Estudio estadístico de dos muestras de datos independientes entre sí.

GRUPO 11

José Ignacio López Paz C.I:4491310-6
Germán Ruiz Raviales C.I:4317743-2
Federico Mujica Cazenave C.I:4786543-9

INDICE GENERAL

Estudio de Aleatoriedad

Primero veremos si las muestras son aleatorias realizando 2 tests. El test de Rachas de ascensos y descensos y el test de correlación de rangos de Spearman.
Para ambas muestras en estos tests la hipótesis nula Ho corresponde a que los datos son iid (independientes idénticamente distribuidos) y la hipótesis alternativa H1 a que no lo son.

Test de Rachas:
Realizando los respectivos cálculos para la primer muestra llegamos a que el total de rachas es igual a 64 y para la segunda muestra es 68. Como las muestras son de tamaño mayor a 25, entonces se usa la "aproximación normal" del TCL (teorema central del límite) para hallar el p-valor.
[1] El p-valor de la primer muestra es igual a 0.3304003 y el p-valor de la segunda muestra es igual a 0.3900298. Tomando el error de tipo1 igual al 10%(α=0.10), no se rechaza la hipótesis nula para ninguna de las muestras ya que sus respectivos p-valores son mayores que α.
Como las dos muestras aceptan Ho para este test podemos decir que los datos no tienen una tendencia sistemática (muy pocas rachas) y que no reflejan una periodicidad (demasiadas rachas).

[1] Para ver los cálculos realizados para los estadísticos y p-valores de los tests ver en los anexos el script de TestDeAleatoriedad

Test de Spearman:

El estadístico de Spearman es igual a:

$Rs= 1-[6(\sum_{1\leq i\leq n} (R(X_i)- i)^2)/(n(n^2-1))]$ siendo n la cantidad de datos en la muestra y $R(X_i)$ el rango del dato i.

Haciendo los cálculos llegamos a que el estadístico[2] de Spearman Rs1= -0.0240264 para la muestra1, y para la muestra2 Rs2=0.110003.

Usando la "aproximación normal" del TCL (teorema central del límite) llegamos a que el p-valor de la primer muestra nos dio 0.4055296 y el de la segunda dio 0.1368644. Para las dos muestras el p-valor es mayor que el error de tipo1 (α) por lo cual no se rechaza la hipótesis nula del test para ambas muestras.

Al haber aceptado Ho las dos muestras podemos decir que los datos no tienen ni una tendencia creciente ni decreciente.

Conclusión: Luego de realizar estos dos tests conjuntamente y como las dos muestras aceptaron llegamos a que los datos de ambas muestras se pueden considerar independientes idénticamente distribuidos.

[2] Para ver los cálculos realizados para los estadísticos y p-valores de los tests ver en los anexos el script de TestDeAleatoriedad

Tests de Ajuste

Primero realizamos graficas e histogramas para estimar cual distribución se adaptaba más a cada muestra.[3]
Luego de observar las graficas notamos que las dos podrían corresponder a una distribución exponencial, para verificar esta idea nos propusimos realizar algunos tests empezando con el test de Lilliefors de ajuste a exponenciales.
Para los tests presentados a continuación la hipótesis nula Ho corresponde a que los datos de la muestra tienen una distribución F=exp(λ)=Fo y la hipótesis alternativa que no.

Test de Lilliefors para exponenciales:
A partir de la muestra estimamos por el método de máxima verosimilitud (EMV) el parámetro λ de dicha distribución. El estadístico de esta prueba es igual a $Dn= \sup_{t\in R} |F_n(t) - F_0(t)|$ siendo la función de distribución empirica de la muestra
$F_n(t)= \#\{i: 1 \le i \le n: X_i \le t\}/n$ (n es la cantidad de datos de la muestra).
El estadístico[4] del test para la primer muestra nos dio Dn1 = 0.0585 y de la segunda nos dio Dn2 = 0.0395.
Como la cantidad de datos es mayor a 30 usamos la "aproximación normal" para hallar el valor critico

[3] Ver en anexos las graficas Histogramas_Densidades,EstimacionMuestra1 y EstimacionMuestra2
[4] Para los cálculos ver sript de TestDeAjuste de los anexos

asintótico que corresponde al error de tipo1 $\alpha=0.1$, que lo llamamos Cn. Llegamos a que Cn=0.096.
La región critica de la primer muestra es igual a R={Dn1>Cn} y para la segunda es R={Dn2>Cn}.
El estadístico de ambas muestras es menor a Cn por lo tanto no caen en la región entonces no se rechaza que los datos de las muestras tengan distribución exponencial.
A continuación realizamos el test de Kolmogorov-Smirnov para ambas muestras.

Test de Kolmogorov-Smirnov:

El estadístico y la región crítica de este test es el mismo que en el de Lilliefors. La diferencia es que el parámetro λ se estima con la mitad de la muestra y luego se usa la otra parte para aplicar el test. Decidir que parte de la muestra se usara para estimar los parámetros y que parte para aplicar el test, es una arbitrariedad. Para evitarla, lo hacemos en dos casos. En el primero aplicamos el procedimiento estimando los parámetros con la primera mitad de la muestra y aplicamos la prueba de ajuste con la segunda y en el segundo hacemos al revés, la segunda mitad para estimar y la primera para aplicar el test. De esta manera aseguramos más el rechazo o la aceptación de Ho.
Para el primer caso llegamos a que el estadístico para la primer muestra es [5]Dn1=0.1195 y para la segunda es Dn2=0.1684. En este caso el valor de

[5] Para los cálculos ver sript de TestDeAjuste en los anexos

Cn= 0.1725341. Dn1 es claramente menor que Cn por lo tanto no se rechaza Ho en este caso para la primer muestra. Dn2 también es menor que Cn pero por muy poca diferencia, por lo tanto hay que probar con el segundo caso planteado para corroborar que los datos de la segunda muestra no rechazan Ho.
En el segundo caso [6]Dn1=0.0777 que también es claramente menor que Cn por lo tanto para la primer muestra no se rechaza Ho. Luego Dn2=0.1413.En este caso dio bastante menor que en el anterior por lo tanto nos aseguro que la segunda muestra no rechaza Ho.

Conclusión: Luego de aplicar estos tests llegamos a la conclusión de que las dos muestras se pueden ajustar a distribuciones exponenciales ya que no se rechazo la hipótesis nula planteada en ninguno de ellos.

[6] Para los cálculos ver sript de TestDeAjuste en los anexos

Comparación de las distribuciones

En la parte anterior concluimos que los datos de las dos muestras tienen el mismo tipo de distribución por lo tanto en esta sección haremos el test de Kolmogorov-Smirnov (KS) para dos muestras para ver si tienen la misma distribución.
Para este test la hipótesis nula corresponde a que la distribución de la primer muestra F_1 es igual a la de la segunda F_2, la hipótesis alternativa es que son distintas.[7]

Test K-S para dos muestras:
El estadístico de este test es
$Dnm = \sup_{t \in R} |F_1(t) - F_2(t)|$ con n y m la cantidad de datos de la primer y segunda muestra respectivamente. Como en este caso las dos muestras son de igual tamaño lo llamaremos simplemente D.
Luego llegamos a que [8]D=0.22 y el Cn para α=0.1 es Cn= 0.5456006. La región critica de este test para rechazar Ho es R={nmD>Cn} (nmD el producto entre ellos y el m y n el tamaño de las muestras).Luego nmD=545.60 que es claramente mayor que Cn por lo tanto se rechaza Ho para las muestras.

[7] Ver en los anexos grafica de Funciones DeDistribución
[8] Ver la comparación de las muestras en los anexos en el script de Comparación para más detalles

Conclusión: Cómo se rechazo Ho concluimos con mucha seguridad que no tienen distribución idéntica.

Test Paramétricos de comparación

En esta sección del informe ya suponemos que las muestras tienen una distribución exponencial de parámetro $\lambda 1$ y $\lambda 2$ estimados por el método de máxima verosimilitud (EMV), que son iguales a 1/Xn (Xn=Promedio de los datos de la muestra1), y 1/Yn (Yn=Promedio de los datos de la muestra2) respectivamente.
Tenemos los datos de la muestra1 X =X1,. . .,Xn, y de la muestra2 Y=Y1,. . .,Yn que los suponemos iid y que tienen una distribución que no es la normal. La cantidad de datos es grande entonces si EXi = $\mu 1$, EYi = $\mu 2$, Var (Xi) = $\sigma 1^2 > 0$ y Var(Yi) = $\sigma 2^2$ vale el TCL entonces podemos calcular los intervalos de confianza para la media de dichas muestras.

Intervalos de confianza para la media:
El intervalo de confianza al 95% para $\mu 1$ es igual a [9]IC1 = [1.046226, 2.925395] y para $\mu 2$ es igual a IC2 = [1.437317, 4.629810].
A primera vista se puede observar que el IC1 es más chico que el IC2. El valor mínimo del IC1 es más chico del IC2 por poco pero en cambio el máximo es claramente menor al máximo del IC2. Para poder

[9] Ver en los anexos los cálculos de los intervalos en el script de TestParametricos

verificar que μ1 es efectivamente menor que μ2 aplicamos unos tests de comparación.

Test de comparación de las medias:

Tomamos la hipótesis nula como Ho: $\mu 1 = \mu 2$ y la hipótesis alternativa como H1: $\mu 1 \neq \mu 2$. La región critica del test es $R = \{|E| \geq Z\alpha/2\}$ con el estadístico $E = (Xn - Yn) / \sqrt{((Sn^2(X) + Sn^2(Y)) / n)}$.

Haciendo los respectivos cálculos llegamos a que [10] $E = -2.912908$ y para 5% ($\alpha=0.05$) $Z\alpha/2 = 1.96$.

Haciendo el valor absoluto de E resulta que se cumple la condición de la región critica por lo tanto se rechaza Ho para las muestras.

Luego dejamos Ho: $\mu 1 = \mu 2$ y cambiamos a H1: $\mu 1 \leq \mu 2$. El estadístico E es el mismo y la región critica en este caso es $R = \{E \leq -Z\alpha\}$.

Para $\alpha=0.05$ en este caso $Z\alpha = 1.65$ entonces nuevamente se cumple la condición de R por lo tanto se rechaza Ho.

Conclusión: Luego de aplicar estos dos tests pudimos confirmar lo que se había observado en los intervalos de confianza. Efectivamente la media de la primer muestra μ1 es menor que la de la segunda muestra μ2.

[10] Ver en anexos los cálculos reaalizados en el script de TestParametricos

Conclusión Final

Para este estudio los datos de las dos muestras corresponden al tiempo necesario para realizar la misma tarea en dos computadoras diferentes.
Luego de realizar todo este estudio estadístico pudimos concluir que los datos de ambas muestras podrían ser independientes idénticamente distribuidos para luego poder ajustar los datos a una distribución. Concluimos que ambas siguen una distribución exponencial de parámetro λ estimado por máxima verosimilitud. También concluimos que no tienen distribución idéntica. Los muestras se suponen independientes entre sí por lo tanto pudimos aplicar tests paramétricos para comparar las medias que serian la esperanza de cada dato. La media de los datos de la primer muestra resulto menor que el de la segunda.
Con todo esto en consideración pudimos concluir que los tiempos de ejecución del programa generados por la primer computadora tienden a ser menores que los generados por la segunda.
Como conclusión final nosotros consideraríamos que la primera computadora es más eficiente que la segunda realizando la tarea en cuestión.