DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 1

5 de septiembre de 2017 — Ciencia de datos

Integrante	LU	Correo electrónico
Christian Cuneo	755/13	chriscuneo93@gmail.com
Julián Bayardo	850/13	julian@bayardo.com.ar

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - Pabellón I
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Argentina
Tel/Fax: (54 11) 4576-3359
http://exactas.uba.ar

Índice

1. Exploración Preliminar

Comenzamos graficando los datos para ver bien con que estamos trabajando.
Primero un simple gráfico de lineas:

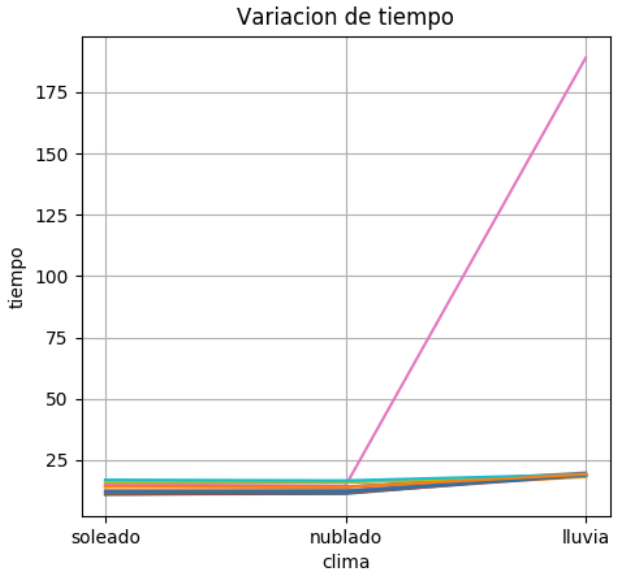

Figura 1: Vista general

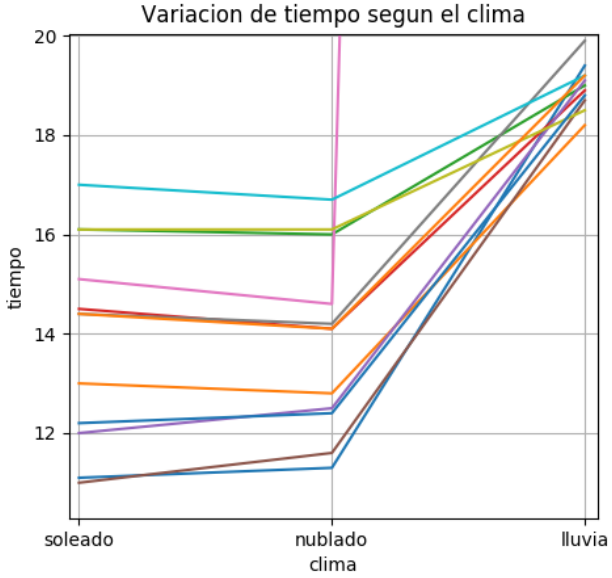

Figura 2: Zoom

Ya aquí podemos ver a simple vista que hay un dato que no tiene sentido, ya que estamos hablando de humanos corriendo, por lo tanto desde nuestro punto de vista no tiene sentido una diferencia tan amplia de tiempo.

A continuación procedimos a limpiar los datos, únicamente este tiempo en lluvia para el sujeto 7. La intuición inicial nos diria que directamente borremos toda esa fila, y demos como invalida las mediciones del sujeto 7, pero al tener tan pocos datos, decidimos solamente limpiarlo, suponiendo que lo que se olvidaron de poner es el punto decimal, trasnformando el 189 en un 18,9

A continuación veremos los datos limpios:

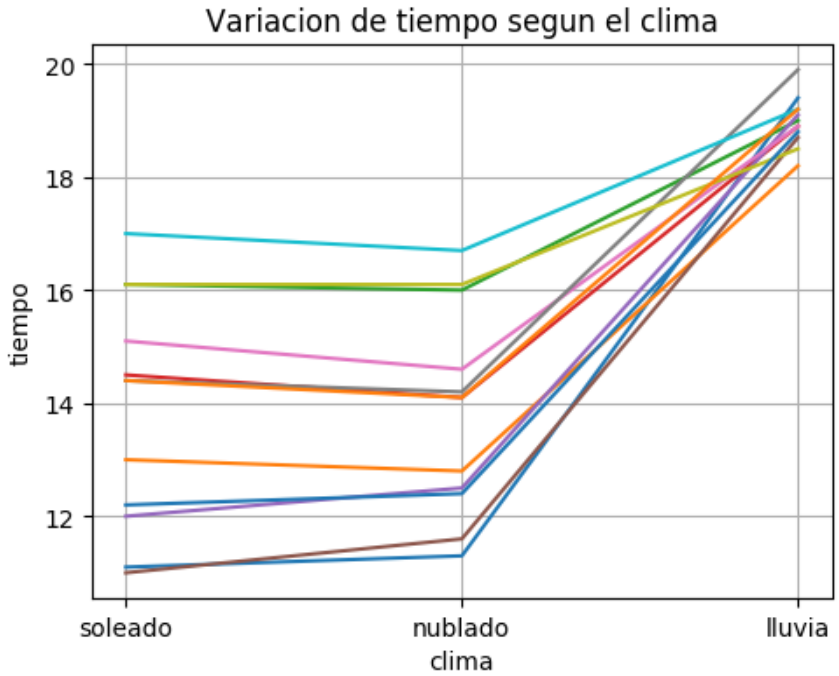

Figura 3: Datos limpios

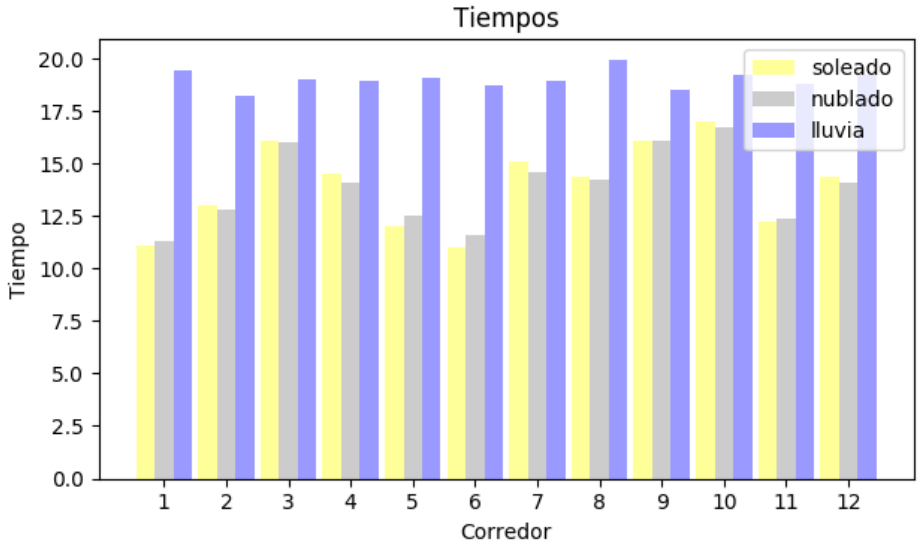

Figura 4: Grafico barras datos limpios

2. Tests

Ya con datos confiables procedemos a hacer pruebas estadísticas. Primero hicimos lo mas básico que fue plotear boxplots de los sampleos para cada clima:

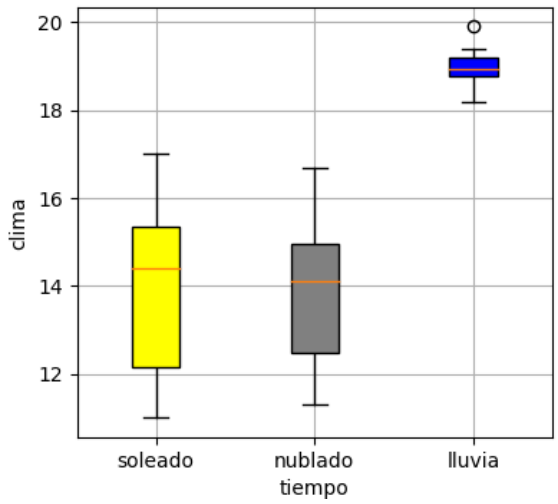

Figura 5: Media y desviación estándar según el clima

Ya podemos ver a simple vista que no hay diferencia perceptible entre la corrida con sol y la nublada, sino que la variación principal se puede ver al correr bajo lluvia. Acá podemos ver las diferencias entre las corridas, también en formato boxplot, la diferencia se calcula para cada sujeto:

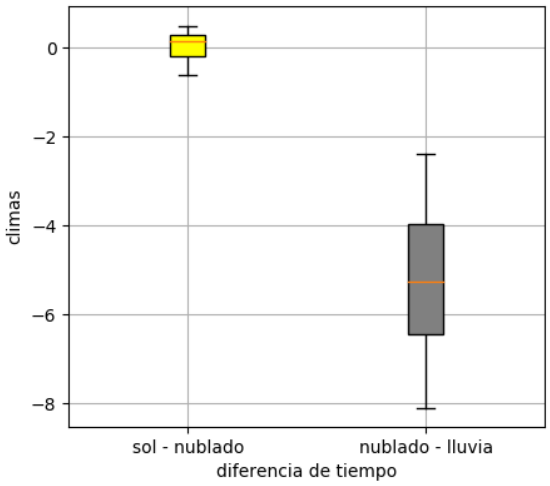

Figura 6: Media y desviación estándar de las diferencias

Vemos que la diferencia es insignificante entre el soleado y el nublado, pero no así entre el nublado y la lluvia (o el solado y la lluvia)

A continuación empezamos a realizar las pruebas, para este punto queremos utilizar tests de muestras apareadas, ya que lo que nos interesa es comparar entre las muestras en distintos climas. Si podemos utilizar el t-test de Students vamos a hacerlo, y si los datos no cumplen las hipótesis necesarias utilizaremos otro. En este caso a simple vista no parece que vaya a incumplir la hipótesis de la distribución Gaussiana pero si el de la varianza (mas que nada entre soleado vs lluvia y nublado vs lluvia) aunque por supuesto vamos a medir ambas variables. Si no comparten varianza utilizaremos el t-test de Welch.

T-test

Test sol vs nublado:
Distributions pass normality test
Distributions seem to have equal variance
Student t-test apareado: pvalue=0.6881556115604558
Averages scores are identical

Test nublado vs lluvia:
Distributions pass normality test
Distributions dont seem to have equal variance
Welch t-test independent: pvalue=2.408757741370807e-09
Averages scores differ significantly

Test sol vs lluvia:
Distributions pass normality test
Distributions dont seem to have equal variance
Welch t-test independent: pvalue=2.2013598111131482e-08
Averages scores differ significantly

El resultado muestra (con pvalues que indican resultados significativos) que podemos descartar la hipótesis nula (que diría que no hay una diferencia entre las medias de los grupos) para los pares nublado vs lluvia y soleado vs lluvia. Y muestra que para el caso de nublado vs sol no se puede rechazar esta hipotesis (que tiene sentido ya que notamos que no había diferencia en el gráfico preliminar)

Luego decidimos hacer un test no paramétrico, el signed rank de Wilcoxon (testea la hipotesis nula que ambos muestreos apariados vienen de la misma distribucion):

Wilcoxon

Sol vs Nublado: 28.5 con pvalue=0.6882065030738291
Nublado vs Lluvia: 0.0 con pvalue=0.002217721464237049
Sol vs Lluvia: 0.0 con pvalue=0.002217721464237049

Viendo de nuevo que la probabilidad de la hipotesis nula para los casos nublado vs lluvia y sol vs lluvia es rechazada, con un pvalue < 0.01.

Luego, realizamos un test de permutaciones, con 1000 permutaciones distintas, solo para la diferencia entre soleado y lluvia, ya que a esta altura ya vemos que la diferencia relevante se encuentra ahí y obtuvimos que la probabilidad de la hipótesis nula (que los tiempos no difieren entre soleado y lluvia) es 0.001:

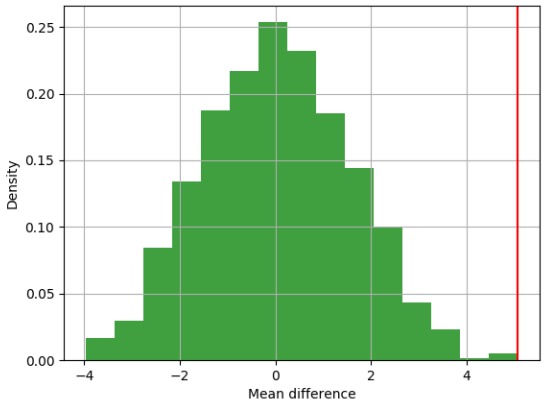

Figura 7: Histograma test de permutaciones, linea roja es la media inicial

Por ultimo calculamos el coeficiente de correlación de Pearson:

Pearson

Sol vs Nublado: 0.9908237767557433 con pvalue=5.045646399806607e-10
Nublado vs Lluvia: 0.04228906602480422 con pvalue=0.896177007833118
Sol vs Lluvia: 0.0525893206142193 con pvalue=0.8710570171791218

Aquí podemos ver que el único par que muestra correlación casi exacta es la del sol vs nublado, que vimos que no tenían una diferencia significativa, además el pvalue es muy chico, dándonos mucha confianza. Para los otros dos no da correlación significativa, y además, si existiese esa leve correlación, por el pvalue que dio, seria por azar.

3. Conclusiones

1. Si, y podemos ver este resultado con todos los test estadísticos que corrimos (no el de pearson, porque no aplicaría a este caso). En todos vimos que la hipótesis nula (que las diferencias no son significativas) fue rechazada. Mas específicamente en el de permutaciones, podemos ver que la probabilidad que ambas muestras vengan de distribuciones iguales es prácticamente nula.

2. Si tomamos como clima normal que este soleado, entonces podemos confirmar a partir de los datos que el cielo nublado no influye en los resultados, ya que mostramos con el coeficiente de correlación de Pearson que los muestreos con clima soleado y con clima nublado están exactamente correlacionados.

3. A simple vista parecería que no, ya que es consistente la velocidad en los otros climas, con la velocidad con lluvia, solo que se alentece para todos los corredores. El coeficiente de pearson dio una correlación positiva, por lo que se mostraría que se comportan de igual manera, pero el pvalue nos dice que puede no ser significativo, aunque con tan pocas muestras no es tan confiable. En el caso que esos tiempos no estén correlacionados podríamos concluir que, como en caso soleado los tiempos son mejores, y no están correlacionados con los malos tiempos en lluvia, no solo se afecta negativamente la performance, sino que no se los resultados en lluvia no reflejan el verdadero potencial de los deportistas

4. Esta conclusión no podemos sacarla de esta investigación ya que el clima puede ser muy variado y no analizamos todos los casos acá, ya que tenemos mediciones limitadas a 3 casos de clima. Con todos los t-tests pudimos ver que el tiempo en lluvia no viene de la misma distribución que con sol o nubes, pero no entre sol y nublado, por lo tanto la diferencia esta en ciertos casos, pero no están correlacionados los que difieren, por lo tanto sabemos que influye pero no de que forma.

4. Consejos

La primera impresión es que tienen razón los deportistas, ya que no logran llegar a su potencial en lluvia. Esto se ve con la falta de correlación entre los resultados de lluvia contra sol y nublado; y de la amplia diferencia en las medias y varianzas. Igualmente esta conclusión es muy débil por la calidad de los resultados de correlación. Para poder dirimir la cuestión también mediría la temperatura, y el tipo de suelo. Por sentido común el problema en velocidades puede venir por un tema de adherencia al piso y porque la temperatura corporal no llega a una temperatura de trabajo mas performante.