Universidad
de La Laguna

Distribución de Poisson

Alba Tomé Rodríguez
Dácil Batista García
Desireé Praena Pacheco

Grupo (3D)

Técnicas Experimentales. 1^{er} *curso.* 2^{do} *semestre*

Lenguajes y Sistemas Informáticos

Facultad de Matemáticas

Universidad de La Laguna

La Laguna, 11 de mayo de 2014

Índice general

Índice de figuras

Índice de cuadros

Resumen

El presente informe recoge información sobre la distribución de Poisson. En él se muestran definiciones fundamentales para la comprensión de ésta,así como ejemplos, aplicaciones y el diseño de un experimento en el cual se aplica dicha distribución.

De ella podemos decir que es una de las distribuciones más importantes de variable discreta (como la conocida binomial) pues los valores que puede tomar la variable aleatoria son números naturales. Es muy útil cuando la muestra o segmento 'n' es grande y la probabilidad de éxitos 'p' es pequeña.

La distribución de Poisson se utiliza en situaciones donde los sucesos son impredecibles o de ocurrencia aleatoria. En otras palabras no se sabe el total de posibles resultados.

Capítulo 1

Motivación y objetivos

En la vida cotidiana aparecen distintas situaciones que conviene analizar para realizar informes, estudios, predicciones, etc. Ejemplos de ellas podrían ser el 'Número de hijos por pareja de un país' o el 'Número de piezas defectuosas en un lote de 100 bombillas'. Dada la naturaleza de estas variables y considerando la definiciones que se muestran a lo largo del informe, se puede admitir que seguirán una distribución de Poisson.
Por esto el objetivo principal de este trabajo va a ser la implementación en Python de una distribución de Poisson, para así poder encontrar solución a este tipo de problemas.

1.1. Ejemplo 1. Distribución de los aminoácidos en las proteínas

Antes del descubrimiento del código genético aceptado actualmente, se habían propuesto ciertos códigos basados, entre otras, en consideraciones de dependencia entre los aminoácidos de una proteína. (Código Triángulo y código Diamante) Una experiencia de Gamow parece indicar que no se dan tales relaciones, al menos para las proteínas estudiadas. Se compilaron los dipéptidos que podían formarse tomando aminoácidos vecinos, en una muestra de proteínas con unas proporciones de aminoácidos bastante equilibradas, descartando algunas con una composición altamente especializada. Con la muestra utilizada podían formarse 177 dipéptidos, de este tipo. Si se consideran 20 aminoácidos distintos, pueden formarse $20 * 20 = 400$ combinaciones diferentes de dos aminoácidos, con lo que, por término medio se observarán:

$$\frac{177}{400} = 0{,}442$$

dipéptidos de un tipo.

Los 400 pares posibles pueden considerarse como "casilla", a las que un dipéptido "pertenece" si está formado por aquellos dos aminoácidos, pero se repite la experiencia un número de veces $N = 177$, grande (número de dipéptidos realmente observados), la variable aleatoria x= "número de dipéptidos en una casilla", seguirá una distribución de Poisson

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!} \text{ (x = 0,1,2,3,...,n,...)}$$

siendo $\lambda = \frac{177}{400} = 0{,}442$, de manera que el número esperado de casillas con x dipéptidos será $Nf(x)$.

Los resultados se resumen en la tabla anexa (véase el cuadro 1.1). Como se puede apreciar en ella, la distribución de dipéptidos observada se ajusta en gran medida a la esperada según la distribución de Poisson, cosa que no ocurre con la distribución esperada según los códigos "Diamante" y "Triangulo".

N	F	CT	CD	D. Poisson
0	264	276	305	264.2
1	103	86	55	116.0
2	27	26	23	25.6
3	4	9	7	3.77
4	2	3	3	0.42
5	0	0	3	0.037
6	0	0	2	0.0027
7	0	0	0	$1{,}7x10^{-4}$
8	0	0	2	$9{,}5x10^{-6}$

Cuadro 1.1: Datos

1.2. Ejemplo 2. Carnavales

La probabilidad de que en carnavales una persona se desmaye es de 0.001. Considerando que acuden unas 5000 personas, ¿cuál es la probabilidad de que se desmayen 25 personas?

Solución:

Se trata de una binomial con $n = 5000$ y $p = 0{,}001$. La probabilidad sería:

$$F(x = 25) = \binom{5000}{25} 0{,}001^{25}(1 - 0{,}001)^{5000-25}$$

que resulta complejo de calcular. Por eso se prefiere aproximar a una distribución de Poisson, con $\lambda = 5000 * 0{,}001 = 50$, y quedaría:

$$F(x = 25) = (e^{-50})\frac{50^{25}}{25!} = 3{,}70 * 10^{-5}$$

1.3. Ejemplo 3. Cheques sin fondo

Si un banco recibe en promedio 6 cheques sin fondo por día, ¿cuáles son las probabilidades de que reciba:

1. cuatro cheques sin fondo en un día dado?
2. diez cheques sin fondo en cualquiera de dos días consecutivos?

Solución:

1. x = variable que nos define el número de cheques sin fondo que llegan al banco en un día cualquiera=0,1,2,3,.... $\lambda = 6$ cheques sin fondo por día.

$$p(x = 4, \lambda = 6) = \frac{6^4 2{,}718^{-6}}{4!} = \frac{(1296)(0{,}00248)}{24} = 0{,}13392$$

2. x = variable que nos define el número de cheques sin fondo que llegan al banco en dos días consecutivos=0,1,2,3,.... $\lambda = 6 * 2 = 12$ cheques sin fondo en promedio que llegan al banco en dos días consecutivos.

$$p(x = 10, \lambda = 12) = \frac{12^{10} 2718^{-12}}{10!} = \frac{(6{,}1917364 * 10^{10})(0{,}000006151)}{3628800} = 0{,}104953$$

Capítulo 2

Fundamentos teóricos

2.1. Historia de la Distribución de Poisson

La distribución fue introducida por primera vez por Simeón Denis Poisson, físico y matemático francés al que se le conoce por sus diferentes trabajos en el campo de la electricidad y que también hizo publicaciones sobre la geometría diferencial y la teoría de probabilidades, en la que nos centraremos. La distribución fue publicada, junto con su teoría de la probabilidad, en 1837 en su obra "Recherches sur la probabilit des jugements en matire criminelle et en matire civile", un trabajo importante en el cual describe la probabilidad como un acontecimiento fortuito ocurrido en un tiempo o intervalo bajo las condiciones siguintes: la probabilidad de que un acontecimiento ocurra es muy pequeña, pero el número de intentos es muy grande. Entonces el evento ocurre algunas veces. El resultado había sido dado previamente por Abraham de Moivre en "De Mensura Sortis seu, de Probabilitate Eventuum en Ludis un Casu fortuito Pendentibus" en Philosophical Transactions de la Royal Society, p. 219.

Una aplicación práctica de esta distribución fue hecha por Ladislao Bortkiewicz en 1898 cuando se le dio la tarea de investigar el número de soldados en el ejército prusiano matados accidentalmente por tiro de caballos. Este experimento introdujo la distribución de Poisson para el campo de la ingeniería de confiabilidad. (Veáse la figura 2.1)

2.2. Conceptos fundamentales

Para entender la distribución de Poisson hemos de tener en cuenta conceptos fundamentales como por ejemplo:

- **Variable aleatoria:** En palabras simples, podemos decir que, dado un experimento aleatorio y asociado al mismo un espacio probabilístico, una variable

Figura 2.1: Simeón Denis Poisson

aleatoria es una aplicación que a cada valor del espacio muestral le hace corresponder un número real. Se clasifican en:

- **Discretas:** si los números asociados a los sucesos son puntos aislados. Por ejemplo:
  "Lanzar una moneda 3 veces y que salga cara". Los posibles resultados son (0,1,2,3).
- **Continuas:** los valores asignados pueden ser cualesquiera dentro de ciertos intervalos. Por ejemplo:
  "Tomando la variable nivel de agua de un embalse", pueden obtenerse valores entre 0 y ∞.

2.2.1. Distribución de Poisson

Sea X una variable aleatoria de una distribución discreta definida sobre un espacio de probabilidad, se dice que X tiene una distribución de Poisson de parámetro λsi su función de densidad (es decir, función que describe el comportamiento probabilística de la variable) es:

$$f(x) = \begin{cases} P[X = x] = e^{-\lambda}\frac{\lambda^x}{x!} & , \text{si } x \in \mathbb{N} \\ 0 & , \text{en otro caso} \end{cases}$$

donde λ es un parámetro característico de la distribución. A dicha distribución se denomina $P(\lambda)$.

2.2.2. Aproximación de la binomial a la Poisson

La Poisson se puede obtener de la binomial en determinadas condiciones. Sea X una variable aleatoria con distribución binomial $B(n, p)$, cuya función de densidad es:

$$f(X) = \binom{n}{k} p^k (1-p)^{n-k}$$

Cuando el número de pruebas $n \to \infty$ y la probabilidad del suceso tiende a cero y $np \to \lambda$ entonces:

$$\lim_{n \to +\infty p \to 0 np \to \lambda} f(x) = e^{-\lambda} \frac{\lambda^k}{k!}$$

que es la función de distribución de Poisson, es decir, $B(n,p) \to P(\lambda)$ bajo las condiciones anteriores.

2.3. Propiedades

1. **Esperanza**

   La media o esperanza matemática de $P(\lambda)$ es:

   $$E(X) = \lambda$$

2. **Varianza**

   Respecto a la varianza,

   $$Var(X) = \lambda = E(X)$$

3. **El parámetro λ**

   El parámetro λ de una distribución de Poisson caracteriza a la misma:

   $$\lambda = \lim_{n \to +\infty p \to 0} np$$

   Se puede obtener de varias formas:

   - n y p conocidas
     $$\lambda = np$$
   - A partir de la esperanza
     $$\lambda = E(X)$$
   - Estimando a partir de una muestra
     $$\lambda = m(X_1, \ldots, X_s) = \frac{1}{s} \sum_{i=1}^{s} X_i$$
   - A partir de la probabilidad del suceso $[X = 0]$ ya que

$$P[X = 0] = f(0) = e^{-\lambda}\frac{\lambda^0}{0!} = e^{-\lambda}$$

luego $\lambda = -\ln P[X = 0]$

4. **Función característica**

   La función característica viene dada por:

$$\varphi = e^{\lambda(e^{it}-1)}$$

2.4. Aplicaciones

- Estimación del número de diferencias génicas entre dos especies a partir de datos sobre la identidad electroforética de sus proteínas.
- Liberación de cuantos de acetilcolina en las terminaciones nerviosas. (Sigue $P(\lambda)$)
- Consideraciones sobre las hipótesis explicativas del proceso de extinción de 10 órdenes de reptiles en el Cretácico.
- Distribución de los aminoácidos en las proteínas.
- Contaje del número de glóbulos rojos en una muestra de sangre.

Capítulo 3

Procedimiento experimental

El experimento realizado consiste en aproximar la distribución de los aminoácidos en las proteínas mediante la Poisson. (Ejemplo 1 del primer capítulo) Además, una vez obtenidos los datos, se podrán comparar con experimentos realizados antes de la aparición del modelo de Poisson y con los datos reales observados.

3.1. Descripción de los experimentos

Teniendo en cuenta la descripción dada en el Ejemplo 1, Capítulo 1 (Motivación y objetivos), se ha diseñado un modelo preparado para aplicar la distribución de Poisson a un número entero 'n', que será introducido por el usuario, y que representa el número de dipéptidos (unión de dos aminoácidos, estructuras principales de las proteínas). Además, la función de distribución consta de un parámetro 'lambda' que es constante pero que hay que calcular en base a los datos dados del experimento.
En este caso, se toma una muestra con la que podrían formarse 177 dipéptidos de un total de 400 combinaciones diferentes (por ser dipéptidos -2- y haber 20 proteínas posibles: 20x20=400), por tanto, nuestro parámetro será la cantidad constante $177/400 = 0.442$.
Una vez introducido el número 'n', el algoritmo calcula la probabilidad de que exista una casilla con 'n' dipéptidos, entendiendo por casilla el total (400). Posteriormente, muestra una tabla en la que aparecen, desde 0 hasta 'n' el número de dipéptidos, su probabilidad según la distribución y la frecuencia total (400x'n').

El algoritmo es el siguiente:

```
#!/usr/bin/python
#!encoding: UTF-8
```

```
import sys
import math

def calcular_poi(n):
  lam = 0.4425
  f_i=1 #factorial
  for i in range(n+1):
    if i==0:
      f_i=f_i
    else:
      f_i=f_i*i;
  ex = math.exp(-lam)
  ln = lam**n
  v1 = ex * ln
  valor_poi = v1 / f_i
  return (valor_poi)

#programa principal

argumentos = sys.argv[1:]

if (len(argumentos) == 1 ):
  n = int (argumentos[0]);
else:
  print 'Introduzca el número de  dipéptidos(0<n<10):'
  n = int (raw_input());

if (n>0):

  lista = [] #para indicar que es una variable de tipo lista
  for i in range (n+1):

    valor = calcular_poi(i)
    lista.append (valor) #para añadir valores a la lista

  print "numero de dipeptidos por casilla\tdistribucion de Poisson (f(x))\t400xf(x)

  for i in range (n+1):
    print "%d\t\t\t\t\t%1.10f\t\t\t%1.10f" % (i, lista[i],400*lista[i])


else:
```

```
print 'el número de dipeptidos debe ser mayor que 0'
```

Este estudio se ha hecho para el número de dipéptidos 'n' = 8 , con lo que se solicita al usuario que el número esté en el intervalo [0,10]. No obstante, el algoritmo es válido para números naturales mayores.

En el Apéndice 1 se especifica el nombre del archivo de extensión Python que debe ejecutarse para realizar el experimento.

3.2. Descripción del material

Para la implementación del algoritmo, como ya se ha mencionado en varias ocasiones, se ha utilizado Python.

Las características de la máquina desde la que se ha realizado el informe son las siguientes:

('default', 'Feb 27 2014 20:00:17')
Linux-3.2.0-61-generic-i686-with-Ubuntu-12.04-precise
('Linux', 'PROA', '3.2.0-61-generic', '93-Ubuntu SMP Fri May 2 21:33:33 UTC 2014', 'i686', 'i686')
2.7.3
Intel(R) Celeron(R) M CPU 520 @ 1.60GHz
GenuineIntel
1595.908 Hz
1024 KB

Además, se ha estudiado el tiempo que tarda la máquina en ejecutar el algoritmo principal en función del dato introducido. Ver Figura 3.1

El Apéndice 1 incluye también los nombres de los archivos con los algoritmos implementados para obtener los datos de la máquina y los tiempos de ejecución.

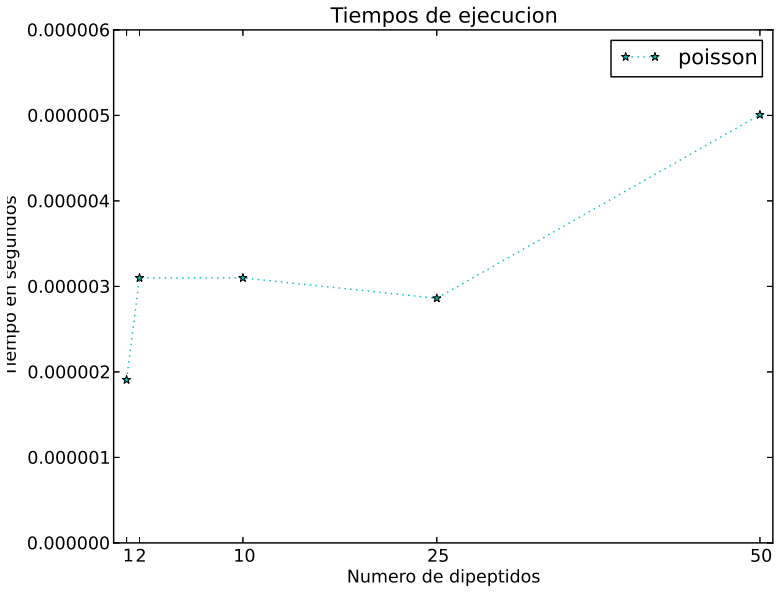

Figura 3.1: Tiempos de ejecucion del algoritmo

3.3. Resultados obtenidos

Las siguientes tablas y gráficos representan los resultados obtenidos en el experimento y además una comparación con resultados reales observados y experimentos anteriores a la aplicación de la distribución de Poisson. Véase el Cuadro 3.1 y la Figura 3.2

3.4. Análisis de los resultados

Haciendo un análisis detallado de los datos que muestran las tablas, puede observarse que la distribución de Poisson es la que describe en mejor medida el experimento.
Es claro que al comparar las gráficas, la función que representa la distribución de Poisson se acerca mucho más a los datos reales observados.
Sin embargo, los experimentos anteriores (Codigo Diamante y Codigo Triangulo) son algo menos acertados.
Se infiere de la observación que a medida que aumenta el número de dipéptidos, es menos probable encontrar casillas que los contengan, es decir, existen muchas más casillas con una cantidad pequeña de dipéptidos que casillas con grandes cantidades.
De hecho, se ha elegido como 'n' = 8 porque a partir de esta cantidad, la probabilidad de que las casillas esté vacías es casi total.

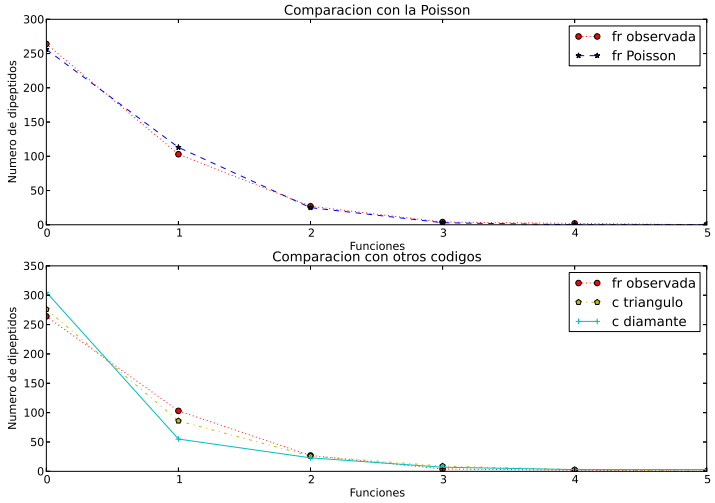

Figura 3.2: Comparacion Poisson-C Diamante-C Triangulo

N dipeptidos	Frec. observada	P(x) (Poisson)	Frecuencia Total 400xP(x)
0	264	0.6424283410	256.9713363872
1	102	0.2842745409	113.7098163514
2	27	0.0628957422	25.1582968677
3	4	0.0092771220	3.7108487880
4	2	0.0010262816	0.4105126472
5	0	0.0000908259	0.0363303693
6	0	0.0000066984	0.0026793647
7	0	0.000004234	0.0001693741
8	0	0.000000234	0.0000093685

N dipeptidos	Frec. observada	Codigo Triangulo	Codigo Diamante
0	264	276	305
1	102	86	55
2	27	26	23
3	4	9	7
4	2	3	3
5	0	0	3
6	0	0	2
7	0	0	0
8	0	0	2

Cuadro 3.1: Comparacion

Este hecho podría ser útil en diferentes campos de la biología o la medicina, entre otros.

Capítulo 4

Conclusiones

Como conclusion principal puede decirse que implementando un algoritmo que representa el problema que se quiere resolver, se ahorra tiempmo al obtener los resultados y ademas estos pueden ser representados en graficas y tablas para realizar comparaciones con otros experimentos
Ademas, el algoritmo no es valido unicamente para el ejemplo en el que se centra el informe, ya que al describir la misma funcion, simplemente cambiando lso datos de entrada puede adaptarse a gran variedad de experimentos con las mismas caracteristicas, con lo cual abarca un campo de estudio mucho mas amplio.
Respecto a los propios datos, se llega a la conclusion de que una distribucion de Poisson descrie mucho mejor el eperimento que algunas de las tecnicas mas antiguas llevadas a cabo como pueden ser los mencionados 'Codigo Diamante' y 'Codigo Triangulo' , ya que los datos obtenidos se aproximan con mas exactitud a los reales observados. Este hecho asegura cierta confianza en posibles futuras aplicaciones del experimento en diferentes campos.

Apéndice A

Archivos adjuntos

A.1. Algoritmos utilizados

A continuacion se muestran los nombres de los ficheros adjuntos a este informe que han sido utilizados para implementar las funciones y crear las tablas y graficos.

1. poisson.py : Algoritmo en Python en el que se ha desarrollado la funcion principal
2. informacion.py : Algoritmo en Python que proporciona la informacion de la maquina
3. tiempos.py : Algoritmo en Python que calcula el tiempo que tarda la maquina en ejecutar Poisson.py.
4. graf tiempo.py : Algoritmo en python que genera la grafica del tiempo que de tiempos.py
5. tablas.tex : Archivo LaTeX que genera las tablas de datos de comparacion de la Poisson

Bibliografía

[1] Fundamentos de Probabilidad en Estadistica. G.Alonso. J.Ocaña. C.M.Cuadras.

[2] Estadistica Teorica y Aplicada. Andres Nortes Checa.

[3] http://www.itch.edu.mx/academic/industrial/sabaticorita/private/05Distr

[4] http://www.aulafacil.com/CursoEstadistica/Lecc-29-est.htm

[1]

[1]Facultad de Matemáticas. Técnicas Eperimentales. Universidad de La Laguna