udp
UNIVERSIDAD DIEGO PORTALES
facultad de ingeniería

Examen Probabilidades y estadística

Profesor: Demian Schkolnik

Ayudante: Nicolás Reyes

Diego Riffo y Felipe Ulloa

8-7-2020

udp | facultad de ingeniería
UNIVERSIDAD DIEGO PORTALES

Tabla de contenido

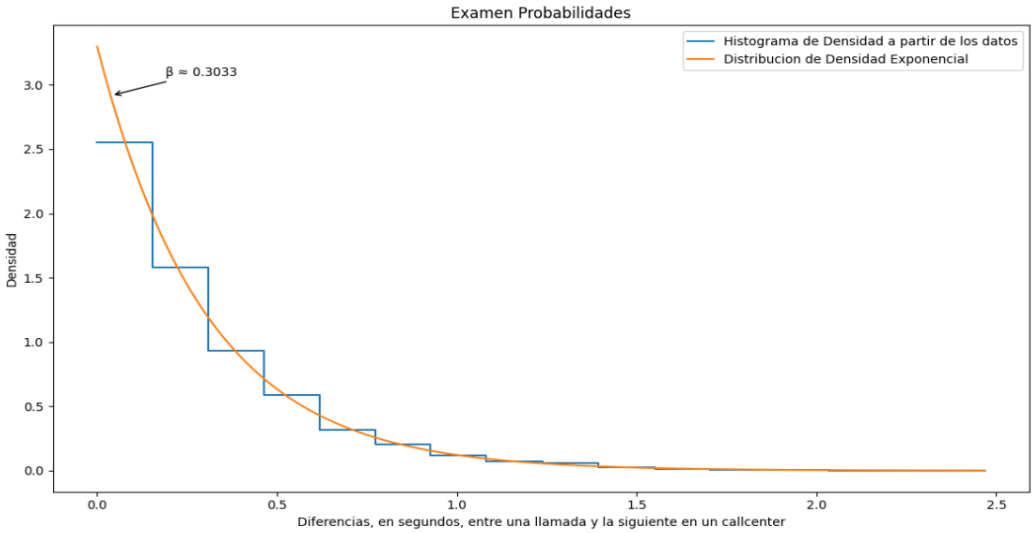
Introducción

Para el examen se no dio un problema el cual consiste en un call center en donde se nos entrega un set datos dándonos la diferencia en segundos en cuanto se tardan en recibir una llamada, una en el cual tenemos que saber la probabilidad que se tiene de que la siguiente llamada sea en menos de 0.5 segundos.

Para este caso lo más representativo a cómo se distribuyen los datos, fue la distribución exponencial, ya que dando un vistazo al histograma se puede observar que la mayoría de los datos de los intervalos entre llamadas de un call center, son mayores entre más cercano están del 0, por lo que de primeras necesitamos una distribución que tenga su mayor valor tendiendo al eje Y, y que vaya decreciendo, revisando las distribuciones que vimos en clase, la distribución que se asemejaban a este comportamiento fueron 2, la distribución exponencial y la distribución gamma, agregando por otro lado que necesitamos conocer cuál es la probabilidad de que la llamada se de en un determinado tiempo, podemos decantarnos soluciona con una distribución exponencial, ya que la distribución exponencial describe el tiempo hasta la ocurrencia de un evento y en cambio la función gamma es el tiempo (o espacio) que transcurre hasta que ocurre un número específico de eventos.

Para poder realizar una representación de cómo se distribuyen los datos en una distribución, tenemos que realizar un ajuste a los parámetros de esta distribución, que en este caso es la distribución exponencial, pero no basta con solo realizar el ajuste, también se debe realizar una comprobación en la confiabilidad de que el ajuste que realizamos a la distribución implica que nuestra distribución es una buena representación a la distribución de un set de datos.

Gráficos y análisis



Para poder comparar visualmente si la distribución que se propuso, con el parámetro correspondiente es representativa de los datos, se graficó el histograma de densidad de la variable aleatoria X, que en este caso representa las diferencias, en segundos, entre una llamada y la siguiente en un determinado call center.

En estadísticas, un histograma es una representación gráfica de una variable en forma de barras, donde el área de cada barra es proporcional a la frecuencia de los valores representados.

El número de barras, representa el número de intervalos en los cuales se divide la cantidad total de datos, los cuales para este caso fueron 16 intervalos, número propuesto por la fórmula de Sturges, que es una regla práctica acerca del número de clases que deben considerar al elaborar un histograma, para 10.000 datos nos arroja un número entre 14 y 15, y el histograma se está comparando con una distribución continua (Exponencial), por lo que entre más intervalos, más se acercará a la forma de la distribución, y para que la amplitud de los intervalos sea el mismo para todos los tramos, se aumentó el número obtenido a 16, por lo que se realizaron 16 intervalos, que dividirán los 10.000 datos.
Ahora, la distribución propuesta, que representa de mejor manera el comportamiento de los datos de nuestra variable aleatoria X, es la distribución exponencial, y la función de densidad, de la distribución exponencial tiene de fórmula:

$$e(x;\beta) = \begin{cases} \frac{1}{\beta} e^{\frac{-x}{\beta}} & x > 0 \\ 0 & e.o.c \end{cases}$$

, por lo que aparte del valor que tome x, la distribución densidad exponencial, solo depende del parámetro constante β, por lo que, para poder ajustar la función de densidad, de nuestra distribución exponencial, se tuvo que solo ajustar β

β se refiere al valor esperado que se espera obtener, para los valores de la variable aleatoria X, que en este caso, la variable aleatoria X representa las diferencias, en segundos entre una llamada y la siguiente de un determinado call center.

Por lo que para poder ajustar nuestra distribución exponencial, calculamos el valor esperado=Promedio, de todos los datos recolectados de la variable aleatoria X, leídos de un csv que corresponde al β e ingresarlo a la distribución exponencial, y así poder graficar la función de densidad de la distribución exponencial y compararla con el histograma de densidad creado a partir de la misma variable aleatoria X.

Test de Hipótesis

Para poder comparar la confiabilidad de nuestro ajuste realizado a la función de densidad de nuestra distribución exponencial, se realizó el test de hipótesis “Chi-cuadrado”, que consta en calcular un valor chi, resulta de comparar la distribución de las frecuencias de los datos observados (Los datos del CSV), con las frecuencias según la distribución teórica propuesta (Distribución exponencial),

Las frecuencias observadas, son cuantas veces se presenta el dato j, en el CSV, y la frecuencia esperada es la frecuencia en la cual se espera obtener a partir de la distribución exponencial, con ajuste realizado.

La frecuencia esperada se calcula de la forma:

Sea un intervalo [$y_{j-1} - y_j$)

$$Frecuencia\ Esperada = Cantidad\ de\ datos \cdot \int_{y_{j-1}}^{y_j} DensidadExponencial(x)\,dx = \left(1 - e^{-\frac{final}{media}} - \left(1 - e^{-\frac{inicio}{media}}\right)\right) \cdot 10.000$$

En el código sería:

```
55  def FrecExponencial(media, start, final):
56      return ((1-math.e **(-final/media))-(1-math.e **(-start/media)))*10000
```

El valor de chi se calcula con la siguiente fórmula:

$$X^2 = \sum_{Intervalos} \frac{(FrecuenciaObservada - FrecuenciaEsperada)^2}{Frecuencia\ Esperada}$$

Nuestra hipótesis nula, que será aceptada o rechazada en cierto intervalo, es de que la distribución exponencial con ajuste de media ≈ 0.3033 es o no, una buena representación de cómo se distribuyen los datos de nuestra variable aleatoria X.

El intervalo en el cual se juzgará será para ciertos valores de α que es la probabilidad de equivocarse rechazando una hipótesis correcta.

Para poder hacer el uso de la tabla de distribución de chi cuadrado, que tiene como parámetros el un valor de chi y un al α, usaremos p, para poder hacer afirmaciones sobre nuestro valor obtenido de Chi.

P = Probabilidad de encontrar un valor mayor o igual que el chi cuadrado
Si el nivel de significación del test es α,
Valor p < α ⇒ se rechaza la hipótesis nula.
Valor p > α ⇒ no se rechaza la hipótesis nula.

```
felipe@felipe-HP-ENVY-Laptop-13-ah0xxx:~/Escritorio/Examen Proba/Examen_proba$ python3 TestChiCuadrado.py
Intervalo                          |        Frecuencia Obtenida                  |      Frecuencia esperada  |     X^2(Chi Cuadrado) Del intervalo

2e-06 0.154436                          3943                        3990.558312794088                        0.5667861333009003
0.154436 0.308866                       2437                           2398.044611276574                        0.632816547055577
0.308866 0.463258                       1439                           1440.831659123457                        0.0023284990465745345
0.463258 0.617632                       910                            865.8980852361608                        2.2461983910109633
0.617632 0.77235                        485                           521.3460927158409                        2.5338992162132348
0.77235 0.926817                        311                           312.6158690423997                        0.00835220799949736 4
0.926817 1.081785                       184                            188.31023583745997                        0.09865705330303963
1.081785 1.237339                       108                            113.29130881739813                        0.24713236428579047
1.237339 1.393017                       92                            67.87184805532421                        8.577454908710035
1.393017 1.549873                       42                            40.854776869979624                        0.03210239090786453
1.549873 1.704253                       22                            24.061259892601548                        0.1765822888665186
1.704253 1.860173                       12                            14.572350284695057                        0.4540781588348436
1.860173 2.034764                       5                           9.488388694414152                        2.1231880059893182
2.034764 2.201871                       4                           5.164029357628275                        0.2623850972920881
2.201871 2.390121                       3                           3.2489451429129623                        0.019075017106749432
2.390121 2.469023                       3                           0.865111579817501 3                        5.268393896184813
Resultado de chi cuadrado es T :  23.249430176107786
```

En la tabla de valores para Chi cuadrado, con 16-1=15 intervalos de confianza, tenemos un valor de chi, con α=0,05 de 24.9958, mayor que el valor obtenido de chi 23.249, por lo que no se rechaza la hipótesis al nivel α = 0.05, lo que implica que en ese α, no da razón de rechazar la hipótesis propuesta, pero nuestro chi, si es menor que en el α=0,1 y esto indica que en el intervalo de α=0,1 si se rechaza nuestra hipótesis .

En estadística, que nuestra hipótesis sea aceptada en el intervalo α=0,05, quiere decir que las diferencias entre las proporciones observadas entre los grupos (Hipótesis propuesta verdadera e hipótesis propuesta falsa) , son estadísticamente significativas, esto quiere decir que hicimos una representación de cómo se distribuyen los datos a una distribución estudiada(exponencial), estadísticamente representativa a cómo se distribuyen las diferencias, en segundos entre una llamada y la siguiente de un determinado call center

Solución de problema

El problema se solucionará ocupando el modelo de probabilidad que se comentó anteriormente, el cual es el la distribución exponencial.
con el que lo calcularemos usando la fórmula de la función de distribución la cual es la siguiente:

$$F(x) = P(X \leq x) = \int_0^x e(t;\beta)dt = \frac{1}{\beta}\int_0^x e^{-t/\beta}dt = 1 - e^{-x/\beta}.$$

donde β es la media de nuestros datos y x es la cantidad de segundos que queremos que se demora la cual el problema que se nos dio es de 0.5 segundos.
La media de todos los datos es de 0.303 segundos

$$P(X<0.5) = 1 - e^{-\frac{0.5}{0.303}}$$

la cual nos da como resultado 0.807 de probabilidad de que en la siguiente llamada sea en menos de 0.5 segundos.

Código

Para la realización del código se ocupó Python con lo que se hizo el gráfico que podemos encontrar en “Gráfico y análisis”, para poder ocupar el data set entregado, se leyó el data set y se introdujo cada dato en un espacio de un arreglo convirtiendo el String en un Float. Para corroborar el número de datos del data set, introdujo un contador.

```
datosCsv = []
contador = 0
x=0
with open('dataSet9.csv', newline='') as File:
    reader = csv.reader(File)
    for row in reader:
        contador=contador+1
        x=float(row[0])+x
        datosCsv.append(float(row[0]))
```

Después se tiene que crear el histogramas que se hace con los datos recogidos del data set, histogramas se dividirá en 16 intervalos en los cuales todos tendrán el mismo rango.

```
AmplitudIntervalo=(max(datosCsvOrdenados)-min(datosCsvOrdenados))/16

IndiceCsv=0
IndiceCsv2=0

for i in range(16):
    inicio=IndiceCsv
    FrecIntervalo=0
    while(datosCsvOrdenados[IndiceCsv]<(datosCsvOrdenados[inicio]+AmplitudIntervalo)):
        IndiceCsv+=1
        FrecIntervalo+=1
        if(IndiceCsv==10000):
            IndiceCsv-=1
            break


    lista1.append(datosCsvOrdenados[inicio])
    lista2.append(((FrecIntervalo/10000))/AmplitudIntervalo)
    lista1.append(datosCsvOrdenados[IndiceCsv])
    lista2.append(((FrecIntervalo/10000))/AmplitudIntervalo)

plt.plot(lista1, lista2, label="Histograma de Densidad a partir de los datos")
```

Luego se hará la gráfica de la distribución de densidad exponencial ocupando la media de el data set, con el que se tiene que ingresar en la fórmula de densidad exponencial vista anteriormente.

```
def exp(datoX, media):
    exp = (1/media)* math.e **(-datoX/media)
    return exp
```

Hay que unir el histograma con la curva exponencial en un solo gráfico para poder hacer las comparaciones y analizar si la distribución propuesta es correcta, la cual se hizo de la siguiente manera.

```
expX=[]
expY=[]
#Llenar las listas con los valores de al distribucion binomial
datosCsvOrd=sorted(datosCsv)
for i in range(contador):
    expX.append(datosCsvOrd[i])
    expY.append(exp(datosCsvOrd[i], media))


plt.plot(expX,expY, label="Distribucion de Densidad Exponencial")
plt.annotate("β ≈ "+str(round(media, 4)),
             xy=(0.04, exp(0.04,0.3)),
            xytext=(0.04+0.15, exp(0.04,0.3)+0.15), textcoords='data',
            arrowprops=dict(arrowstyle="->",
                            connectionstyle="arc3")
            )
plt.xlabel('Diferencias, en segundos, entre una llamada y la siguiente en un callcenter')
plt.ylabel('Densidad')
plt.legend(loc='upper right')
plt.title('Examen Probabilidades')
plt.show()
```

Conclusión

Podemos concluir que representar un set de datos, en una distribución conocida nos otorga facilidad al realizar análisis y generar hipótesis sobre el comportamiento de nuestra variable aleatoria a analizar.

Por otro lado, podemos concluir que no solo basta con generar un ajuste a una distribución, sino que, también hay que comprobar que la distribución con el ajuste realizado, es representativa a cómo se distribuyen nuestro set de datos, en este informe , la función de distribución de la distribución exponencial, es lo suficientemente representativa para que con esta resolver el problema, esto lo podemos observar en el gráfico en donde comparando las gráficas del histograma y la función de densidad de la distribución exponencial y también se puede comprobar en el test de hipótesis realizado.