bases provenientes de refugios, como machos clasificados como Calico o Tortie, categorías que no son genéticamente posibles bajo un modelo XX/XY estándar, salvo casos excepcionales como XXY. Otro limitación que se presentó fueran descripciones de color basadas en la apariencia estética más que en la genética subyacente. Para el análisis del gen (O), estos casos se clasificaron como que no aportan al modelo y se registraron como valores faltantes en el genotipo correspondiente, evitando que distorsionaran las estimaciones de frecuencias.

# 3 Metodología

La metodología general se dividió en cuatro etapas principales: asignación de genotipos, estimación de frecuencias alélicas, construcción de modelos (teórico y empírico) y verificación del equilibrio de Hardy–Weinberg.

## 3.1 Asignación de genotipos

Se diseñó una función en R para inferir el genotipo del gen (O) a partir del sexo y del color base:

- **Machos (XY)**

  - (X^OY): machos naranjas
  - (X^oY): machos no naranjas

- **Hembras (XX)**
  - (X^O X^O): hembras de pelaje naranja sólido
  - (X^O X^o): hembras calicó/tortie (detectadas mediante valores como Calico, Dilute Calico, Tortie, Torbie en `Color_Base`)
  - (X^o X^o): hembras sin color naranja

Los registros donde un macho aparecía con descripciones de tipo Calico o Tortie se consideraron inconsistentes y se asignaron como NA en el genotipo (O). Estos casos se excluyeron del análisis de frecuencias alélicas.

### 3.2 Estimación de frecuencias alélicas

A partir de los genotipos depurados se contó el número de alelos (O) y (o) en la población. Sea:

- (n_O): número total de alelos (O)
- (n_X): número total de cromosomas X en la muestra

La frecuencia del alelo naranja se definió como:

$$p = \frac{n_O}{n_X},$$

mientras que la frecuencia del alelo no naranja se obtuvo como:

$$q = 1 - p.$$

### 3.3 Modelo teórico (Hardy–Weinberg)

Bajo el supuesto de equilibrio de Hardy–Weinberg, para las hembras se esperan las siguientes proporciones genotípicas:

- (p^2) para (X^O X^O)
- (2pq) para (X^O X^o)
- (q^2) para (X^o X^o)

Para los machos, al tener un solo cromosoma X, las proporciones esperadas son simplemente:

- (p) para (X^OY)
- (q) para (X^oY)

Con estas proporciones se construyó un modelo de simulación teórico en el que se generaron 100 machos y 100 hembras según las probabilidades definidas por (p) y (q). Cada pareja de gatos produjo 100 crías simuladas y se repitió el proceso para obtener proporciones estables de genotipos en la descendencia.

### 3.4 Modelo empírico

El modelo empírico utilizó directamente los genotipos observados en la base de datos. El procedimiento fue seleccionar al azar conjuntos de machos y hembras reales según su genotipo (O), simular camadas utilizando una función de cruce que para cada cría se elige al azar uno de los dos cromosomas X de la madre. Luego, se sortea el sexo de la cría con un 0.5 de probabilidad de cada género, y finalmente, se combinan los cromosomas para obtener el genotipo de la cría. Este modelo refleja cómo se comportaría la población si se reprodujera respetando las frecuencias genotípicas observadas, sin suponer equilibrio.

### 3.5 Verificación de equilibrio de Hardy–Weinberg

Para evaluar si la población se encontraba en equilibrio de Hardy–Weinberg se compararon, para cada categoría genotípica, los conteos:

- **Observados** (a partir de los datos depurados), y
- **Esperados** (calculados con (p) y (q)).

Se aplicó una prueba de chi-cuadrado:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

donde (O_i) y (E_i) representan los valores observados y esperados para el genotipo (i). Se calcularon p-valores por separado para machos y hembras, usando los grados de libertad correspondientes.

## 4 Resultados

### 4.1 Equilibrio de Hardy–Weinberg

Los gráficos de genotipos observados versus esperados (véase anexo A) muestran discrepancias visibles en varias categorías, tanto en machos como en hembras. La prueba de chi-cuadrado arrojó:

- ($p < 0.05$) para machos,
- ($p < 0.05$) para hembras.

Estos resultados permiten rechazar la hipótesis de equilibrio de Hardy–Weinberg en la población analizada. Este hallazgo es coherente con la naturaleza de los datos, ya que la muestra proviene de refugios y no de una población aleatoria. Además, existen errores en la clasificación de color (por ejemplo, machos marcados como Calico). Así mismo, es probable que no se cumplan supuestos como apareamiento aleatorio o ausencia de selección.

### 4.2 Resultados del modelo teórico

El modelo teórico basado en (p) y (q) (véase anexo B) produjo una distribución de genotipos compatible con las proporciones mendelianas esperadas bajo equilibrio:

- Mayor proporción de hembras (X^o X^o) (no naranjas).
- Baja proporción de hembras (X^O X^O) (naranjas sólidas).
- Proporción intermedia de hembras (X^O X^o) (calicó/tortie).
- Proporciones de machos (X^OY) y (X^oY) cercanas a (p) y (q).

Estas simulaciones sirven como referencia idealizada de cómo se distribuirían los genotipos en una población en equilibrio.

### 4.3 Resultados del modelo empírico

El modelo empírico (véase en el anexo C), al basarse en los genotipos observados, mostró:

- Una frecuencia relativamente alta de hembras (X^O X^o), consistente con la abundancia de descripciones tipo Calico y Dilute Calico en la base de datos.
- Diferencias en las proporciones de (X^o X^o) y (X^O X^O) respecto a las expectativas teóricas.
- Pequeñas discrepancias en la proporción de machos naranjas (X^OY) y no naranjas (X^oY).

Estas diferencias reflejan tanto el sesgo de muestreo propio de los refugios como los errores de clasificación de color.

### 4.4 Comparación entre modelo teórico y empírico

Al comparar las distribuciones de genotipos resultantes de ambos modelos se observaron diferencias claras (véase en el anexo D), especialmente en las categorías de hembras. Este comportamiento es coherente con el resultado de la prueba de chi-cuadrado: la población real no satisface los supuestos de Hardy–Weinberg. La comparación entre ambos enfoques permite ilustrar cómo una población real puede desviarse de la teoría por efectos de muestreo, selección y errores de medición.

## 5 Conclusiones

El gen (O) sigue un patrón de herencia ligada al cromosoma X, lo que genera distribuciones de color distintas entre machos y hembras. Esto se puede apreciar en las proporciones de cálculo de frecuencias alélicas. Al calcular estos datos, se determinó que la base de datos analizada presenta errores de clasificación, especialmente en machos etiquetados como Calico o Tortie, lo que indica que las descripciones de color son principalmente estéticas y no genéticas. La prueba de chi-cuadrado muestra que la población no cumple el equilibrio de Hardy–Weinberg, lo cual es esperable dada la procedencia de los datos y la ausencia de una muestra aleatoria.

El modelo teórico reproduce correctamente las proporciones genotípicas esperadas bajo equilibrio, mientras que el modelo empírico refleja el comportamiento real de la población observada. La comparación entre ambos modelos evidencia cómo las poblaciones reales pueden desviarse de las expectativas

mendelianas por factores como sesgos de muestreo, selección y errores en los registros. Debido a los resutados obtenidos, se puede concluir que a pesar de no cumplir el equilibrio de Hardy Weirberg, este modelo se puede enfocar en poblaciones reales de cualquier especie que no necesariamente estén equilibradas por causas de migración o cambios en su habitat.

# 6 Recomendaciones

Para crear un modelo probabilístico de manera exitosa a partir de la teoría Hardy-Weirberg es importante profundizar en la limpieza y validación de datos, incorporando criterios más estrictos para la clasificación de colores. Utilizar bases de datos más amplias y, en lo posible, con información genética confirmada y no solo descripciones visuales. Además, replicar el análisis en otras regiones geográficas o especies para comparar patrones de herencia en diferentes contextos poblacionales.

# 7 Bibliografía

All Science. (2024). *Crazy orange cats*. https://allscience.substack.com/p/crazy-orange-cats

LibreTexts. (2024). *Hardy–Weinberg principle of equilibrium*. https://bio.libretexts.org/

# 8 Anexos

## 8.1 Anexo A. Gráfico de chi-cuadrado

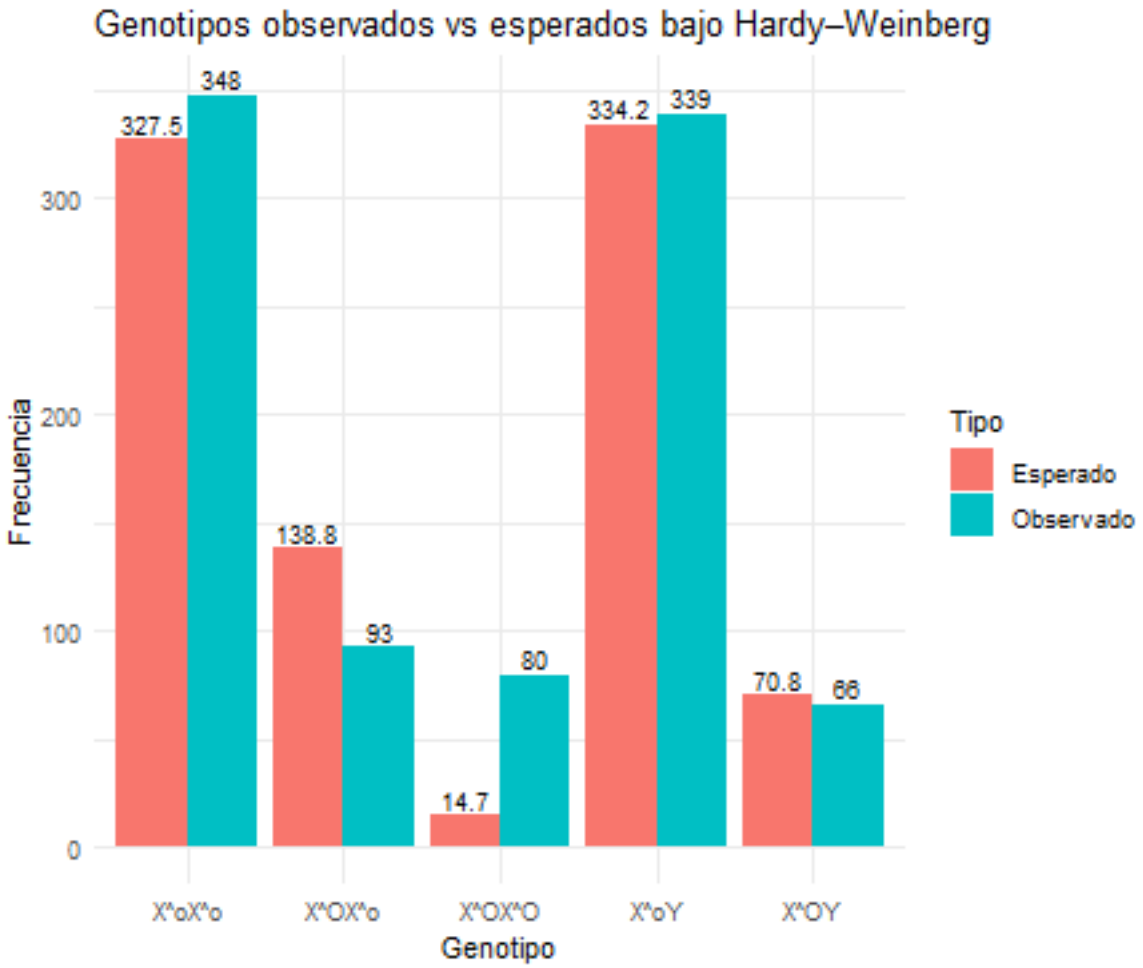


Figura 1: Gráfico de chi-cuadrado que compara valores observados y esperados.

## 8.2 Anexo B. Simulación teórica bajo Hardy–Weinberg

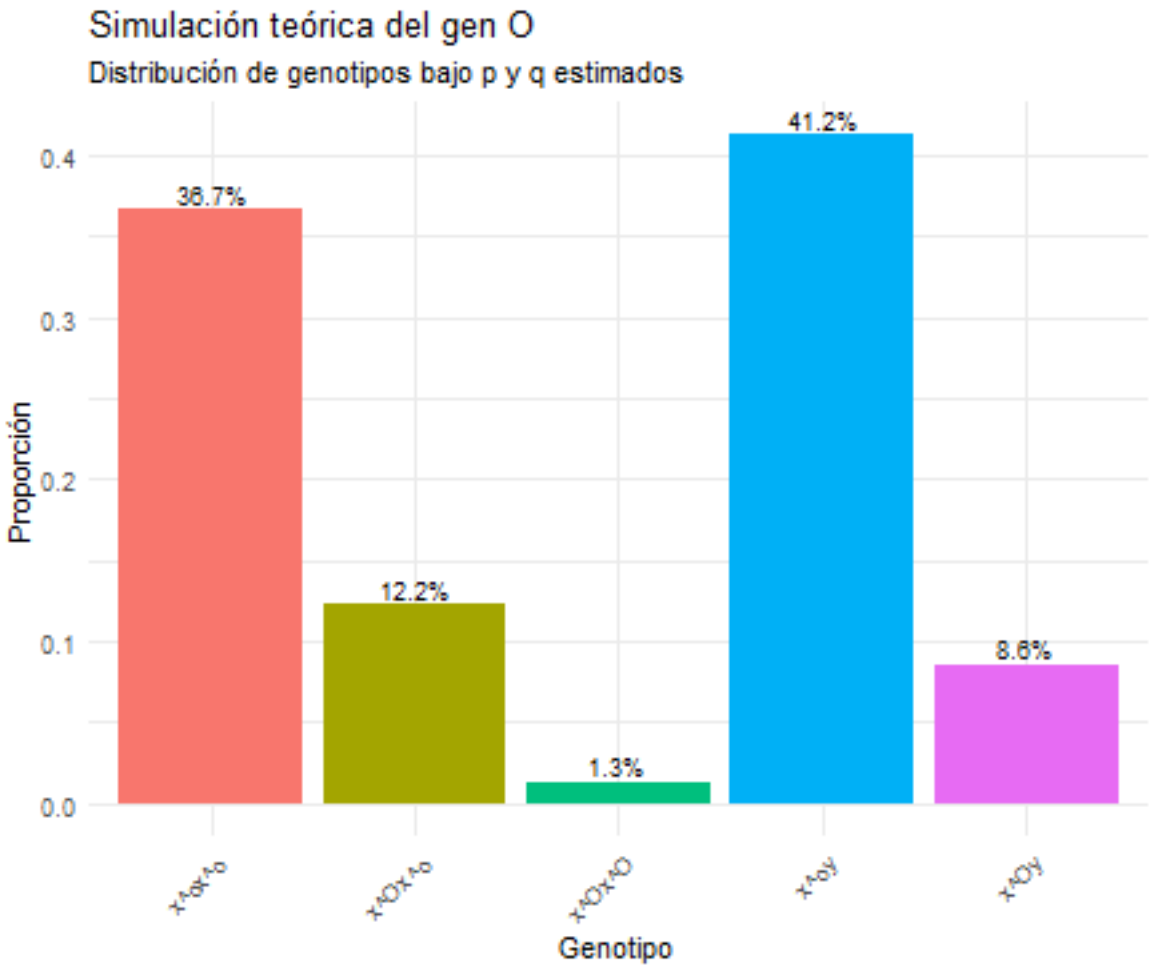


Figura 2: Distribución de genotipos en el modelo teórico.

## 8.3 Anexo C. Simulación empírica con datos reales

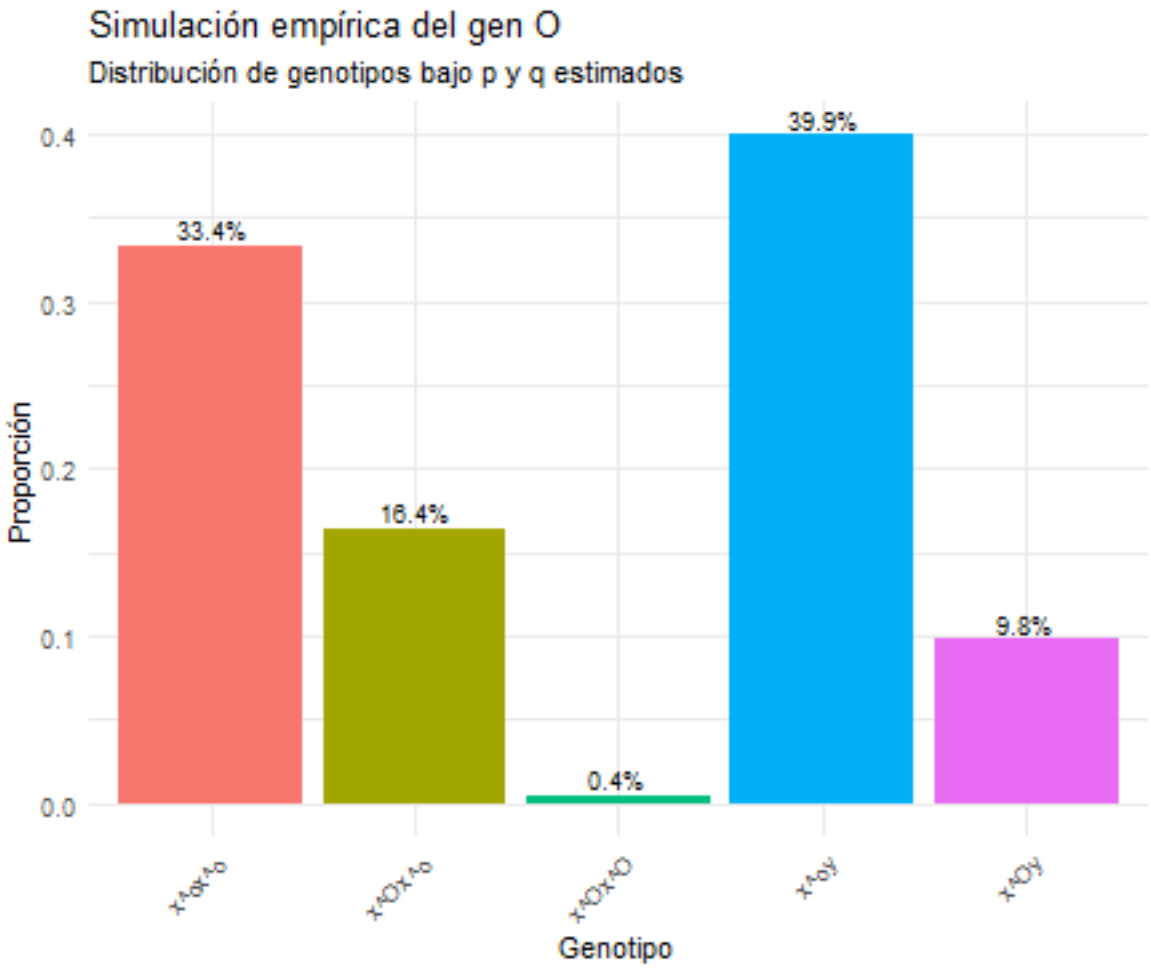


Figura 3: Distribución de genotipos basada en datos observados.

## 8.4 Anexo D. Comparación entre simulación teórica y empírica

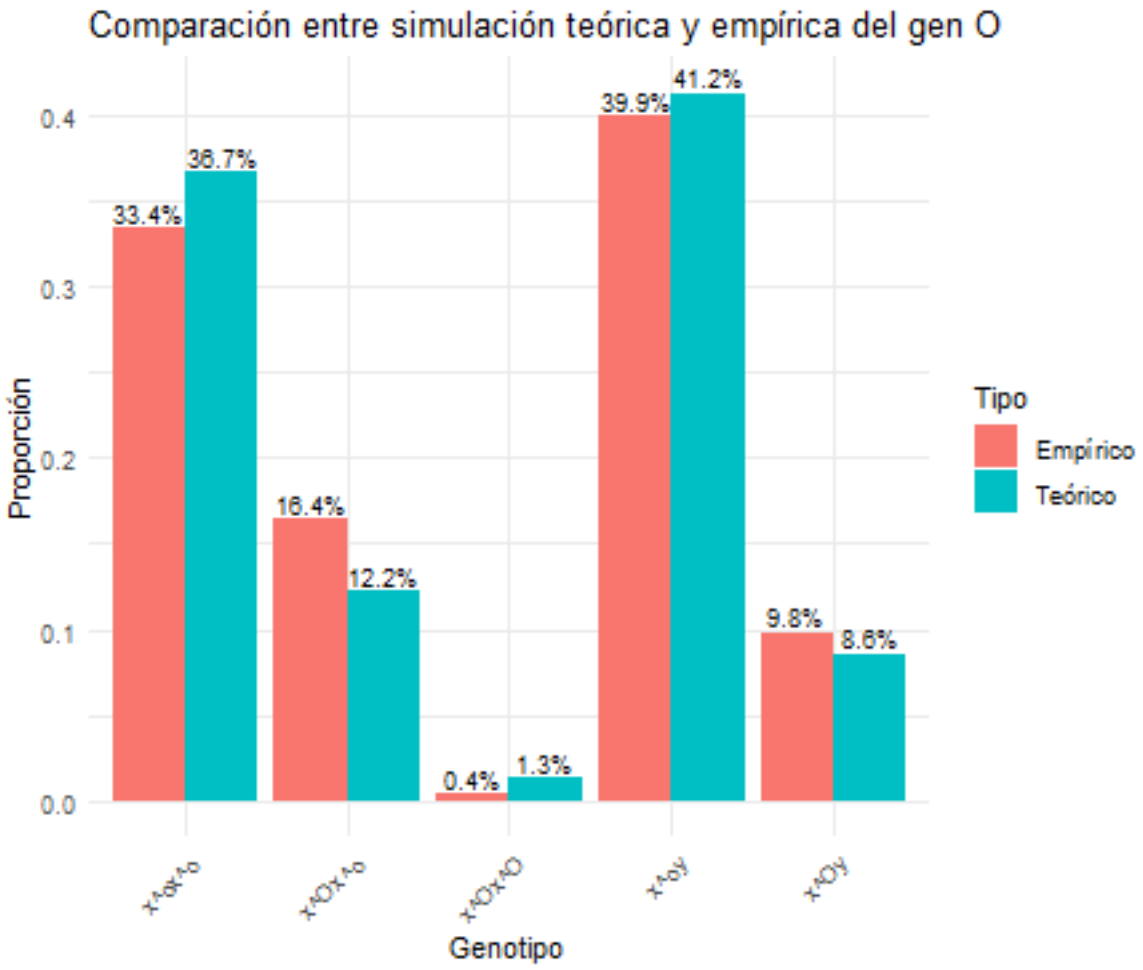


Figura 4: Comparación entre modelos teórico y empírico.