La duración de las canciones con los años

Joan Manel Garcia-Reyero Sais

Tutor:
Bernat Gascón Cabestany

Índice

Resumen

In this research we are going to study the duration of songs over the last sixty years. We start with an exploratory graphic analysis. The distribution of the songs' durations is almost Gaussian, with 4 min being the mode. We then create a graphic with the averages of the durations per year. In it we see that before 1970 the average song duration stayed between 2,5 min and 3,5 min. But in the seventies it completely changed, and after the seventies we get averages between 3,5 min and 4,5 min. The graphic of standard deviations shows us that an increase in the dispersion of song durations precedes the increase of the averages: in the sixties the deviations increase from values between 0 min and 1 min to values between 1,5 min and 2,5 min.

Finally, we introduce a hypothesis of what could have happened. A rule in radio stations, which said that songs could only last 2,5 min or less, kept the duration of songs short. During the sixties, with the creation of symphonic rock, the dispersion increases. Once in the seventies, the *2.5 minutes rule* gets cancelled and bands are able to produce longer songs. Adding this to a popularization of symphonic rock, we see the averages grow.

This is, in a synthetic way, our theory of the evolution of song duration. In further sections of the paper we defend it with graphics that help to see in a clearer way what was happening and why our theory could be true.

1. Hipótesis

Green Day es un grupo formado en los ochenta en los Estados Unidos. Un día, mientras estaba escuchando música, me fijé en que sus canciones recientes duraban más que las antiguas. Se me ocurrieron varias hipótesis por las cuales podría pasar esto: la primera, que las canciones de *Green Day* se habían hecho más comerciales a medida que se iba popularizando el grupo. Empezaron tocando punk, que es un género en el cual las canciones suelen ser más cortas, y acabaron tocando rock, una música más comercial donde son generalmente más largas. La segunda hipótesis que me vino a la cabeza fue que podría haber incrementado la duración de las canciones, en general, con el paso del tiempo. Para comprobar lo dicho fui a ver otros grupos que tenía en mi reproductor, unos más antiguos, como *Ramones* (1976-1996), y otros más modernos, como *My Chemical Romance*, fundado en el siglo XXI. Una vez hecha esta pequeña ojeada vi que mi hipótesis parecía cumplirse, los grupos más antiguos tenían canciones mucho más cortas que los modernos.

Elegí hacer este estudio para ver si lo que se cumplía en mis escasas 150 canciones pasaba con todas las canciones relativamente modernas. (Entendiendo como moderna cualquier canción compuesta en los últimos cincuenta años).

Este estudio hará un análisis estadístico basado en los datos de *Million Song Dataset* y analizará la hipótesis de que la duración de las canciones ha aumentado a medida que avanzamos en el tiempo. Una vez analizados todos los gráficos vemos que desde los años sesenta a los ochenta hay un incremento importante de las medias de las duraciones por año, y por tanto nos vamos a centrar en este período principalmente.

2. Obteniendo la información

La base de datos usada ha sido descargada de una página web llamada *Million Song Dataset (labrosa.ee.columbia.edu/millionsong)*, que pertenece a una fundación llamada *Labrosa*.

Según su página principal *Million Song Dataset* es: *"Una colección gratuita de metadata de un millón de canciones contemporáneas y populares"*. Como nuestro objetivo es hacer un análisis estudiando si la duración de las canciones ha crecido con el tiempo, necesitamos una base de datos donde haya cuantas más canciones mejor. Y que contenga, además, canciones lo

más antiguas posible. La entrada de cada canción debe tener la duración y la fecha en que fue compuesta. La base de datos encontrada cumple todos los requisitos: contiene un numero elevado de canciones (un millón, aunque una vez quitamos las que no tienen el año o la duración nos quedan 515394).

Solo nos faltan un par de condiciones por comprobar: si es un proyecto fiable y si las canciones son de géneros heterogéneos, ya que si no, no estaríamos estudiando la evolución de la duración de la música, lo haríamos de géneros concretos, y la muestra no sería válida.

Siempre siguiendo lo que pone en su página web, han creado la base de datos por diversos motivos: para animar la investigación de algoritmos; para aportar una base de datos referente para trabajos de investigación; como un apoyo para crear una base de datos más grande, y para ayudar a arrancar a los nuevos investigadores. Esto nos da una primera idea de que es un proyecto serio, aunque hay dos datos más que dan soporte a este hecho.

El primero es que hay más de una página web que recomienda *Million Song Dataset*.

El segundo es que son clientes de *The Echo Nest (http://the.echonest.com/)*, que es una red que acumula datos de canciones para poder hacer aplicaciones e investigación. Estos ya tienen más de 35 millones de canciones, de más de dos millones de artistas diferentes, y tienen como clientes compañías como Twitter, Spotify, MTV, la BBC o VEVO. Por tanto, si *Million Song Dataset* ha sacado buena parte de su información de *The Echo Nest* será un proyecto fiable.

Respecto a nuestra segunda condición, la que dictaba que debía haber variedad de géneros, encontramos la respuesta en la sección de *FAQ (Frequently Asked Questions)* de la propia página web, que nos explica el complejo proceso de extracción de datos. Un 48 % de las canciones las extrajeron mirando quiénes eran los 100 artistas más populares en cada año según *The Echo Nest*. A partir de estos artistas fueron enlazando a otros mediante un proceso aleatorio, y descargando los datos de sus canciones. Al haber un elemento aleatorio, y dado el hecho que el número total final de artistas es mucho mayor que los 100 primeros, nos hace deducir que en la base de datos hay diferentes géneros.

Million Song Dataset cumple todos los requisitos para ser la base de datos en la que se va a basar el trabajo, por lo tanto solo falta descargar su base de datos y eliminar las canciones que no contienen los datos que nos interesan.

3. Método de trabajo

Tras eliminar todas las canciones que no contienen todos los datos que nos interesan, nos quedan 515576 canciones, que sigue siendo una cifra muy elevada. Por lo tanto, queda totalmente descartada la opción de manipular los datos a mano o con Excel, ya que el máximo de filas que puedes introducir con éste son 65536. Teniendo en cuenta todos estos factores, opté por hacerlo en Python, un lenguaje para programar el ordenador. Lo primero que tuve que hacer fue asignar todas las duraciones a una lista de Python, y las fechas a otra. Ambas listas quedarán ordenadas de forma que la fecha de la canción número i quede en la posición $i-1$ (en Python se empiezan a nominar las listas por el 0), de la lista de fechas, y la duración de esa misma canción en la posición $i-1$ de la lista de duraciones. Por tanto nos quedará algo parecido a esto:

```
duraciones = [duracion_1, duracion_2, ... , duracion_n]
fechas = [fecha_1, fecha_2, ... , fecha_n]
```

3.1. Gráfico utilizando Matplotlib

Al pensar hoy en un teléfono, poca gente se lo imagina sin aplicaciones que lo complementen. Pasa igual con un lenguaje de programación. Existen miles de librerías, programas que están en internet que hacen la función para la que están programados, de manera que el usuario de Python solo debe descargarlas, importarlas y "llamarlas" en su propio programa para que hagan la función deseada. Hay una librería de Python llamada *Matplotlib* para graficar datos; solo se le introducen los datos de tu gráfico, señalando cual es la variable independiente y la dependiente y *Matplotlib* dibuja el gráfico. A continuación, tenemos un ejemplo del código necesario para hacer un gráfico simple:

```
import matplotlib as plt

lista_x = [1, 2]
lista_y = [4, 8]

plt.hist(llista_x, llista_y)
plt.show()
```

Lo primero es importar *Matplotlib*; el `as plt` nos permite llamarlo como `plt` en vez de `matplotlib`, tan solo una cuestión de comodidad. A continuación definimos las listas que serán los diferentes puntos que aparecerán en el gráfico, atribuyendo a la `lista_x` los parámetros independientes y a la `lista_y` los dependientes. Por lo tanto, nos quedarán los puntos $(1, 2)$ y $(4, 8)$. A continuación, le decimos a *Matplotlib* que nos cree un histograma (`hist`) con los integrantes de las listas, y finalmente le decimos que nos lo muestre. Para tener un buen gráfico aún nos quedaría ajustar las dimensiones y la escala, pero esos detalles nos desviarían demasiado del tema.

4. Gráfico de distribución de las duraciones

Lo primero que hay que hacer es un gráfico de distribución de las duraciones para hacernos una primera idea de como estarán repartidas. Lo haremos utlilizando *Matplotlib* y con un código parecido al del ejemplo anterior pero con pequeñas modificaciones.

Al gráfico solo le entramos una lista, a diferencia del anterior que le habíamos entrado dos, ya que para esta figura solo nos interesan las duraciones de las canciones. También hay que decirle como queremos que nos dibuje las barras del histograma. No hay un criterio inteligible para saber de cuantos "bins" deben ser, pero después de probar varias veces me decidí por 300.

```
plt.hist(duraciones, bins = 300)
```

En la siguiente línea le damos el título del gráfico, y dentro de éste, queremos que nos diga el número de canciones con las que tratamos. El porcentaje le indica a Python que allí hay una palabra que aún está por definir. La definimos en el segundo porcentaje como `len(duraciones)`. *len* proviene de la palabra inglesa *length*, que significa longitud. Por lo tanto el `len(duraciones)` nos va a dar el número de duraciones que tenemos, que va a ser igual al número de canciones.

```
plt.title("Distribucion global (%d canciones)" % len(duraciones))
```

En la siguiente línea decidimos la longitud de los ejes. La forma de saber la longitud buena es ir probando, hasta que tienes una que te engloba todos los datos. Los dos primeros números son para el eje de las x, que va a ir de

0 min a 15 min. Los dos últimos nos indican el eje de las y, que van de 0 a 35000 canciones.

```
plt.axis([0, 15, 0 , 35000])
```

A continuación le pondremos el título al eje de las x (`xlabel`) y el de las y (`ylabel`)

```
plt.xlabel("Duraciones (minutos)")
plt.ylabel("Numero de canciones")
```

Finalmente le pedimos que nos lo enseñe (`plt.show`), y que se cierre (`plt.close`). El gráfico nos queda tal como vemos en la figura 1.

```
plt.hist(durades, bins = 300)
plt.title("Distribucion global (%d canciones)" % len(duraciones))
plt.axis([0, 15, 0 , 35000])
plt.xlabel("Duraciones (minutos)")
plt.ylabel("Numero de canciones")
plt.show()
plt.close()
```

La figura 1 nos muestra que hay muy pocas canciones cortas, y a la vez pocas muy largas. El dibujo que nos queda tiene un cierto parecido con una distribución normal o *gausiana*, aunque no lo es ya que no acaba de ser simétrica, se nos disparan las canciones de larga duración. El rango es [0,083, 50], por tanto va de 0,083 min, es decir 5 s, a 50 min. La media de la duracion es de 4,11 min.

5. Media y desviación tipo

Ahora ya hemos visto cómo están distribuidas las duraciones y tenemos una idea del rango. Los siguientes pasos serán hacer un gráfico con la media de las duraciones de cada año para tener una idea general de cómo han evolucionado las duraciones en el tiempo. Este gráfico lo complementaremos con otro de las desviaciones tipo de cada año, ya que podría ser que un año nos dé una media muy alta a causa de unas pocas canciones muy duraderas. Antes de hacer los gráficos hay que hacer un pequeño código en Python que nos haga las medias y las desviaciones tipo.

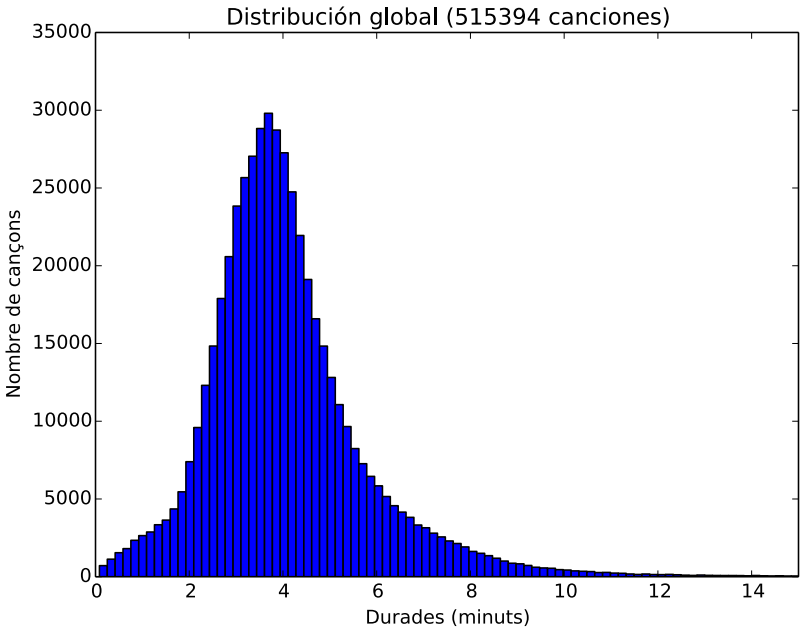

Figura 1: Distribución global

Lo que tenemos ahora mismo es una lista con todas las duraciones y las fechas ordenadas correlativamente. Pero para hacer la media y la desviación tipo necesitamos un diccionario. Los diccionarios ordenan sus entradas usando claves, que en nuestro caso serán los años. Cada clave tiene un contenido que la acompaña, que va a ser una lista con las duraciones de las canciones de ese año. Por lo tanto hemos de crear el diccionario.

```
def add_to_dic(dic, year, duration):
    if year in dic:
        dic[year].append(duration)
    else:
        dic[year] = [duration]

def build_dic(years, durations):
    dic = {}
    i = 0
```

```
    while i < len(years):
        add_to_dic(dic, years[i], durations[i])
        i = i + 1
    return dic

dic = build_dic(anys, durades)
```

Las funciones son lo que se utiliza en Python para hacer un algoritmo que desarrolle una acción. Nosotros hemos creado dos funciones. La primera mira en el diccionario, llamado `dic`, si el año en el que estamos ya es una entrada en `dic`. Si lo es, añadirá a la lista que esté relacionada con ese año la duración correspondiente al año que hemos utilizado primero. Si el año no está en el diccionario creará una entrada con ese año y le relacionará una lista con un solo elemento, la duración correspondiente al año. Lo que hace la segunda función es hacer el procedimiento que hace la primera con todas las entradas de la lista para así construir el diccionario.

5.1. Media

Ahora ya tenemos la información que nos interesa para calcular la media,

$$\overline{x} = \frac{\sum x_i n_i}{n}.$$

Ya podemos proceder a calcular la media de las duraciones de cada año. Primero hay que hacer una función que nos haga la media de diferentes elementos de una lista.

```
def average(array):
    if len(array) == 0:
        return None
    total_sum = sum(array)
    return total_sum / float(len(array))
```

Suma todos los componentes de la lista y divide el resultado por el número de componentes de la lista. Si el `len` de la lista es 0, nos devuelve `None`, ya que no hay elementos.

Ahora nos hace falta que haga la media de cada una de las listas guardadas en cada año del diccionario.

```
def final_average(years, durations):
    a_years = []
    averages = []
    for e in range(min(years), max(years)):
        if e in dic:
            a_years.append(e)
            averages.append(average(dic[e]))
    return a_years, averages
```

Este programa recorre todas las entradas del diccionario y por cada entrada llama a la función que nos hacía la media, y la calcula de todos los componentes de la lista relacionados con esa entrada. Finalmente, coge todos los años y los pone en una lista (`a_years`), y las medias en otra lista (`averages`), correlativa con los años.

Ya le podemos pedir a *Matplotlib* que nos dibuje el gráfico,

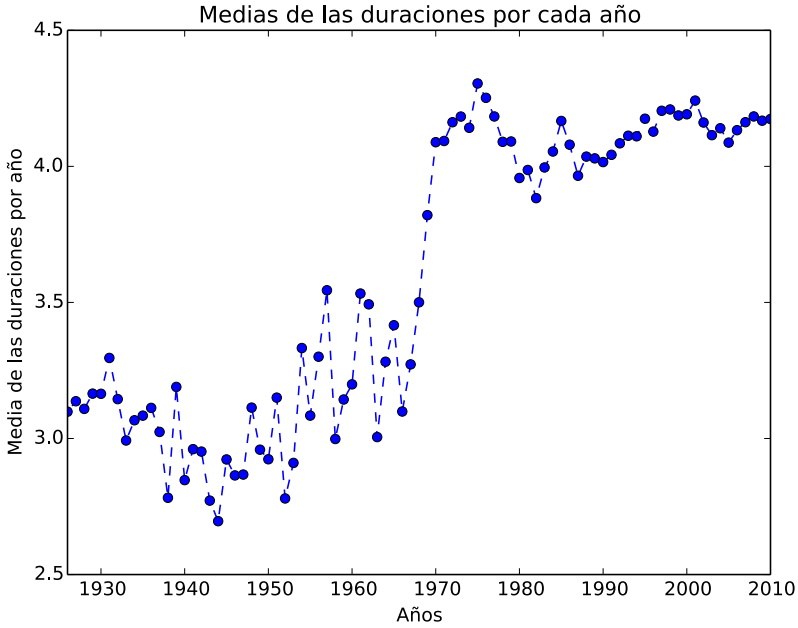

Figura 2: Media de las duraciones por año

En la figura 2 vemos que las medias de todos los años varían entre 2.5 i 3.5

hasta el 1970, aunque justo en aquel año pasa un hecho muy curioso: la media de las duraciones sube de manera notable para situarse entre 4min y 4,5min y estabilizarse allí, llegando a un máximo de 4,3min. Esto podría haber pasado por diferentes hechos de tipo social: podría ser que de repente hubiera habido un cambio de tendencias debido a hechos que hasta el momento desconocemos y que hayan hecho aumentar las duraciones de las canciones; o podría haber ayudado la creación de un nuevo género musical con las canciones muy largas, tanto que hagan subir las medias. Si fuera solo esto el gráfico no sería tan representativo como parece al principio. Para salir de dudas utilizaremos el gráfico de las desviaciones tipo.

5.2. Desviación tipo

Por lo tanto, antes de ponernos a investigar sobre los cambios que padeció la música en los setenta hay que hacer el gráfico de la desviación tipo de cada año,

$$\sigma = \frac{\sum (x_i - \overline{x})^2}{n},$$

para ver si ya podemos descartar alguna de las teorías anteriores.

Para hacerlo podemos aprovechar el código que hemos hecho anteriormente para hacer el diccionario `dic`. Entonces solo nos queda hacer un programa que haga las desviaciones tipo de todas esas entradas.

```
def stdev(array):
    media = average(array)
    calc = []
    for e in array:
        c = e - media
        y = pow(c, 2)
        calc.append(y)
    return math.sqrt(sum(calc) / len(array))
```

Este programa resta la media calculada anteriormente a todas las entradas de la lista, y eleva el resultado al cuadrado. Una vez ha hecho este paso en todas las entradas, hace el sumatorio de todos los resultados y divide el resultado de éste por el numero de integrantes en la lista. Para acabar hace la raíz cuadrada del producto de la división para darnos la desviación tipo.

```
def final_deviation(years, durations):
    s_years = []
    deviations = []
    for e in range(min(years), max(years)):
        if e in dic:
            s_years.append(e)
            deviations.append(stdev(dic[e]))
    return s_years, deviations
```

Lo que hace esta segunda función es, para todas las listas asociadas a las entradas de un diccionario, hacer la desviación tipo y ponerla en una lista ordenada correlativamente al año al cual pertenece.

Ahora ya sí que tenemos hecho el programa y ya le podemos decir a *Matplotlib* que nos haga el gráfico. Figura 3

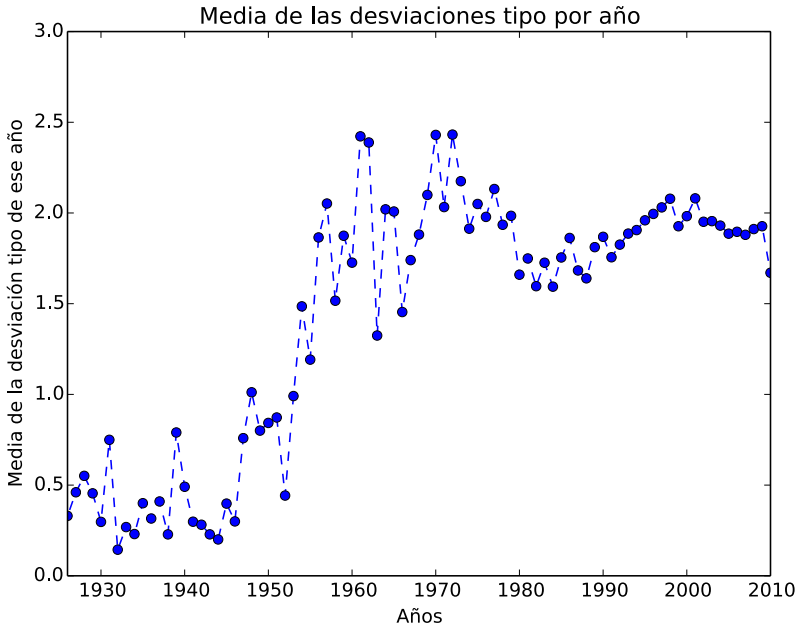

Figura 3: Media de las desviaciones tipo por año

6. Análisis de los resultados

Como hemos visto antes, observar cada gráfico de manera individual no tiene demasiado sentido, ya que el uno sin el otro no está completo. Si observamos la relación entre los dos podemos ver diversos aspectos interesantes. Primero de todo dividiremos nuestros datos en tres grupos: los *Primeros Años*, des del inicio a los sesenta; *la Transición*, de los sesenta a los ochenta, y los *Últimos años*, de los ochenta hasta la actualidad.

6.1. Primeros años

Consideramos las canciones anteriores a los 60 como las más antiguas. Lo que vemos es que las medias varían entre 2,5 min y 3,5 min, y que las desviaciones tipo lo hacen entre 0 min i 1 min. Este primer dato nos muestra que todas las canciones antiguas tienen una media muy parecida, y que todas tienen desviaciones pequeñas, por lo tanto no solo todas las canciones eran generalmente cortas, sino que todas tenían duraciones parecidas.

Una de las hipótesis de por qué pasa esto es que todas las canciones de estos escasos veinte años tenían una duración parecida, pero que había algunas canciones largas que hacían variar la media, y hacían subir un poco el valor de la desviación tipo. Si esta hipótesis fuera cierta implicaría que si la media es más grande de lo habitual en este periodo la desviación tipo también lo sería. Entonces si hiciéramos un gráfico de las desviaciones respecto las medias, nos tendría que dar una figura creciente.

Una vez hecha la figura 4 vemos que no nos queda ninguna figura definida, y por lo tanto no hay una relación directa en los primeros años entre la media y la desviación tipo.

Una vez descartada esta hipótesis solo nos queda remarcar lo obvio: las canciones son generalmente cortas, y no hay mucha diferencia entre diferentes duraciones de canciones. Las razones por las que las canciones son cortas son diversas. Por una parte había presión de los promotores musicales, ya que solo dejaban poner canciones de 2,5 min o menos en la radio musical. Otra razón es porque en los discos en donde estaban grabadas las canciones tenían muy poca capacidad, por lo tanto no podían hacerlas muy largas si querían poner más de un par de canciones en los discos.

Lo que podemos hacer para acabar de completar esta parte es coger un año al azar y hacer la distribución general. El elegido va a ser el 1955.

En la figura 5 vemos de una manera más representativa lo que nos dicen

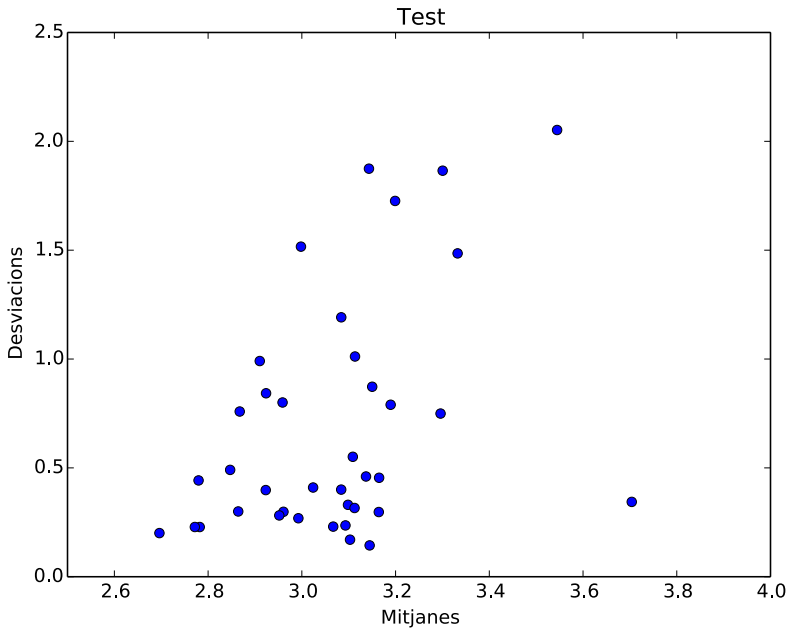

Figura 4: Relación entre las desviaciones y las medias

los gráficos de las medias *(Figura 2)* y de las desviaciones tipo *(Figura 3)*. La mayoría de canciones se encuentran entre los 2 min y los 4 min, aunque tenemos muchas más canciones en él hacia el dos que hacia el cuatro.

6.2. La transición

En los sesenta las desviaciones tipo se disparan, haciendo subir solo un poco las medias. En cambio en los setenta la duración general de las canciones aumenta.

En los años sesenta las desviaciones tipo aumentan considerablemente *(Figura 3)*, pero no lo hacen así las medias *(Figura 2)*, que es verdad que incrementan un poco, aunque no de una manera tan notable. En los setenta, en cambio, las desviaciones tipo siguen más o menos igual que en los sesenta, pero las medias de las duraciones aumentan mucho. En este punto nos vamos a fijar en este incremento, tanto en las medias como en las desviaciones. Primero de un modo numérico y luego histórico.

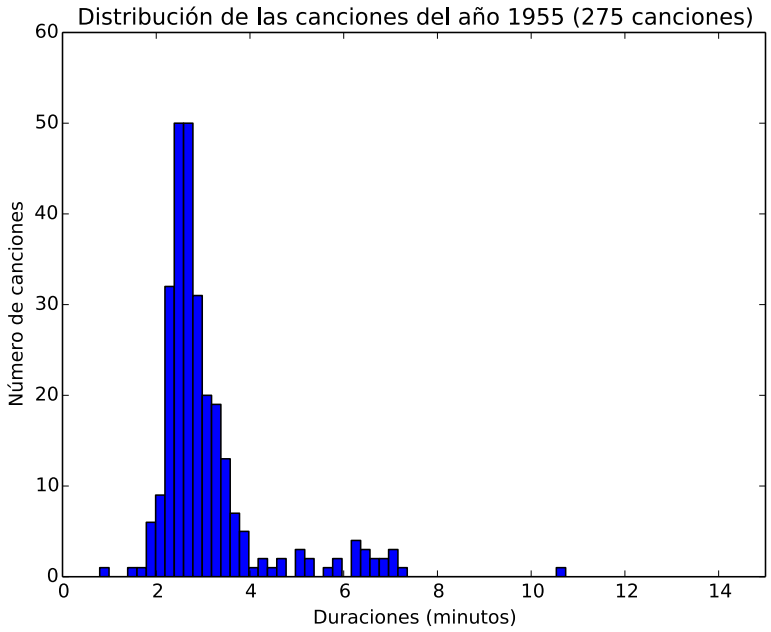

Figura 5: Distribución de las duraciones en 1955

6.2.1. Análisis numérico

Haremos primero un gráfico para ayudar a hacer esta análisis numérica, que nos va a servir para la histórica también.

En la figura 6 vemos como en el año 1965 la gran mayoría de los valores se encuentran entre los 2 min y los 3 min, cosa que cumple lo que veíamos en el gráfico de las medias *(Figura 2)*. Vemos también que a diferencia del gráfico de 1955 mostrado anteriormente, hay una serie de valores extremos que hacen variar la desviación tipo, sin hacer subir de manera tan notable la media. En cambio en 1971, año en que empiezan a aumentar las duraciones, vemos que la mayoria de canciones se concentran entre los 3 min y los 4 min, cosa que nos hace subir las medias.

6.2.2. Análisis histórico

El rock sinfónico empieza a aparecer en los sesenta, aunque al principio la producción era mínima, se hizo popular en la década siguiente. Las canciones

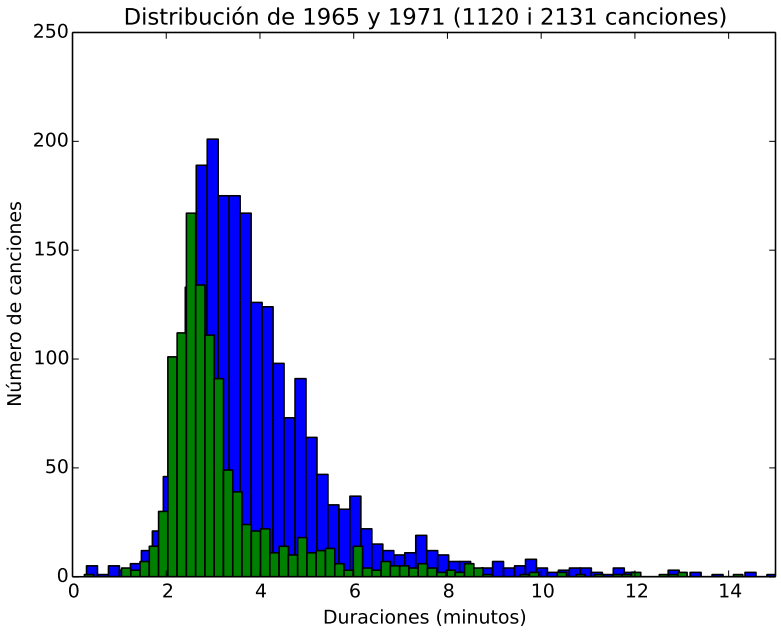

Figura 6: Comparación de las duraciones de los sesenta (verde) y de los setenta (azul)

de este género musical eran muy largas. Esta pequeña producción de rock sinfónico, que puede venir también acompañado de otros géneros de larga duración, nos puede hacer variar mucho la desviación tipo que hasta los sesenta era tan pequeña. Además, si la producción aumentara nos podría hacer cambiar la media. A partir de esta pequeña explicación podemos extraer una hipótesis de porque aumentan así las medias.

Antes de los sesenta había una norma en las radios musicales que decía que no podían poner canciones que duraran más de 2,5min. Muchas canciones más duraderas tenían una versión más corta para emitir en la radio, y es que convertir una canción de 3min a 2,5min no es difícil. En cambio convertir una canción de 7 min, que es lo que solían durar las canciones de rock sinfónico, a 2,5 min es casi impossible. Tampoco había ninguna otra forma de hacer popular tu música que fuera sencilla y gratuita para el oyente, ya que la falta de Internet y televisión lo dificultaba mucho. El rock sinfónico se empezó a popularizar en los setenta, coinciciendo con la suspensión de esta norma.

La derogación de *La norma de los dos minutos* comportó que no hubie-

ra una duración máxima para las canciones que sonaban en la radio. Esto comportó una serie de cambios.

En primer lugar, la popularización del rock sinfónico, con grupos como *Pink Floyd*, que acabaron teniendo una gran popularidad. Esto contribuyó a subir el valor de las medias *(figura 2)* de cada año, ya que cuantas más canciones largas haya, más alta será la media.

Esta teoría sugiere una pregunta: si el número de canciones duraderas sube, ¿Por qué no hay un cambio radical de los valores de las desviaciones tipo de forma consecutiva? Como la *Norma de la radio* quedó derogada, los artistas tenían mucha más libertad para poderse expresar y, a la vez, darse a conocer a través de la radio. Este hecho nos lleva a una subida general de la duración de las canciones. Por lo tanto, el aumento general de las medias, y que las desviaciones tipo no lo hagan, queda explicado, ya que la subida general de las duraciones contrasta con la de los estilos de música con duraciones extremadamente altas que están en auge en este período.

En el gráfico de distribuciones que hemos hecho antes *(Figura 3)* vemos que todas las fechas históricas dadas coinciden con el gráfico. Por lo tanto no podemos refutar esta hipótesis.

6.3. Últimos años

A partir de 1980, las medias de estabilizan y empiezan a oscilar entre 4 min y 4,5 min. Las desviaciones tipo bajan un poco respecto a la década anterior, aunque siguen más altas que en los sesenta. Esto se debe a una estabilización de las canciones de larga duración.

Si miramos el gráfico de las medias y de las desviaciones tipo de los últimos diez años vemos que casi no varían. Esto implica que casi todas las canciones duraban entre 6 min y 2 min, ya que la media es cuatro y la desviación tipo es dos. Es un hecho curioso que todas las canciones estén en unos parámetros tan marcados.

Si hacemos un gráfico de la distribución de las canciones de los últimos diez años, el 2005 por ejemplo, vemos que se cumple la tendencia de los gráficos anteriores.

En la figura 7 vemos que la mayoría de los valores se concentran entre los 2 min y los 6 min, y la moda es 4. Por lo tanto los gráficos de las desviaciones tipo y de las medias nos demuestran que los gráficos de todos los años tendrán una estructura parecida a éste.

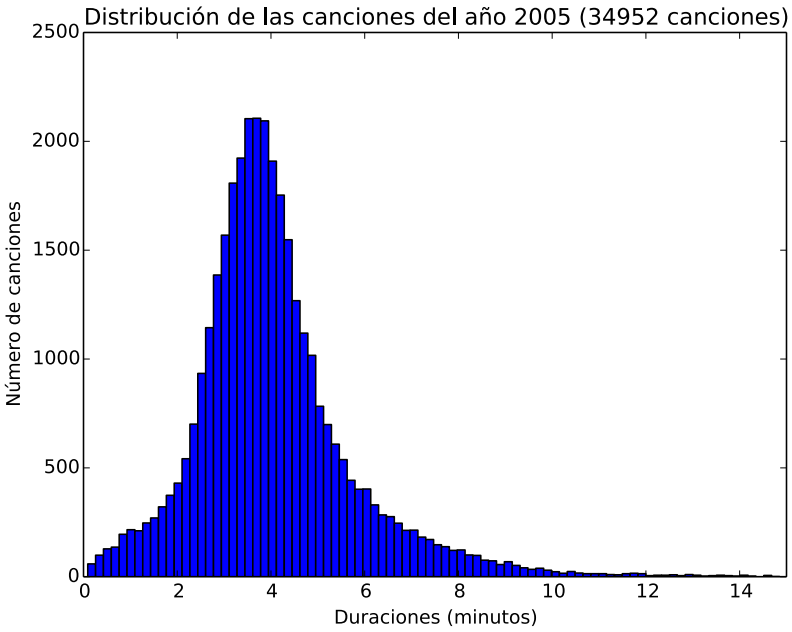

Figura 7: Distribución de las duraciones en el 2005

Por lo tanto las canciones modernas, considerando moderno desde 1980 hacia adelante, se pueden dividir en dos grupos según su desviación tipo: antes y después del 2000. Como ya hemos visto un gráfico posterior al 2000, lo vamos a comparar con uno anterior, como el 1990.

Vemos en la figura 8 que casi no cambia respecto a la figura 7. Hay pequeñas variaciones, pero la forma de las barras del histograma es muy parecida. Vemos en el gráfico de las desviaciones tipo *(figura 3)* que hay una pequeña variación entre el 1990 y el 2005, pero que cuando lo pasamos a una distribución la diferencia es poco notable. Por lo tanto, podemos afirmar que no solo las canciones de los últimos diez años tienen una distribución muy parecida, sino también las de los últimos treinta, es decir, desde los ochenta hasta la actualidad.

Aquí podríamos encontrar otra hipótesis histórica: en los ochenta se crea la primera cadena de televisión dedicada exclusivamente a la música. A partir de este momento se empieza a crear una red, tanto en la televisión como en Internet unos años más tarde, que permitirá al artista darse a conocer. Esto le va a dar total libertad a los artistas para hacer música de diferentes

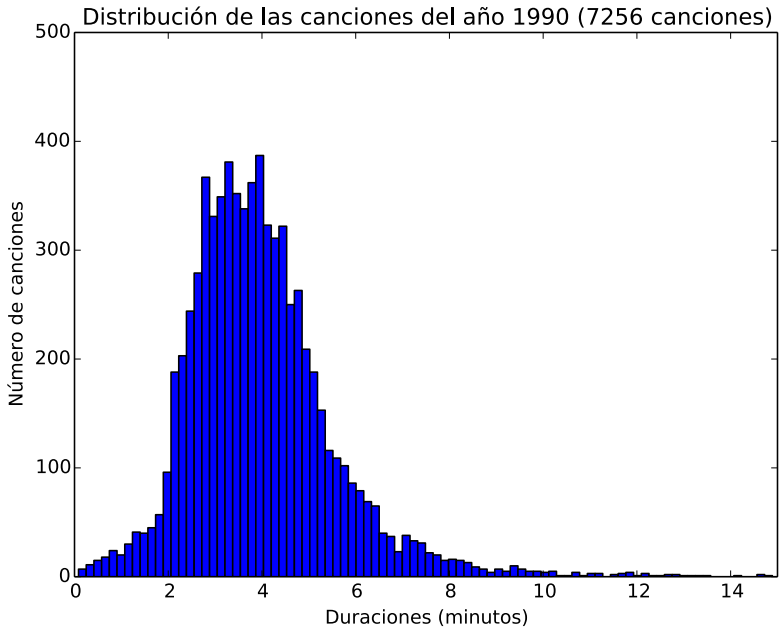

Figura 8: Distribución de las duraciones del 1990

duraciones y que todas sean aceptadas por los medios. Seguirán siendo la moda las canciones entre 3,5min y 4min, aunque habrá muchas más canciones con duraciones más extremas, tanto largas como cortas. Esta estructura se mantendrá hasta la actualidad.

7. Conclusiones

La hipótesis que hemos presentado al inicio del trabajo era cierta: las duraciones de las canciones aumentan de modo general a medida que avanzan los años. Hemos visto primero un gráfico de las distribuciones que ya nos sugería que había muchas más canciones de 3 min y 4 min que de 2 min, por tanto ya nos daba una primera idea de lo que podía acabar pasando.

A continuación, el gráfico de las medias de las duraciones de cada año nos dejaba bien claro que las duraciones de las canciones aumentan a medida que nos acercamos a la actualidad. No lo hacen de la manera que en un principio pensábamos que harían, no nos esperábamos que las antiguas estuvieran

tan igualadas entre sí, igual que pasa con las modernas, y que hubiera un incremento tan significativo entre las dos, la llamada transición. Después el gráfico de las desviaciones tipo complementaba al de las medias. Éste también tenía una subida muy grande, pero aumentaba en los sesenta, antes que el de las medias que aumentaba en los setenta.

Después de hacer una investigación conseguimos encontrar una explicación histórica por la que esto sucedía. La aparición del rock sinfónico en 1960 y su popularización en 1970, y la derogación de *La norma de los dos minutos*, también en 1970 son la clave.

A partir de aquí dividimos nuestros datos en tres partes, las más antiguas, desde el inicio de los sesenta; *La Transición*, de los sesenta a los ochenta; y finalmente las más modenas, de los ochenta hacia adelante. A partir de aquí hacemos una serie de gráficos centrándonos en diferentes años, tanto para poder presentar de manera visual lo que pasa en los dos períodos, como para comprobar nuestra teoría. Toda la información que nos dan los gráficos la apoyan, de manera que la acabamos dando por buena.

En resumen, hemos visto cómo evoluciona la música y que nuestra hipótesis inicial era cierta. Hemos conseguido una buena base de datos para sacar unos gráficos y unas conclusiones sólidas.

A. Código

```
# -*- coding: utf-8 -*-
import math
import trackdb
import matplotlib.pyplot as plt
meta = trackdb.read('track-meta.db', ['year', 'duration'])
nl_years = list(meta['year'])
nl_durations = list(meta['duration'])


#  Poping duratins < 5'' #

years = []
durations = []

def pop_durations():
    c = 0
    while c < len(nl_durations):
        if nl_durations[c] >= 5:
            durations.append(nl_durations[c])
            years.append(nl_years[c])
        c = c + 1
    return durations

pop_durations()



#  CONVERTING DURATIONS FROM SECONDS TO MINUTES  #

def minutes(durations):
    i = 0
    while i < len(durations):
        durations[i] = durations[i] / 60
        i = i + 1
    return durations
```

```
durations = minutes(durations)



# GENERAL GRAPHIC  #

# graphic of the distribution of all the durations

plt.hist(durations, bins = 300)
plt.title(u"Distribucio global (%d cancons)" % len(durations))
plt.axis([0, 15, 0 , 35000])
plt.xlabel(u"Durades (minuts)")
plt.ylabel(u"Nombre de cancons")
#plt.show()
plt.savefig('general.pdf')
plt.close()



#  AVERAGE  #

# calculing the average of...

def average(array):
    total_sum = sum(array)
    if total_sum == 0:
        return None
    return total_sum / float(len(array))


# STANDARD DEVIATION  #

# function for the standart deviation of...

def stdev(array):
    media = average(array)
    calc = []
    for e in array:
```

```
        c = e - media
        y = pow(c, 2)
        calc.append(y)
    almostfinished = sum(calc) / len(array)
    return math.sqrt(almostfinished)



#AVERAGE OF THE DURATIONS PER YEARS #

# looking to the general list, and creating
#a dictionary of all the years, and the durations
#related to them

def add_to_dic(dic, year, duration):
    if year in dic:
        dic[year].append(duration)
    else:
        dic[year] = [duration]

def build_dic(years, durations):
    dic = {}
    i = 0
    while i < len(years):
        add_to_dic(dic, years[i], durations[i])
        i = i + 1
    return dic

build_dic(years, durations)

dic = build_dic(years, durations)



# calculing the average of all the duratons related to its year

def final_average(years, durations):
```

```
    a_years = []
    averages = []
    for e in range(min(years), max(years)):
        if e in dic:
            a_years.append(e)
            averages.append(average(dic[e]))
    return a_years, averages

#print final_average(years, durations)



# GRAPHIC OF THE AVERAGE OF THE DURATIONS PER YEARS #


a_years, averages = final_average(years, durations)

plt.plot(a_years, averages, linestyle='--', marker='o', color='b')
plt.axis([1926, 2010, 2.5, 4.5])
plt.title(u"Mitjana de les durades per any")
plt.xlabel(u"Anys")
plt.ylabel(u"Mitja de les durades d'aquell any")
#plt.show()
plt.savefig('averages.pdf')
plt.close()



# GRAPHIC OF THE STANDART DEVIATION PER YEARS  #


def final_deviation(years, durations):
    s_years = []
    deviations = []
    for e in range(min(years), max(years)):
        if e in dic:
            s_years.append(e)
```

```
            deviations.append(stdev(dic[e]))
    return s_years, deviations

#print final_average(years, durations)


s_years, deviations = final_deviation(years, durations)

plt.plot(s_years, deviations, linestyle='--', marker='o', color='b')
plt.axis([1926, 2010, 0, 3])
plt.title(u"Mitjana de les desviacions tipus per any")
plt.xlabel(u"Anys")
plt.ylabel(u"Mitja de les desviacions tipus d'aquell any")
plt.savefig('deviation.pdf')
#plt.show()




# GRAPHIC OF THE DISTRIBUTION OF THE DURATION OF THE SONGS PER YEAR #

def search(year, dic):
    return dic[year]


fiveties = search(1955, dic)
#print fiveties

sixties = search(1965, dic)
#print sixties

seventies = search(1971, dic)
#print seventies

nineties = search(1990, dic)
#print nineties
```

```
twothousandfive = search(2005, dic)


# 1955's distribution

plt.hist(fiveties, bins = 50)
plt.axis([0, 15, 0 , 60])
plt.title(u"Distribucio de les
cancons de l'any %d (%d cancons)" % (1955, len(fiveties)))
plt.xlabel(u"Durades (minuts)")
plt.ylabel(u"Nombre de cancons")
#plt.show()
plt.savefig('1955.pdf')
plt.close()


# 1965 vs 1971 distribution

plt.hist(seventies, bins = 100)
plt.hist(sixties, bins = 100)
plt.axis([0, 15, 0 , 250])
plt.title(u"Distribucio de 1965 i 1971
(%d i %d cancons)" % (len(sixties), len(seventies)))
plt.xlabel(u"Durades (minuts)")
plt.ylabel(u"Nombre de cancons")
#plt.show()
plt.savefig('1965-71.pdf')
plt.close()


# 2005 distribution

plt.hist(twothousandfive, bins = 300)
plt.axis([0, 15, 0 , 2500])
plt.title(u"Distribucio de les cancons de l'any %d (%d cancons)"
% (2005, len(twothousandfive)))
plt.xlabel(u"Durades (minuts)")
plt.ylabel(u"Nombre de cancons")
```

```
#plt.show()
plt.savefig('2005.pdf')
plt.close()


# 1990 distribution

plt.hist(nineties, bins = 100)
plt.axis([0, 15, 0 , 3000])
plt.title(u"Distribucio de les cancons de l'any %d
(%d cancons)" % (1990, len(nineties)))
plt.xlabel(u"Durades (minuts)")
plt.ylabel(u"Nombre de cancons")
#plt.show()
plt.savefig('1990.pdf')
plt.close()




# RANGE #

#print min(durations)
#print max(durations)




#_____________________ TEST_____________________ #

#a_years
#averages

def test():
    a = 0
    test_years = []
    test_averages = []
```

```
    test_deviations = []
    for e in range(a_years[0], a_years[37]):
        test_years.append(e)
        test_averages.append(averages[a])
        test_deviations.append(deviations[a])
        a = a + 1
    return test_deviations, test_averages

td, ta = test()



plt.plot(ta, td, 'ro', marker='o', color='b')
plt.axis([2.5, 4, 0, 2.5])
plt.title(u"Test")
plt.xlabel(u"Mitjanes")
plt.ylabel(u"Desviacions")
#plt.show()
plt.savefig('test.pdf')
plt.close()
```