Introducción estadística
Seminario de Ciencia de Datos

Mirko Yves Bahoz Torrico, Andrés Genoud

September 6, 2017

1. Problema

Doce atletas entrenan para una competencia de $100mts$ llanos. El entrenamiento se realiza aún en condiciones climáticas adversas. El archivo tiempos.txt contiene los tiempos en segundos de cada atleta para un entrenamiento en un día soleado, en un día nublado y en un día de lluvia intensa.

2. Exploración Preliminar y Gráficos

Se tomaron los datos del archivo **tiempos.txt** y se realizo la gráfica siguiente:

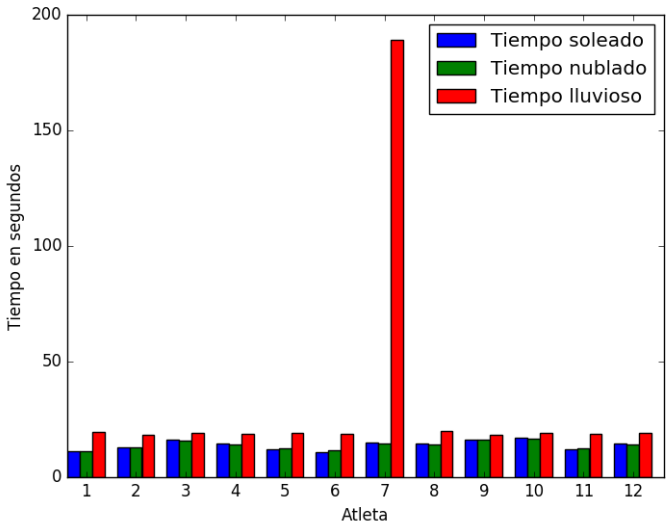

Figure 1: Tiempos de atletas

Como se puede observar hay un dato, en el atleta 7, que difiere ampliamente del resto de las medidas obtenidas. Al ver en detalle el archivo tiempos.txt se observó que el dato en cuestión tiene un valor de 189. Se llegó a la conclusión del que el valor no es un valor limpio y que se debe a un mal copiado de datos. Se decidió corregirlo y de esta manera tener un valor 18.9 (este valor se encuentra estrechamente cercano con el resto de los datos obtenidos).

Una vez corregido esto obtuvimos el siguiente gráfico:

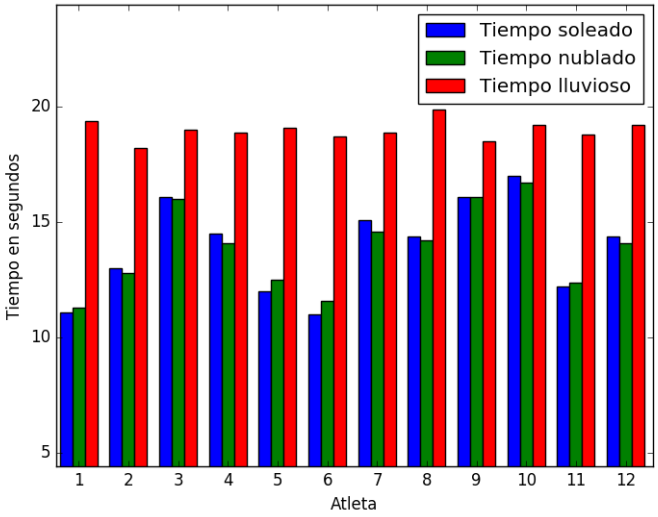

Figure 2: Tiempos de atletas

3. Análisis de datos

Un primer análisis de los datos se basó en obtener los tiempos promedios para los tiempos según el estado del día (soleado, nublado o lluvioso).

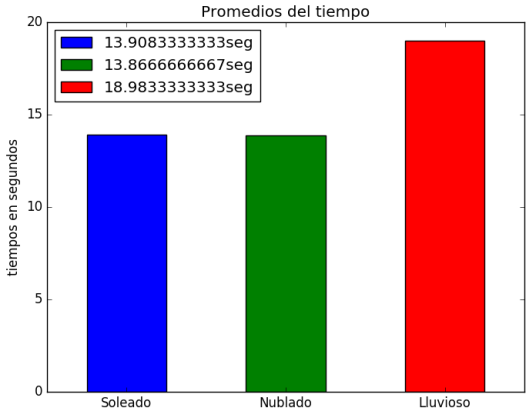

Figure 3: Tiempos promedios según el estado del día

Es claramente notorio que hay dos medias son casi identicas y una restante que difiere notablemente.

3.1 T-Tests de muestras independientes

Para verificar lo observado se realizó un t-test, con el supuesto de muestras independientes. En este test la **hipótesis nula** es:

Las medias de soleado y nublado son valores idénticos.

Obteniendo la siguiente información:

statistic= 0.053403167764020903, **pvalue**= 0.95789265561280046.

El p-valor $grande$ indica que no hay evidencia suficiente para rechazar la hipótesis nula. Osea que no tenemos información suficiente para decir que las medias sean distintas.

Una vez obtenido el resultado anterior queremos ver si existe evidencia suficiente para decir que la media del día soleado es diferente a la del día lluvioso. Para ello realizamos el mismo tests que el anterior. Luego nuestra hipótesis nula es que las medias del día soleado y lluvioso son iguales.

El resultado fue el siguiente:

statistic= -8.4828396056353661, **pvalue**= $2.2013598111131482e^{-08}$

El p-value $chico$ indica que la probabilidad de que las medias en cuestion sean iguales es extremanadamente baja. Entonces podemos rechazar la hipótesis nula.

Ambos resultados condicen con lo observado en la **Figura 2**.

3.2 T-Test de Muestras apareadas

Como podemos observar los datos que tenemos pertenecen a un conjunto de datos considerados apareados, ya que tenemos como constante a los atletas, y un cambio de condiciones (el clima, en el cual se realizaron las medidas). Entonces podemos realizar un t-test de muestras apareadas. La misma es más fuerte que el test de 2 muestras independientes.

Se realizaron nuevamente dos tests donde las hipótisis nulas fueron las mismas.

Hipótesis nula	soleado y nublado mismas medias	soleado y lluvioso mismas medias
statistic	0.41213824986058739	-8.576227572594302
p-valor	0.68815561156045579	$3.3516511607218045e^{-06}$

Table 1: Resultado. Tests de Muestras Apareadas.

Los datos obtenidos realizados por este test no contradicen las conclusiones obtenidas para el test de muestras independientes. Sin embargo se puede observar que los valores no son los mismos.

3.2.1 Test de Wilcoxon

El test de Wilcoxon es una prueba no paramétrica para comparar el rango medio de dos muestras relacionadas y determinar si existen diferencias entre ellas. En este caso la **Hipótesis nula** es: No existen diferencias signidicativas entre el tiempo de un corredor un día soleado y un día lluvioso.

Se obtuvo:

statistic= 0.0, **pvalue**= 0.0022177214642370492

El p-valor nos dice que rechazemos el H_0. Entonces podemos decir que existe una diferencia significativa entre el tiempo de un corredor un día soleado y un día lluvioso.

3.2.2 Test de Permutación

El concepto detrás del test de permutación es que los métodos paramétricos de inferencia estadística no son mas que aproximaciones al resultado que obtendríamos usando permutaciones de los mismos.

De esta forma nuestra hipótesis nula es que, **no hay** diferencia entre correr un día soleado y un día lluvioso.

Esto viene como resultado de suponer que las etiquetas son irrelevantes, y a que cualquier diferencia encontrada se debe al azar.

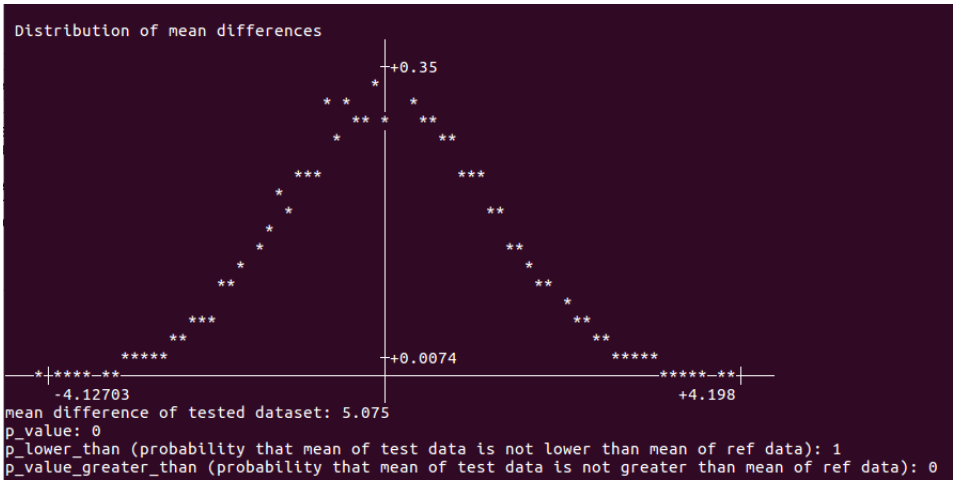

Figure 4: Resultado por consola de realizar el test de permutación soleado vs nublado

Con un p-valor=0, (el cero se debe a una error computacional, se deduce que es un valor muy pequeño) podemos rechazar la H_0. Y de esta forma decir que los días de lluvia los corredores son mas lentos.

4. Concluciones

Conclusiones estipuladas:

- **Los atletas son más lentos en días de lluvia que en días soleados.** Desde el primer gráfico se vio que la media de los tiempos correspondientes a un día lluvioso es mayor que los otros. También el statistic con signo negativo hacen ver lo mismo(en varios de los tests).
- **El cielo nublado no influye en los tiempos de los atletas.** Al igual que el hecho anterior no se encontro evidencia suficiente para refutar que los tiempos obtenidos en días soleados o nublados sean correspondientes a modelos distintos.
- **El clima influye en la velocidad de los atletas.** En particular este item quedó claro que sí el clima influye, se puede ver claramente en el análisis que se realizó con el test de permutación.

4.1 Consejos para el entrenador

Los atletas sostienen que no tiene sentido entrenar con lluvia, pero el entrenador asegura que es de gran utilidad. Dada la información dada no hay manera de afirmar o negar el interrogante. Para el mismo se propone al entrenador realizar otro tipo de experimentación. Por ejemplo, podria dividir a los corredores de forma aleatoria, y realizar el entrenamiento en días de lluvia con uno y el otro dejarlo como grupo de referencia.