



UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICA

Estudio Estadístico de las Frecuencias de Conexiones para decidir si las Redes Biológicas son Libres de Escala

Trabajo Especial de Grado presentado ante la ilustre Universidad Central de Venezuela por la **Br. Wendy M. Montero R.** para optar al título de Licenciado en Matemática.

Tutor: Dra. María Margarita Olivares.

Caracas, Venezuela
Noviembre 2008

Nosotros, los abajo firmantes, designados por la Universidad Central de Venezuela como integrantes del Jurado Examinador del Trabajo Especial de Grado titulado “**Estudio Estadístico de las Frecuencias de Conexiones para decidir si las Redes Biológicas son Libre de Escala**”, presentado por la **Br. Wendy Melissa Montero Ramirez**, titular de la Cédula de Identidad **17855761**, certificamos que este trabajo cumple con los requisitos exigidos por nuestra Magna Casa de Estudios para optar al título de **Licenciado en Matemática**.

Dra. Maria Margarita Olivares
Tutor

Dr. Ricardo Rios.
Jurado

Dra. Glaysar Castro
Jurado

Dedicatoria

A mi mamá Flor de Montero por su apoyo desde que comencé este trabajo y por ser mi amiga y estar conmigo en los días buenos y no tan buenos.

A mi papá Guido Montero, hermanos Wilfredo Montero, Willy Montero y mi bella sobrina Greshell Montero, por confiar en mí y apoyarme siempre.

A Koki, mi eterno compañero de todas las noches de estudio. Espero estés en el cielo feliz de que termináramos este trabajo.

Agradecimiento

A Dios por darme la salud y el regalo de despertar cada día.

A la prof. María Margarita Olivares y prof. Glaysar Castro por confiar en mi el desarrollo de este trabajo.

Al prof. Juan M. Guevara por su valiosa ayuda y colaboración en el desarrollo de este trabajo y por su apoyo incondicional sin importar la hora y el día.

Al prof. Ricardo Rios por su orientación y apoyo en la elaboración de este trabajo.

A mi familia por el estímulo constante y por el apoyo que me brindaron para alcanzar esta meta. Gracias por su paciencia en las largas noches con la luz encendida.

Índice general

Índice de figuras	VII
Índice de cuadros	VIII
Introducción	1
Capítulo 1. Modelo Matemático	2
1. Introducción	2
2. Método de Máxima Verosimilitud	2
3. Prueba Estadística	4
4. Prueba Chi-Cuadrado	5
Capítulo 2. Sistemas Biológicos	8
1. Introducción	8
2. Redes Complejas	8
3. Redes Biológicas	11
Capítulo 3. Métodos Computacionales	14
1. Introducción	14
2. Método de Newton-Raphson	14
3. Análisis de Redes	16
Capítulo 4. Análisis	21
1. Introducción	21
2. Representación de la red	21
3. Método de Máxima verosimilitud para cada Red	27
4. Resultados	37
5. Conclusión	40
Bibliografía	41

Apéndice	43
1. Apéndice A	43
2. Apéndice B	54

Índice de figuras

2.1.Red	12
2.2.Interacción	12
3.1.Método de Newton	15
4.1.Recta ajustada para la red Rain	22
4.2.Recta ajustada para la red Guelzim	22
4.3.Recta ajustada para la red Uetz	23
4.4.Recta ajustada para la red Li-data	23
4.5.Recta ajustada para la red Lee	24
4.6.Recta ajustada para la red Ito	24
4.7.Datos ajustados para la red Rain utilizando la Distribución Exponencial	31
4.8.Datos ajustados para la red Guelzim utilizando la Distribución Exponencial	31
4.9.Datos ajustados para la red Uetz utilizando la Distribución Exponencial	32
4.10.Datos ajustados para la red Li-data utilizando la Distribución Exponencial	32
4.11.Datos ajustados para la red Lee utilizando la Distribución Exponencial	33
4.12.Datos ajustados para la red Ito utilizando la Distribución Exponencial	33

Índice de cuadros

1. Estimadores de la distribución de ley de potencia.	37
2. Resultados de la Prueba χ^2 para cada red.	37
3. Estimadores de la Distribución de Ley de Potencia Exponencial Truncada.	38
4. Resultados de la Prueba χ^2 con la Distribución Exponencial Truncada.	39
5. Frecuencias de Rain. Tamaño de la red $n = 732$.	44
6. Frecuencias de Ito. Tamaño de la red $n = 3278$.	46
7. Frecuencias de Guelzim. Tamaño de la red $n = 478$.	47
8. Frecuencias de Uetz. Tamaño de la red $n = 1044$.	48
9. Frecuencias de li-data. Tamaño de la red $n = 3228$.	50
10. Frecuencias de Lee. Tamaño de la red $n = 2398$.	53

Introducción

El estudio matemático de las redes surge de manera natural para representar la interacción entre diferentes elementos de un sistema. Las redes se estudian de manera sistemática en la llamada teoría de grafos. En los últimos años los estudios de sistemas muy complejos como la internet, mecanismos biológicos y la propagación de epidemias, por citar algunos ejemplos, han demostrado que los grafos que surgen de estos sistemas no corresponden a los tradicionalmente estudiados y es por ello que se han buscado nuevos métodos de estudio e instrumentos de trabajo para el análisis de modernas redes complejas. Así introducen la noción de red libre de escala, un tipo específico de red compleja compuesta de nodos y enlaces que tiene la particularidad de que los enlaces están distribuidos de forma no uniforme donde algunos nodos están altamente conectados a diferencia de otros nodos que poseen pocos enlaces, de este tipo de red surge de la necesidad de utilizar métodos probabilísticos para describir o analizar problemas en los cuales las propiedades de la red se conocen de forma incompleta.

El concepto de redes libres de escala es un poderoso paradigma unificador en el estudio de redes biológicas, sus parámetros son independientes del tamaño del sistema, además son robustas a rupturas aleatorias de enlaces. Se cree que conocer la estructura o topología de las redes biológicas ayudará a entender el diseño y la dinámica del proceso evolutivo que generaron estas redes. La distribución de grados y las frecuencias de conexiones es fundamental para las conclusiones biológicas. Es por ello que en el siguiente trabajo se realiza un análisis estadístico para probar si las redes biológicas son libres de escala basado en el artículo *How Scale-Free Are Biological Networks* realizado por Raya Khanin and Ernst Wit. De igual forma se estudian las propiedades de estas redes.

CAPÍTULO 1

Modelo Matemático

1. Introducción

En este capítulo se presenta el Método de Máxima Verosimilitud el cual es una de las herramientas estadísticas más utilizadas para calcular estimadores de cierta distribución. De la misma forma se explica en qué consiste una prueba estadística. Seguidamente se describe la prueba *Chi – Cuadrado*.

2. Método de Máxima Verosimilitud

Antes de introducir el Método de Máxima Verosimilitud es importante tener claro que un estimador es una función de la muestra que sirve para obtener un estimado de los parámetros de una población en base a una muestra de la misma.

El método de máxima verosimilitud es un método sistemático y una de las herramientas de estimación más utilizadas que permite obtener estimadores puntuales de los parámetros desconocidos de una distribución teórica.

Función de Verosimilitud.

Se define la *función de verosimilitud* L de una muestra observada a la densidad conjunta (o función de probabilidad conjunta en el caso discreto) de la muestra aleatoria $\vec{X} = (X_1, X_2, \dots, X_N)$ considerada como función del parámetro o de los parámetros desconocidos de la distribución evaluada en las observaciones $x_1, x_2, \dots, x_N$ del vector aleatorio $\vec{X}$.

Sea $\vec{\beta} = (\beta_1, \beta_2, \dots, \beta_k)$ el vector de parámetros desconocidos de la distribución entonces la función de verosimilitud viene dada por:

$$L(X; \vec{\beta}) = f_{x_1, x_2, \dots, x_N}(x_1, x_2, \dots, x_N; \vec{\beta})$$

donde $f_{x_1, x_2, \dots, x_N}$ es la densidad conjunta de $X_1, X_2, \dots, X_N$, en el caso de densidad y en el caso discreto viene dada por:

$$L(X; \vec{\beta}) = p_{x_1, x_2, \dots, x_N}(x_1, x_2, \dots, x_N; \vec{\beta})$$

donde $p_{x_1, x_2, \dots, x_N}$ es la densidad conjunta de $X_1, X_2, \dots, X_N$

La función de verosimilitud L es una función del vector $\vec{\beta}$ de parámetros ya que las observaciones o valores muestrales se conocen. Cuando hay un solo parámetro desconocido, L es una función de una variable.

La función de verosimilitud representa en cierto sentido, la probabilidad de observar lo que realmente se observó.

El método consiste en elegir los parámetros $\beta_j, j = 1, 2, \dots, k$ de manera que la probabilidad de observar lo que realmente se observó sea máxima, es decir, se desea elegir los parámetros β_j de forma tal que maximicen la función de verosimilitud.

Sea $X_1, X_2, \dots, X_N$ una muestra de N variables aleatorias independientes de la variable aleatoria X entonces:

1. En el caso de densidad, si $f(x; \vec{\beta})$ es la densidad de X y $x_1, x_2, \dots, x_N$ representa la muestra observada, se tendrá que:

$$L(x_1, x_2, \dots, x_N; \vec{\beta}) = f(x_1; \vec{\beta}) \cdot f(x_2; \vec{\beta}) \cdot \dots \cdot f(x_N; \vec{\beta})$$

2. En el caso discreto, si $g(x; \vec{\beta}) = \mathbb{P}(X = x)$ es la función de probabilidad de X , y $x_1, x_2, \dots, x_N$ representa la muestra observada, $\xi_1, \xi_2, \dots, \xi_r$ con $1 \leq r \leq N$ son los valores distintos observados y $f_1, f_2, \dots, f_r$ son las frecuencias respectivas, con $\sum_{i=1}^r f_i = N$ se tendrá que:

$$L(x_1, x_2, \dots, x_N; \vec{\beta}) = (g(\xi_1; \vec{\beta}))^{f_1} \cdot (g(\xi_2; \vec{\beta}))^{f_2} \cdot \dots \cdot (g(\xi_r; \vec{\beta}))^{f_r}$$

Supongamos construida la función de verosimilitud para $\vec{\beta}$, $L(x_1, x_2, \dots, x_N; \vec{\beta})$. Un procedimiento intuitivo de estimación es escoger aquel valor que haga máxima la probabilidad de aparición de los valores muestrales efectivamente observados; en otros términos, seleccionar

como estimador del parámetro el valor que maximice la probabilidad de lo efectivamente ocurrido.

Esto conduce a obtener el valor máximo de la función

$$L(x_1, x_2, \dots, x_N; \vec{\beta})$$

Suponiendo que esta función es diferenciable y que su máximo no ocurre en un extremo de su campo de definición, el máximo se obtendrá resolviendo

$$\begin{aligned} \frac{\partial L(\vec{\beta})}{\partial(\beta_1)} &= 0 \\ &\dots\dots\dots \\ \frac{\partial L(\vec{\beta})}{\partial(\beta_k)} &= 0 \end{aligned}$$

Como el logaritmo es una función estrictamente creciente, se tiene que tanto la función de verosimilitud como su logaritmo alcanzan el máximo en el mismo punto, por esta razón en la práctica los estimadores de máxima verosimilitud se obtienen derivando el logaritmo de la función de verosimilitud.

$$l(x_1, x_2, \dots, x_N; \vec{\beta}) = \ln L(x_1, x_2, \dots, x_N; \vec{\beta})$$

3. Prueba Estadística

En una prueba estadística están presentes cuatro elementos:

1. La hipótesis nula
2. La hipótesis alternativa
3. El estadístico de prueba
4. La región de rechazo

La Hipótesis Nula

Se indica simbólicamente por H_0 y establece la hipótesis sometida a prueba. H_0 especifica valores para uno ó más parámetros de la población.

La Hipótesis Alternativa

Se indica simbólicamente con H_1 la cual debe reflejar el valor posible ó intervalo de valores del parámetro de interés, si la hipótesis nula es falsa. Así, la hipótesis alternativa representa una forma de negación de la hipótesis nula.

El Estadístico de prueba

La decisión de rechazar o aceptar la hipótesis nula se basa en la información contenida en una muestra extraída de la población de interés. Los valores de la muestra se usan para calcular un solo número, este número actúa como ente que toma decisiones y lo llamaremos estadístico de prueba

Región de Rechazo

El conjunto de valores o rango del estadístico de prueba se divide en dos conjuntos ó regiones. Si el estadístico de prueba calculado a partir de una muestra específica asume un valor dentro de la región de rechazo, entonces la hipótesis nula es rechazada y decidimos a favor de la alternativa. Si este valor cae fuera de la región de rechazo (cae en la región de aceptación), no se rechaza H_0 lo cual no debe interpretarse como equivalente a que H_0 deba ser aceptada sino que simplemente, la información experimental que se tiene no conduce a rechazarla.

4. Prueba Chi-Cuadrado

La Prueba *Chi – Cuadrado* es el contraste de ajuste más antiguo, cuya idea es comparar las frecuencias observadas a partir de unas observaciones con las frecuencias especificadas por el modelo teórico que se contrasta. Esta prueba es válida para todo tipo de distribuciones, discretas y continuas.

La hipótesis H_0 indica que unos datos de una variable aleatoria X provienen de cierta distribución hipotética. Existen dos variantes posibles. En la primera H_0 especifica completamente la distribución, en la segunda H_0 especifica la forma, pero no los parámetros que se estiman a partir de los datos.

Caso Discreto:

Supongamos que X es discreta y se realizan n observaciones ($n \geq 25$) del experimento en investigación. Sea $\xi_1, \xi_2, \dots, \xi_k$ con ($k \leq n$) el número de observaciones distintas de la variable X , $f_1, f_2, \dots, f_k$ son las frecuencias correspondientes, es decir f_i es el número de observaciones iguales a ξ_i . (Eventualmente $f_i = 0$ para algún i)

Sabemos que $f_1 + f_2 + \dots + f_k = n$

Sea $p_i = \mathbb{P}(X = \xi_i), i = 1, 2, \dots, k; \sum_{i=1}^k p_i = 1$.

Si n está fijo, f_i es el número de veces que aparece ξ_i en n repeticiones del experimento y p_i representa la probabilidad de obtener ξ_i , luego, f_i tiene distribución binomial de parámetros (n, p_i) , donde $\mathbb{E}(f_i) = np_i$ es la frecuencia esperada por el modelo teórico

Ahora se calcula la discrepancia entre las frecuencias observadas y las previstas por el modelo mediante:

$$(1.1) \quad \chi^2 = \sum_{i=1}^k \frac{(f_i - \mathbb{E}(f_i))^2}{\mathbb{E}(f_i)}$$

donde

- f_i es la frecuencia observada de la clase i
- $\mathbb{E}(f_i)$ es la frecuencia esperada de la clase i según la distribución hipotética

La diferencia $f_i - \mathbb{E}(f_i)$ mide la desviación entre las frecuencias observadas y las frecuencias esperadas.

Luego (1.1) se distribuye aproximadamente como una χ^2 y sus grados de libertad son:

- Si el modelo especifica completamente las probabilidades p_i que son conocidas antes de tomar la muestra, el número de grados de libertad será $k - 1$
- Si las probabilidades p_i se han calculado estimando r parámetros del modelo por el método de máxima verosimilitud, el número de grados de libertad es $k - 1 - r$

Toma de decisión

Se quiere que f_i esté cercano a $\mathbb{E}(f_i)$, es decir que el valor observado

$$\chi_{obs}^2 = \sum_{i=1}^k \frac{(f_i - \mathbb{E}(f_i))^2}{\mathbb{E}(f_i)}$$

esté cercano a cero, procedemos a calcular el p -valor el cual se obtiene de la siguiente manera:

$$p = \mathbb{P}(\chi^2 > \chi_{obs}^2)$$

Si el p -valor es pequeño, rechazamos la hipótesis nula. Por ejemplo si $p < 0,05$ estamos rechazando con un nivel de significación menor que $\alpha = 0,05$.

Caso Continuo:

Este caso es una generalización del caso discreto. Sean $x_1, x_2, \dots, x_n$ n observaciones de la variable aleatoria X , las cuales tabulamos en una tabla de frecuencias, si k es el número de intervalos de clase, I_i es el i -ésimo intervalo tal que:

$$\sum_{i=1}^k \mathbb{P}(I_i) = 1$$

f_i el número de observaciones que caen en I_i , denotando por $p_i = \mathbb{P}(I_i)$ la probabilidad teórica, el estadístico de prueba será:

$$\chi_{obs}^2 = \sum_{i=1}^k \frac{(f_i - \mathbb{E}(f_i))^2}{\mathbb{E}(f_i)}$$

donde

- f_i es la frecuencia observada del intervalo I_i
- $\mathbb{E}(f_i) = np_i$ es la frecuencia esperada del intervalo I_i según la distribución hipotética.

Al igual que en el caso discreto, los grados de libertad de la distribución dependerá de si se estiman o no parámetros de la distribución teórica.

CAPÍTULO 2

Sistemas Biológicos

1. Introducción

El estudio matemático de las redes surge de manera natural para representar la interacción entre diferentes elementos de un sistema. Las redes se estudian de manera sistemática en la llamada *Teoría de Grafo*. Un grafo es un conjunto de nodos conectados mediante líneas que llamaremos ramas o conectores. Sin embargo, en los últimos años los estudios de sistemas muy complejos tales como la Internet, mecanismos biológicos y la propagación de epidemias, por citar algunos ejemplos, ha demostrado que los grafos que surgen de estos sistemas no corresponden a los tradicionalmente estudiados.

2. Redes Complejas

En la década de los cincuenta del siglo XX aparecen los trabajos de los matemáticos Paul Erdős y Alfred Renyi quienes introdujeron la noción de grafos aleatorios. Este tipo de grafos surgen de la necesidad de utilizar métodos probabilísticos para describir o analizar problemas en los cuales las propiedades de sus grafos se conocen de forma incompleta y solamente se puede pedir la probabilidad de que una cierta propiedad se satisfaga. Al igual que otros objetos matemáticos, los grafos aleatorios habían sido empleados por los científicos tales como Gibbs y Boltzmann en el siglo *XIX* aunque su estudio matemático se inició con Erdős y Renyi. Se dice que un grafo es aleatorio o de Erdős-Renyi si cada rama del grafo existe con una probabilidad p y p sigue una distribución uniforme.

Posteriormente con el desarrollo de la computación ha sido posible estudiar y analizar redes muy complejas en distintos contextos de la ciencia y la tecnología, resultando que los grafos aleatorios cuya distribución del grado de un nodo obedece una ley de potencias son los que surgen más comúnmente en el análisis de problemas reales. Dichos grafos corresponden

a redes libres de escalas y fueron descubiertas por el físico A.L. Barabási.

Ley de Potencia Una función de ley de potencia o función potencial es una expresión de la forma $f(x) = \alpha \cdot x^\gamma$ donde α y γ son los parámetros de dicha función, α es llamado el parámetro de amplitud y γ el exponente.

Medidas Asociadas a las Redes

- *Grado*: Se define el grado de un nodo como el número de ramas o conectores asociadas al nodo. En algunos casos se puede distinguir entre ramas que entran y ramas que salen de un nodo (denominadas redes dirigidas). En el caso de tener una red dirigida se define el grado de entrada de un nodo y el grado de salida de un nodo como el número de ramas que entran y el número de ramas que salen del nodo respectivamente.
- *Distribución de Grado*: La distribución de grado de un nodo es la la distribución de probabilidad del número de conexiones de un nodo, la caracterizamos por $\mathbb{P}(k)$ que representa la probabilidad de que un nodo tenga k conexiones.
- *Grado del Exponente*: Se refiere al valor γ de la distribución de grado de un nodo que satisface una ley de potencias. Dicho exponente es importante porque permite inferir la existencia de nodos principales llamados "hubs" que caracterizan a las redes libres de escala.
- *Distancia mínima o geodésica*: La distancia entre dos nodos a y b se denota por d_{ab} , indica el menor número de ramas o conectores necesarios para unir los dos nodos. En otras palabras indica la longitud del camino más corto entre a y b .

Clasificación de las Redes

En el estudio de fenómenos complejos las redes asociadas evolucionan en el tiempo y raras veces se posee información completa sobre cada uno de los nodos o enlaces que conforman. Estas redes varían fácilmente y por lo general puede ocurrir que nuevos nodos aparezcan o que desaparezcan y de forma similar sucede con los enlaces. Es por ello que a partir de su estructura y comportamiento estas redes se pueden clasificar en tres grandes grupos que son:

Redes Aleatorias: Es el modelo más básico de redes complejas. Los parámetros son el número N de nodos y la probabilidad p de que exista una conexión entre dos nodos. La probabilidad $\mathbb{P}(k)$ de que un nodo tenga k conexiones o tenga k grados corresponde a una distribución binomial. El grado esperado o promedio del número de conexiones es $\langle k \rangle = p(N-1)$. Si p es fijo y N crece indefinidamente ($N \rightarrow \infty$), el promedio $\langle k \rangle$ diverge. Si p es función de N de manera que si N crece indefinidamente $\langle k \rangle$ se mantiene constante ($N \rightarrow \infty, p \rightarrow 0$), la Distribución Binomial se aproxima a la Distribución de Poisson de parámetro $\langle k \rangle$.

Redes Libres de Escala: En este tipo de red algunos nodos están altamente conectados, es decir, poseen un gran número de enlaces a otros nodos ("Hubs"), aunque el grado de conexión de la mayoría de los nodos es bastante bajo.

En las redes libres de escala la probabilidad $\mathbb{P}(k)$ de que un nodo de la red esté conectado con k nodos es proporcional a $k^{-\gamma}$, es decir, sigue una ley de potencias.

- Si $\gamma > 2$ existe el primer momento de la distribución ya que $\sum_{k=1}^{\infty} \frac{1}{k^{\gamma-1}} < +\infty$ si y solo si $\gamma > 2$.
- Si $\gamma > 3$ existe el segundo momento de la distribución es decir, existe la varianza ya que:

$$\bullet E[x^2] = \frac{1}{\xi(\gamma)} \sum_{k=1}^{\infty} \frac{1}{k^{\gamma-2}} < +\infty \text{ si y solo si } \gamma > 3$$

- Por tanto $Var[x]$ nos queda de la siguiente forma:

$$\begin{aligned} Var[x] &= E[x^2] - (E[x])^2 \\ &= \frac{1}{\xi(\gamma)} \sum_{k=1}^{\infty} \frac{1}{k^{\gamma-2}} - \left(\frac{1}{\xi(\gamma)} \sum_{k=1}^{\infty} \frac{1}{k^{\gamma-1}} \right)^2 \end{aligned}$$

- Donde $Var[x] < +\infty$ si y solo si $\gamma > 3$.

Los exponentes γ estimados de muchas redes reales se encuentra entre el rango $2 < \gamma \leq 3$ y la distribución viene dada por:

$$\mathbb{P}(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}, k = 1, 2, \dots,$$

donde $\zeta(\gamma)$ es la función Zeta de Riemann definida por $\zeta(\gamma) = \sum_{k=1}^{\infty} \frac{1}{k^{\gamma}}$ la cual normaliza la distribución, es decir, se cumple que $\sum_{k=1}^{\infty} \mathbb{P}(k) = 1$. En las redes libres de escala no existen nodos huérfanos, es decir, $\mathbb{P}(0) = 0$. Estos modelos se denominan libres de escala pues el cociente $\frac{\mathbb{P}(\alpha k)}{\mathbb{P}(k)}$ depende solamente de α

Redes de Mundo Pequeño: Una red es de mundo pequeño si todo nodo se conecta con cualquier otro a través de un número pequeño de enlaces. Es decir, la mayor parte de pares de nodos en la mayor parte de la red parecen estar unidos (conectados) por un camino corto en la red.

3. Redes Biológicas

A finales de los años noventa, los trabajos de Barabási caracterizaron en términos de mecánica estadística un gran número de grafos que representaban redes reales de diversa naturaleza; estos estudios marcan el inicio de la Teoría de Redes como línea importante de investigación convirtiéndose en una herramienta de análisis poderosa a la biología. El estudio de estas redes es de gran importancia para la biología ya que permite en parte, predecir comportamientos. El patrón más conocido es *Libre de Escala* puesto que su estructura es independiente del tamaño del sistema (i.e, invarianza de escala). Esta topología o estructura de la red se caracteriza por una distribución heterogénea de enlaces, donde pocos nodos están muy conectados y muchos están poco conectados.

Red de interacción de proteína: En esta red las proteínas están conectadas por interacciones físicas.

Red de Metabolismo: Productos metabólicos, sustratos que participan en una reacción.

Red Genética Reguladora: En esta red dos genes están conectados si el nivel de un gen modula la expresión del otro por la activación o represión.

Redes Biológicas como un Grafo

- *Nodo* Representa una Proteína, Gen ó Metabolismo
- *Enlace (Edge)* representa una asociación, es decir, una interacción.

Red

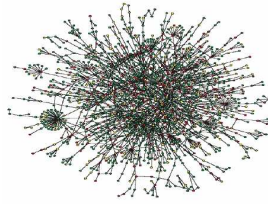


FIGURA 2.1. Red

- *Enlace Dirigido* Muestra la modulación (regulación) de un nodo a otro. Ejemplo: Una flecha de un gen X a un gen Y significa que el gen X afecta la expresión del gen Y .

Interacciones biológicas son estudiadas en diferentes niveles independientes. Las flechas sólidas indican interacciones directas. Flechas punteadas indican las interacciones indirectas que ocurren vía intermediarios.

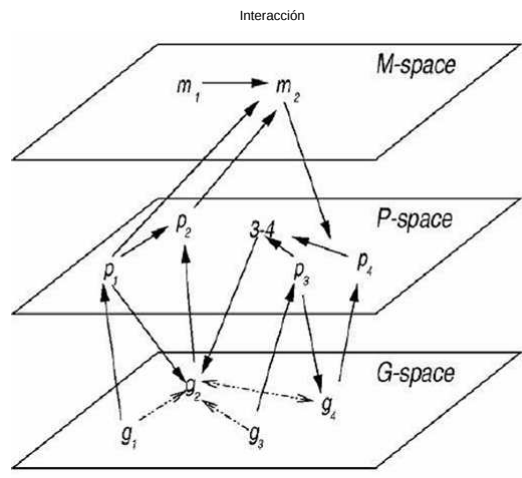


FIGURA 2.2. Interacción

Propiedades en común de las Redes Biológicas

- *Propiedad de Mundo pequeño* (Small-World) Existe un camino corto de cualquier nodo a otro nodo.
- *Presencia de Hubs* Muchos nodos con pocas conexiones y pocos nodos con muchas conexiones.
- *Hubs son enriquecidos por nodos esenciales/mortales*

Nodos Mortales: Son aquellos que cuando mutan el organismo muere.

Nodos Esenciales: Son los que se necesitan para una función específica y si mutan se puede perder la función.

CAPÍTULO 3

Métodos Computacionales

1. Introducción

En éste capítulo se presentan los métodos computacionales empleados en este trabajo así como también algunas características para el análisis de redes. Por último se estudiará la propiedad de autosimilaridad.

2. Método de Newton-Raphson

El método de *Newton-Raphson* es uno de los métodos numéricos más conocidos y poderosos para la resolución del problema de búsqueda de raíces de $f(x) = 0$. También puede ser usado para encontrar el máximo o mínimo de una función, encontrando los ceros de la derivada de la función. A diferencia de otros métodos de Análisis Numérico, el método de Newton-Raphson no trabaja sobre un intervalo sino que basa su fórmula en un proceso iterativo. Existen diversas maneras de introducir el método de Newton-Raphson. La primera forma de introducir el método de Newton-Raphson es desde un enfoque intuitivo basado en el polinomio de Taylor. Otra posibilidad es la de derivar el método de Newton-Raphson como una técnica simple para obtener una convergencia más rápida de la que ofrecen otros tipos de interacción funcional. La tercera manera de introducir el método, la cual estudiaremos a continuación es considerar la técnica gráficamente.

Supongamos que tenemos la aproximación x_n a la raíz x_0 de $f(x)$. Se traza la recta tangente a la curva en el punto $(x_n, f(x_n))$. Esta cruza al eje X en un punto x_{n+1} que será nuestra siguiente aproximación a la raíz x_0 .

Para calcular x_{n+1} calculemos primero la ecuación de la recta tangente.

Es decir

$$m = f'(x_n)$$

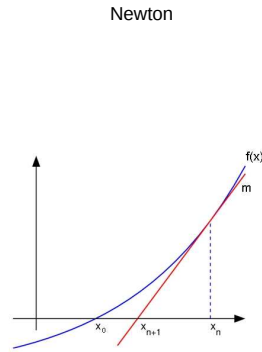


FIGURA 3.1. Método de Newton

entonces

$$y - f(x_n) = f'(x_n)(x - x_n)$$

hacemos $y = 0$

Por tanto tenemos que:

$$x = x_n - \frac{f(x_n)}{f'(x_n)}$$

Esta es la fórmula de Newton-Raphson la cual nos permite calcular la siguiente aproximación.

Luego

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \text{ si } f'(x_n) \neq 0$$

El método de Newton-Raphson no trabaja sobre intervalos donde nos asegure que encontraremos raíz, y de hecho no tenemos ninguna garantía de que nos aproxime a dicha raíz. Existen ejemplos donde este método no converge y en esos casos se dice que el método diverge, sin embargo en los casos donde converge lo hace con mucha rapidez por lo cual es uno de los métodos preferidos por excelencia. En el caso que $f'(x_n) = 0$, el método no se puede aplicar. Geométricamente esto significa que la recta tangente es horizontal y por tanto no interseca al eje X en ningún punto, a menos que coincida con este, en cuyo caso x_n mismo es una raíz de $f(x)$.

3. Análisis de Redes

El análisis de Redes Biológicas y Redes Sociales está relacionado con la ocurrencia de cambios en la red así como también la conexión entre los actores (nodos). En estas redes interesa saber qué pasa cuando se establecen nuevas relaciones entre los actores, cuando se rompen algunas de ellas; cuando se incluyen otros actores; cuando se excluye o cuando sale de la red un actor, el actor de mayor importancia, el mas central de la red entre otros.

Inicialmente el conjunto de datos de las redes biológicas que utilizamos se encontraban en un formato complejo para ser empleado en un programa común, además las redes a utilizar son numerosas por lo que utilizamos *Pajek*, una herramienta poderosa para el análisis de estas redes. Con esta herramienta vamos a tener información más detallada del comportamiento de la red.

Para realizar este análisis es importante tener claro algunas definiciones.

primero recordemos que:

- *Grado*: el grado de un nodo como el número de ramas o conectores asociadas al nodo.
- *Distribución de Grado*: La distribución de grado de una red es la la distribución de probabilidad del número de conexiones de los nodos, la caracterizamos por $\mathbb{P}(k)$ que representa la probabilidad de que un nodo tenga k conexiones.
- *Grado del Exponente*: Se refiere al valor γ de la distribución de grado de una red que satisface una ley de potencias. Dicho exponente es importante porque permite inferir la existencia de nodos principales llamados "hubs" que caracterizan a las redes libres de escala.
- *Distancia mínima*: La distancia entre dos nodos a y b se denota por d_{ab} , indica el menor número de ramas o conectores necesarios para unir los dos nodos. En otras palabras indica la longitud del camino más corto entre a y b .

Medida de Centralidad El problema de la centralidad surge al tratar de responderse preguntas como: Si una red de comunicación es muy grande y heterogénea ¿ Como se logra el intercambio de información entre sus integrantes?. Un ejemplo concreto de una red donde esta pregunta cobra mucho sentido sería; siendo La India una nación tan grande y heterogénea ¿como es posible gobernarla?. La respuesta conduce a la afirmación de que existen centros

en la red que vinculan y entretajan los elementos de diversas categorías. Los nodos que representan mayor importancia desde este punto de vista son aquellos que realizan mayor interconexión entre diferentes nodos en la red. En este sentido mientras mayor sea el número de caminos en los cuales un nodo toma parte, mayor es la importancia del nodo.

Asumiendo que la interacción toma el camino más corto entre dos nodos, representamos la medida de un nodo por su *medida de intermediación* (betweenness centrality). Si $\sigma(j, i, k)$ representa el número de caminos más cortos entre el nodo j y el nodo k que pasan por el nodo i y $\sigma(j, k)$ representa el número total de caminos más cortos entre el nodo j y el nodo k , la medida de centralidad del nodo i viene dada por:

$$B_i = \sum_{j \neq k} \frac{\sigma(j, i, k)}{\sigma(j, k)}$$

Otras medidas de centralidad se definen usando el valor de la distancia entre nodos. Si consideramos $d_{i,j}$ el valor de la distancia entre el nodo i y el nodo j , definimos la centralidad de proximidad por:

$$CC_i = \sum_{i \neq j} \frac{1}{d_{i,j}}$$

El interés de estudiar redes biológicas y redes libre de escala es porque ambas comparten ciertas características las cuales se pueden resumir en las siguientes:

- Una distribución que depende de un parámetro.
- Presencia de Hubs en las redes biológicas y un gran número de nodos con pocas conexiones puede ser considerada como una característica de redes biológicas libre de escala.
- Redes libres de escala poseen la propiedad de mundo pequeño.
- Robustas a rupturas aleatorias.

Una única propiedad de los sistemas libres de escala, es su invarianza para los cambios de escala. El término *libre de escala* refiere a un sistema definido por un funcional de la forma $f(x)$ que no se altera por un factor multiplicativo debido a un cambio de escala de la variable independiente x . Por tanto tenemos el siguiente resultado.

PROPOSICIÓN 3.1. *Sea f una función positiva, definida para $x > 0$, no decreciente, derivable entonces $f(ax) = bf(x)$ para todo $x > 0$ si y solo si f es una función potencial, es decir, $f(x) = cx^r$ con $r > 0$, $x > 0$.*

DEMOSTRACIÓN. ($\implies$)

Veamos que si $f(ax) = bf(x)$ para todo $x > 0$ entonces f sigue una ley de potencia. Observe que la constante b depende de a , la relación $f(ax) = bf(x)$ para todo $x > 0$ nos permite expresar b en función de a mediante la relación $b = \frac{f(a)}{f(1)}$, donde $f(1)$ no puede ser igual a cero. Definamos $g(x) = \frac{f(x)}{f(1)}$, entonces, $g(1) = 1$, observe que

$$g(ax) = \frac{f(ax)}{f(1)} = b \frac{f(x)}{f(1)} = bg(x).$$

Así $g(ax) = bg(x)$ para todo $x > 0$, con $g(1) = 1$, en este caso $b = \frac{g(a)}{g(1)} = g(a)$. Consideremos la función $G(x) = g(\exp(x))$, para cada x_0 fijo,

$$G(x_0 + y) = g(\exp(x_0)\exp(y)) = bg(\exp(y)) = g(\exp(x_0)g(\exp(y))),$$

de estas relaciones se concluye que

$$G(x_0 + y) = G(x_0)G(y).$$

Si $y = 0$, obtenemos $G(0) = \frac{G(x_0)}{G(x_0)} = 1$ Obteniendo que la función G satisface:

$$G(x + h) - G(x) = G(x)(G(h) - 1).$$

$$\lim_{h \rightarrow 0} \frac{G(x + h) - G(x)}{h} = G(x) \lim_{h \rightarrow 0} \frac{G(h) - 1}{h}.$$

Por lo tanto la función G satisface la siguiente ecuación diferencial:

$$G'(x) = G'(0)G(x), \text{ con } G'(0) = r > 0, G(0) = g(1) = 1,$$

note que $G'(0) = g'(1) = \frac{f'(1)}{f(1)} > 0$, pues f es positiva y no decreciente. Hallemos la solución de la ecuación diferencial.

$$y' = ry, \text{ con la condición inicial } y(0) = 1.$$

El factor integrante es $\exp(-ry)$, por tanto la ecuación se puede reescribir como

$$\exp(-ry)(y' - ry) = 0.$$

lo que equivale a

$$\frac{d(\exp(-ry)y)}{dx} = 0.$$

La solución de esta ecuación es $y\exp(-rx) = c$, donde c es una constante. Así,

$$y = c\exp(rx)$$

. Como $y(0) = 1$ obtenemos $c = 1$, entonces $y = \exp(rx)$. Puesto que G satisface la ecuación diferencial resuelta con $G(0) = 1$ se concluye que $G(x) = \exp(rx)$, de donde $g(x) = G(\ln(x)) = \exp(r\ln(x)) = \exp(\ln(x^r)) = x^r$ y la función que nos interesa

$$f(x) = f(1)g(x) = f(1)x^r.$$

Podemos concluir que f es una función potencia.

($\Leftarrow$)

Veamos que si f sigue una ley de potencia entonces $f(ax) = bf(x)$. Como f sigue una ley de potencia, entonces f se puede escribir como $f(x) = \alpha x^k$.

Luego

$$f(ax) = \alpha(ax)^k = \alpha a^k x^k = bx^k = bf(x).$$

Por tanto tenemos que $f(ax) = bf(x)$ donde $b = \alpha a^k, b \in \mathbb{R}$.

□

La proposición anterior nos muestra la propiedad de invarianza en cambios de escala o mejor conocida como *Autosimilaridad*.

El estimador de la distribución de probabilidad para cada red se calcula utilizando un comando en Matlab, de igual forma se realiza un algoritmo para la prueba estadística.

Pasos del Algoritmo

1. Leer la tabla de frecuencias y las clases obtenidas de los datos experimentales.
2. Calcular el estimador de máxima verosimilitud para cada red y graficar los datos experimentales y los datos con la distribución hipotética.
3. Aplicar la prueba estadística
4. Calcular el p valor (con este valor se tomará la decisión si las redes biológicas son libres de escala.)

CAPÍTULO 4

Análisis

1. Introducción

Se hace un análisis estadístico de 6 redes biológicas. Un método extensamente utilizado para encontrar un indicador de una red libre de escala ha sido de ajustar una recta (en escala log-log) a la frecuencia relativa de conectividad.

Las Redes a trabajar son de proteínas (Uetz, Ito, Li, Rain) y genéticas (Lee, Guelzim) obtenidas de la dirección de la revista de Biología Computacional www.stats.gla.ac.uk/raya/howscalefree/howscalefree.htm. Para realizar la prueba estadística vamos a trabajar con las de frecuencias obtenida de la data las cuales se obtuvieron utilizando Pajek y nos indica el número de conexiones de cada nodo en la red.

2. Representación de la red

Anteriormente vimos que si $p(k) = \alpha k^{-\gamma}$ la red es libre de escala y la única solución en este caso es:

$$p(k) = \alpha k^{-\gamma}, k \geq 1,$$

γ es el parámetro de la distribución. El número de conexiones de la red se obtiene experimentalmente y se estima $p(k)$ por la frecuencia relativa observada.

$$N(k) = \frac{\# \text{ de nodos con } k \text{ conexiones}}{\# \text{ total de nodos}}.$$

Ahora para ver si los datos provienen de una red libre de escala ajustamos a $N(k)$ una línea recta.

$$\log(p(k)) = \beta - \gamma \log(k).$$

En la función de ley de potencia, $p(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$, donde $\zeta(\gamma)$ es la función Zeta de Riemann, $k \geq 1$, el parámetro γ se estimó utilizando el método de máxima verosimilitud suponiendo que las conectividades de los nodos son independientes.

A continuación presentamos los gráficos de cada red con las rectas ajustadas. En estos gráficos podemos observar un comportamiento de ley de potencia. Sin embargo, no existe un buen ajuste de todos los datos por lo que no tenemos suficientes razones para suponer un comportamiento libre de escala a priori.

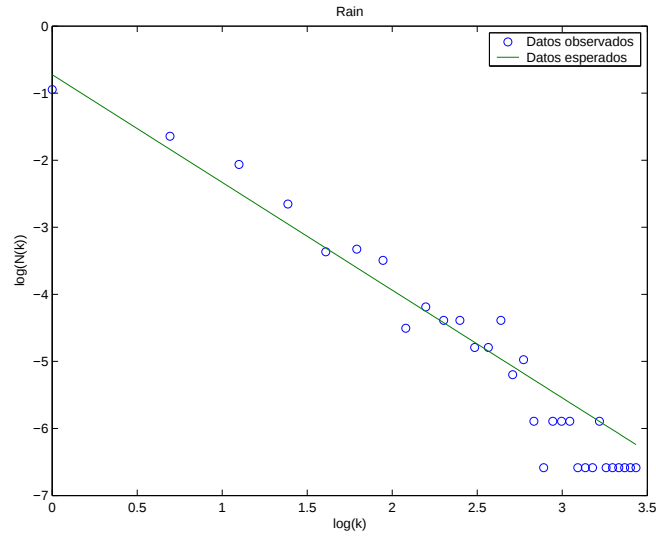


FIGURA 4.1. Recta ajustada para la red Rain

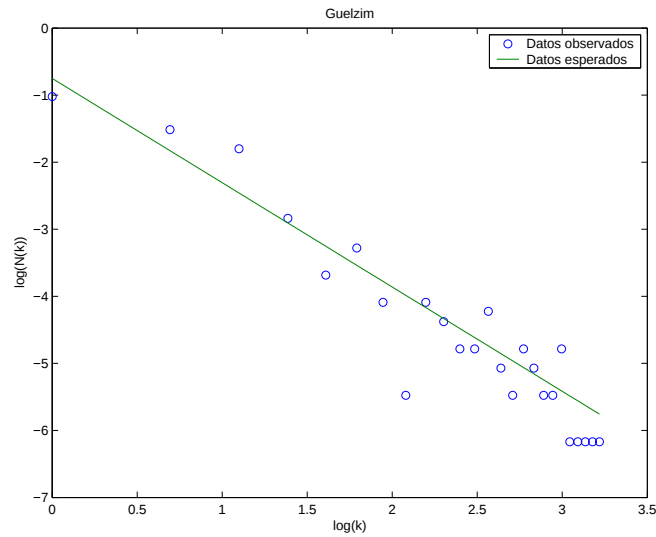


FIGURA 4.2. Recta ajustada para la red Guelzim

En todos los gráficos obtenidos se puede observar un comportamiento similar. A continuación mostramos los graficos de las redes Uetz y Li-data.

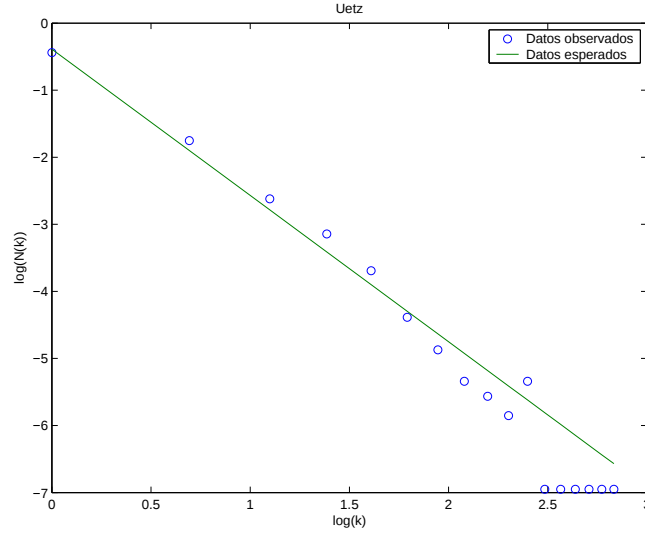


FIGURA 4.3. Recta ajustada para la red Uetz

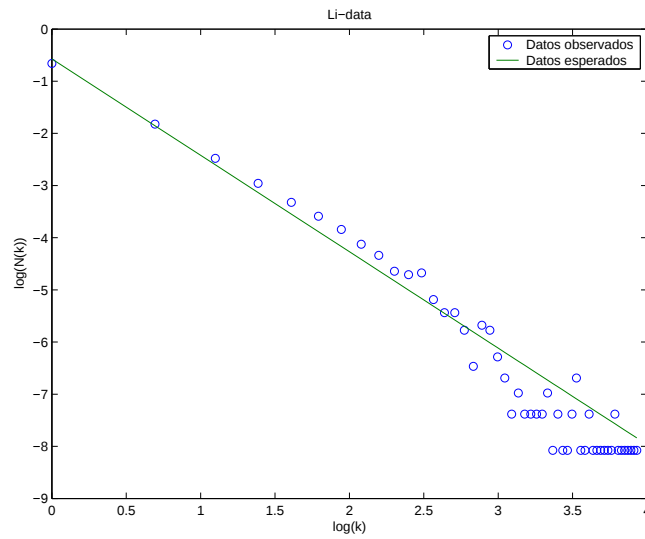


FIGURA 4.4. Recta ajustada para la red Li-data

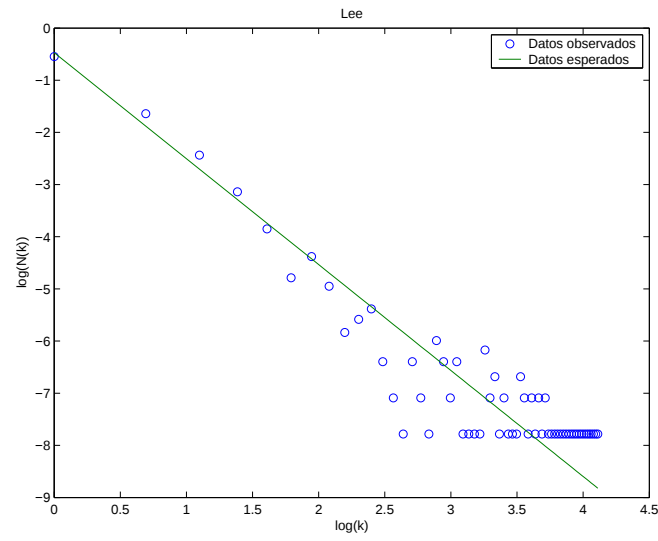


FIGURA 4.5. Recta ajustada para la red Lee

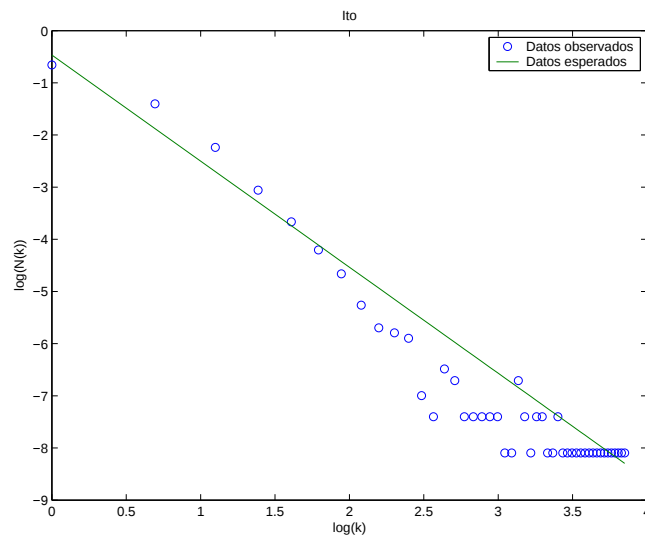


FIGURA 4.6. Recta ajustada para la red Ito

A continuación vamos a estudiar las características de cada red.

Grado: El grado de cada nodo en la red se puede obtener mediante la tabla de frecuencias [Apéndice A]. En estas tablas pudimos observar la presencia de Hubs (pocos nodos están altamente conectados) y por otra parte muchos nodos están poco conectados. Este tipo de conexiones es muy común en una red libre de escala y en consecuencia puede ser un indicador de que las redes que estamos estudiando tengan un comportamiento libre de escala.

Distribución de Grado: En esta oportunidad estamos estudiando las redes biológicas utilizando la Distribución de Ley de Potencia $\mathbb{P}(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$, $k = 1, 2, \dots$, y la Distribución Exponencial Truncada $p(k) = \frac{k^{-\gamma} \exp(-k/k_c)}{C_1(\gamma; k_c)}$, que estudiamos más adelante.

Grado del Exponente: Son los parámetros que se van a estimar.

Distancia mínima: Para cada una de las redes se calcula la distancia mínima entre los nodos mas alejados de la red, es decir, el diámetro de la red y se obtiene:

Guelzim: The longest shortest path from YGL043W (44) to YCL025C (385). Diameter is 11.

Ito: The longest shortest path from YDR072C (531) to YDR477W (742). Diameter is 14.

Lee: The longest shortest path from YBR156C (270) to YOL103W (2101). Diameter is 11.

Li data: The longest shortest path from B0213.11 (26) to K11D12.10 (1836). Diameter is 13.

Rain: The longest shortest path from HP0069 (19) to HP0857 (678). Diameter is 9.

Uest: The longest shortest path from YDR017C (163) to YLR150W (674). Diameter is 18.

En todas las redes se pudo obtener a pesar de ser redes muy numerosas, que el camino entre los nodos mas alejados en la red es pequeño. Esto es una clara evidencia de la característica de pequeño mundo que comparten las redes biológicas.

Centralidad Aproximada:

En la siguiente tabla se presenta el nodo con mayor centralidad aproximada en la red. En todas estas redes se obtuvo una centralidad muy baja.

Red	Centralidad Aproximada	Representante
Rain	0.33943299	128 "HP0775"
Guelzim	0.28649230	27 "YDR207C"
Lee	0.34488376	58 "YKL112W"
Ito	0.28043077	1767 "YJR091C"
Uest	0.12363243	746 "YML064C"
Li-data	0.28978699	2489 "W07B3.2"

Medida de Intermediación (Betweenness Centrality):

En la siguiente tabla se presenta el nodo con mayor betweenness centrality en la red. Recordemos que mientras mayor sea el valor del betweenness centrality mayor es el número de caminos en los cuales un nodo o un enlace toma parte y mayor será su importancia en la red. Al realizar estos cálculos se obtuvo nodos con Betweenness Centrality igual a cero, es decir no tienen gran importancia en la red.

Red	Betweenness Centrality	Representante
Rain	0.15600964	248 "HP1409"
Guelzim	0.16378876	15 "YCR065W"
Lee	0.20705564	822 "YFL002W-B"
Ito	0.21440269	1096 "YGL169W"
Uest	0.12678923	110 "YDL012C"
Li-data	0.14210171	500 Ç35B1.1"

3. Método de Máxima verosimilitud para cada Red

Distribución hipotética de una red libre de escala

Sea $\mathbb{P}(x_i) = \frac{x_i^{-\gamma}}{\zeta(\gamma)}$ la función de probabilidad. Si $x_1, x_2, \dots, x_M$ representa la muestra observada, N es la máxima conectividad observada, $1 \leq N \leq M$, k_i es el número de nodos con i conexiones o frecuencia de la observación i donde k_i puede ser igual cero, $i = 1, 2, \dots, N$ y $\sum_{i=1}^N k_i = M$. Por tanto tenemos que:

$$\begin{aligned} L(x_1, x_2, \dots, x_M/\gamma) &= \mathbb{P}(x_1) \cdot \mathbb{P}(x_2) \cdot \mathbb{P}(x_3) \dots \mathbb{P}(x_M) \\ &= \mathbb{P}(1)^{k_1} \cdot \mathbb{P}(2)^{k_2} \cdot \mathbb{P}(3)^{k_3} \dots \mathbb{P}(N)^{k_N} \\ &= \left(\frac{1^{-\gamma}}{\zeta(\gamma)} \right)^{k_1} \cdot \left(\frac{2^{-\gamma}}{\zeta(\gamma)} \right)^{k_2} \cdot \left(\frac{3^{-\gamma}}{\zeta(\gamma)} \right)^{k_3} \dots \left(\frac{N^{-\gamma}}{\zeta(\gamma)} \right)^{k_N} \\ &= \prod_{j=1}^N \left(\frac{j^{-\gamma}}{\zeta(\gamma)} \right)^{k_j} \text{ Función de Verosimilitud.} \end{aligned}$$

Quedando la función de verosimilitud expresada de la siguiente forma:

$$L(x_1, x_2, \dots, x_M/\gamma) = \prod_{j=1}^N \left(\frac{j^{-\gamma}}{\zeta(\gamma)} \right)^{k_j}.$$

donde N es la máxima conectividad observada en la red.

Como el logaritmo es una función estrictamente creciente, se tiene que tanto la función de verosimilitud como su logaritmo alcanzan el máximo en el mismo punto, por esta razón en la práctica los estimadores de máxima verosimilitud se obtienen derivando el logaritmo de la función de verosimilitud.

Como los cálculos son mas simples utilizamos el logaritmo de la función de verosimilitud y esta función se llamará $l(x_1, x_2, \dots, x_M/\gamma)$

Por tanto nos queda

$$\begin{aligned}
l(x_1, x_2, \dots, x_M/\gamma) &= \ln \left(\prod_{j=1}^N \left(\frac{j^{-\gamma}}{\zeta(\gamma)} \right)^{k_j} \right) = \ln \left(\prod_{j=1}^N \frac{j^{-\gamma k_j}}{\zeta(\gamma)^{k_j}} \right) = \sum_{j=1}^N \ln \left(\frac{j^{-\gamma k_j}}{\zeta(\gamma)^{k_j}} \right) \\
&= \sum_{j=1}^N \ln(j^{-\gamma k_j}) - \ln(\zeta(\gamma)^{k_j}) \\
&= -\gamma \sum_{j=1}^N k_j \ln(j) - k_j \ln(\zeta(\gamma)) \\
&= -\gamma \sum_{j=1}^N k_j \ln(j) - \sum_{j=1}^N k_j \ln(\zeta(\gamma)) \\
&= -\gamma \sum_{j=1}^N k_j \ln(j) - \ln(\zeta(\gamma)) \sum_{j=1}^N k_j; \quad \text{pero} \quad \sum_{j=1}^N k_j = M \\
&= -\gamma \sum_{j=1}^N k_j \ln(j) - M \ln(\zeta(\gamma)), \quad \text{donde} \quad \zeta(\gamma) = \sum_{j=1}^N j^{-\gamma}.
\end{aligned}$$

La función de verosimilitud queda expresada de la siguiente forma

$$l(x_1, x_2, \dots, x_M/\gamma) = -\gamma \sum_{j=1}^N k_j \ln(j) - M \ln(\zeta(\gamma)).$$

Para obtener el máximo de la función de verosimilitud $l(x_1, x_2, \dots, x_M/\gamma)$, derivamos e igualamos a cero y mediante el método de Newton-Raphson conseguimos el estimador de la distribución de probabilidad.

Debemos calcular la segunda derivada de la función de verosimilitud ya que al utilizar el método de Newton-Raphson la función con que estamos trabajando es la derivada de la función de verosimilitud.

Primera derivada

$$\frac{\partial(l(x_1, x_2, \dots, x_M/\gamma))}{\partial\gamma} = M \frac{\sum_{j=1}^N j^{-\gamma} \ln(j)}{\sum_{j=1}^N j^{-\gamma}} - \sum_{j=1}^N k_j \ln(j).$$

Segunda derivada

$$\frac{\partial^2(l(x_1, x_2, \dots, x_M/\gamma))}{\partial \gamma^2} = M \frac{\left[\left(\sum_{j=1}^N j^{-\gamma} \ln(j) \right)^2 - \sum_{j=1}^N (\ln(j))^2 j^{-\gamma} \sum_{k=1}^N k^{-\gamma} \right]}{\left[\sum_{k=1}^N k^{-\gamma} \right]^2}.$$

El método de Newton-Raphson queda expresado de la siguiente forma:

$$\gamma_{n+1} = \gamma_n - \frac{f'(\gamma_n)}{f''(\gamma_n)}.$$

donde

$$f'(\gamma_n) = \frac{\partial(l(x_1, x_2, \dots, x_M/\gamma))}{\partial \gamma}(\gamma_n).$$

y

$$f''(\gamma_n) = \frac{\partial^2(l(x_1, x_2, \dots, x_M/\gamma))}{\partial \gamma^2}(\gamma_n).$$

Sabemos que en las redes libres de escala el parámetro γ se encuentra en el rango $2 < \gamma \leq 3$, por esta razón para optimizar el tiempo de búsqueda del estimador tomamos como punto de inicio ó primera raíz el punto $\gamma_1 = 2$. En este trabajo estamos eliminando los nodos huérfanos, es decir, los nodos que no están conectados con ningún otro y cuya probabilidad es cero.

Distribución de Ley de Potencia Exponencial Truncada

Otra distribución muy importante que sigue una ley de potencia y es aplicada a algunas redes biológicas es la función exponencial truncada y viene dada por:

$$p(k) = \frac{k^{-\gamma} \exp(-k/k_c)}{C_1(\gamma; k_c)}.$$

donde $C_1(\gamma; k_c) = \sum_{j=1}^{\infty} j^{-\gamma} \exp(-j/k_c)$ es la función que normaliza la distribución, γ es el exponente de la ley de potencia y k_c es el parámetro de truncamiento. Esta distribución corresponde a un proceso de autoorganización en el que nuevos nodos (genes o proteínas) se conectan con mayor probabilidad a los nodos más conectados (como en las distribuciones libres de escala), pero con un límite máximo en el número de conexiones que un nodo puede mantener. En otras palabras, la distribución de conectividades viene dada por una ley de potencia hasta llegar a conectividades mayores de cierto umbral k_c . A partir de entonces, la distribución se trunca resultando que valores de conectividad por encima de dicho umbral son más improbables de lo que predeciría una ley de potencia.

Aplicando igual procedimiento a la parte anterior, para encontrar un indicador de una red libre de escala se ajustaron los datos observados o la distribución de frecuencia relativa con los datos esperados por la distribución hipotética (en este caso es la distribución de ley de potencia exponencial truncada).

A continuación presentamos los gráficos de cada red con las rectas ajustadas utilizando la Distribución de Ley de Potencia Exponencial Truncada. No obstante se observó una diferencia considerable entre algunos datos observados y datos esperados. Estos cálculos se obtuvieron utilizando el algoritmo empleado con la distribución de ley de potencia con la diferencia que utilizamos la Distribución de Ley de Potencia Exponencial Truncada.

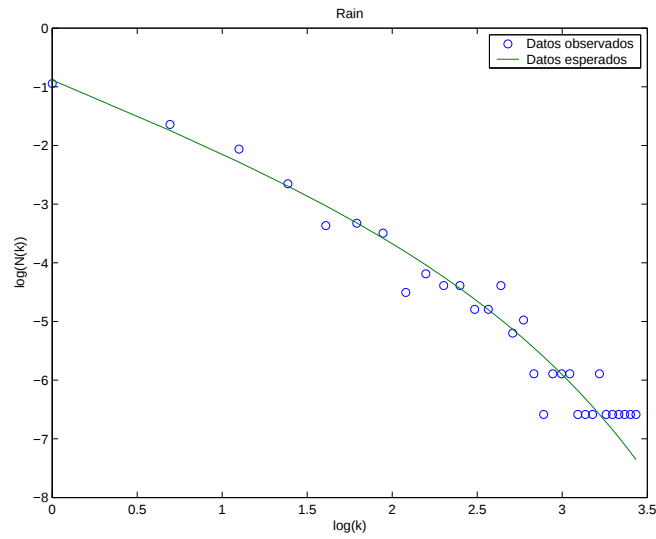


FIGURA 4.7. Datos ajustados para la red Rain utilizando la Distribución Exponencial

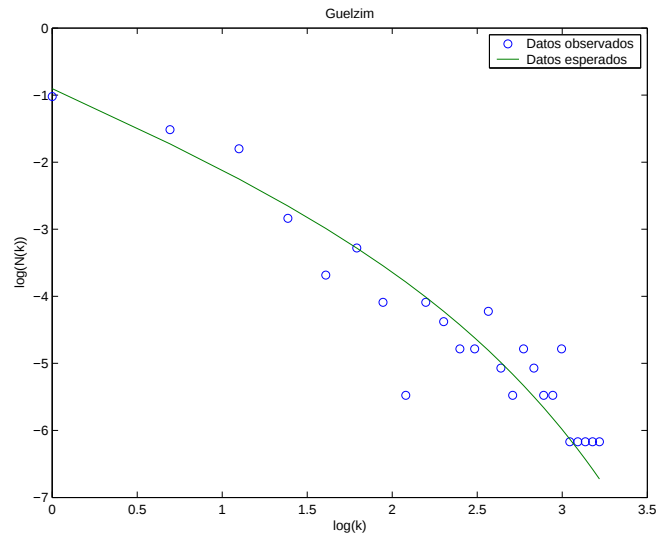


FIGURA 4.8. Datos ajustados para la red Guelzim utilizando la Distribución Exponencial

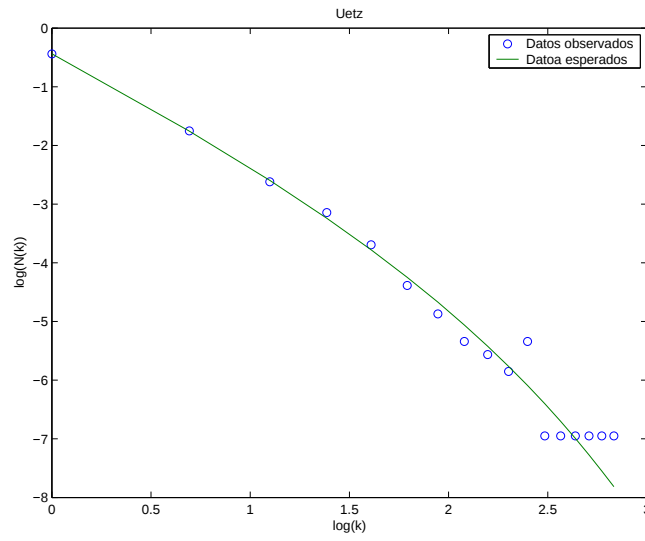


FIGURA 4.9. Datos ajustados para la red Uetz utilizando la Distribución Exponencial

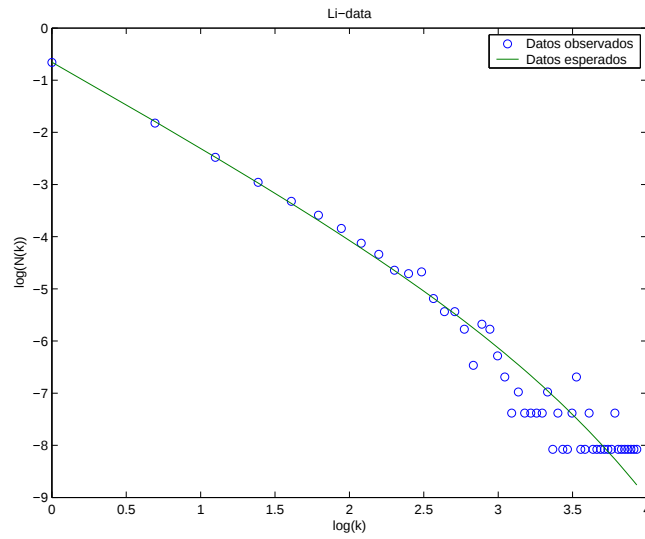


FIGURA 4.10. Datos ajustados para la red Li-data utilizando la Distribución Exponencial

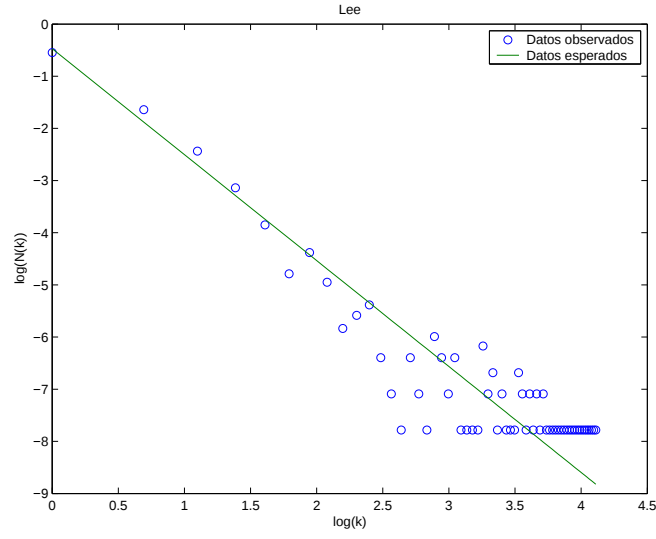


FIGURA 4.11. Datos ajustados para la red Lee utilizando la Distribución Exponencial

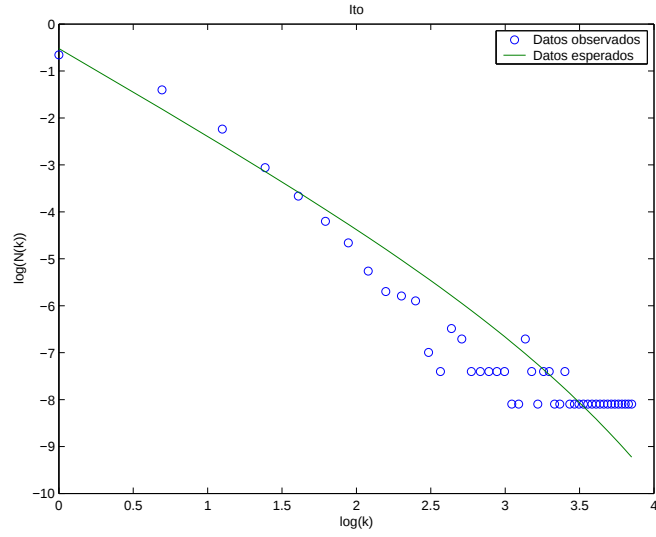


FIGURA 4.12. Datos ajustados para la red Ito utilizando la Distribución Exponencial

Recordemos que en la función $p(k) = \frac{k^{-\gamma} \exp(-k/k_c)}{C_1(\gamma; k_c)}$, donde $C_1(\gamma; k_c) = \sum_{j=1}^N j^{-\gamma} \exp(-j/k_c)$ es la función que normaliza la ley de potencia, se estima γ y k_c utilizando el Método de Máxima Verosimilitud.

Método de Maxima Verosimilitud para la Distribución de Ley de Potencia Exponencial Truncada

Sea $\mathbb{P}(x_i) = \frac{k^{-\gamma} \exp(-k/k_c)}{C_1(\gamma; k_c)}$ la función de probabilidad. Si $x_1, x_2, \dots, x_M$ representa la muestra observada, N es la máxima conectividad observada, $1 \leq N \leq M$, k_i es el número de nodos con i conexiones o frecuencia de la observación i donde k_i puede ser igual cero, $i = 1, 2, \dots, N$ y $\sum_{i=1}^N k_i = M$. Por tanto tenemos que:

$$\begin{aligned} L(x_1, x_2, \dots, x_M/\gamma, k_c) &= \mathbb{P}(x_1) \cdot \mathbb{P}(x_2) \cdot \mathbb{P}(x_3) \dots \mathbb{P}(x_M) \\ &= \mathbb{P}(1)^{k_1} \cdot \mathbb{P}(2)^{k_2} \cdot \mathbb{P}(3)^{k_3} \dots \mathbb{P}(N)^{k_N} \\ &= \left(\frac{1^{-\gamma} \exp(-1/k_c)}{C_1(\gamma; k_c)} \right)^{k_1} \left(\frac{2^{-\gamma} \exp(-2/k_c)}{C_1(\gamma; k_c)} \right)^{k_2} \dots \left(\frac{N^{-\gamma} \exp(-N/k_c)}{C_1(\gamma; k_c)} \right)^{k_N} \\ &= \prod_{i=1}^N \left(\frac{i^{-\gamma} \exp(-i/k_c)}{C_1(\gamma; k_c)} \right)^{k_i} \text{ Función de Verosimilitud.} \end{aligned}$$

La función de verosimilitud queda expresada de la siguiente forma

$$L(x_1, x_2, \dots, x_M/\gamma, k_c) = \prod_{i=1}^N \left(\frac{i^{-\gamma} \exp(-i/k_c)}{C_1(\gamma; k_c)} \right)^{k_i}.$$

donde N es la máxima conectividad observada en la red.

Como es igual maximizar una función que su logaritmo (al ser esta una función estrictamente creciente) este máximo puede calcularse derivando parcialmente con respecto a γ y k_c la función de verosimilitud o bien su logaritmo, y tomando como el estimador de máxima verosimilitud que haga nula la derivada parcial con respecto a cada parámetro .

Como los cálculos son mas simples utilizamos el logaritmo de la función de verosimilitud y esta función se llamará $l(x_1, x_2, \dots, x_M/\gamma)$

Por tanto nos queda

$$\begin{aligned}
l(x_1, x_2, \dots, x_M/\gamma, k_c) &= \ln \left(\prod_{i=1}^N \left(\frac{i^{-\gamma} \exp(-i/k_c)}{C_1(\gamma; k_c)} \right)^{k_i} \right) \\
&= \ln \left(\prod_{j=1}^N \frac{[i^{-\gamma} \exp(-i/k_c)]^{k_i}}{C_1(\gamma; k_c)^{k_i}} \right) \\
&= \sum_{j=1}^N \ln \left(\frac{[i^{-\gamma} \exp(-i/k_c)]^{k_i}}{C_1(\gamma; k_c)^{k_i}} \right) \\
&= \sum_{j=1}^N [\ln(i^{-\gamma} \exp(-i/k_c))^{k_i} - \ln(C_1(\gamma; k_c)^{k_i})] \\
&= \sum_{j=1}^N [k_i \ln(i^{-\gamma} \exp(-i/k_c)) - k_i \ln[C_1(\gamma; k_c)]] \\
&= \sum_{j=1}^N k_i [\ln(i^{-\gamma}) + \ln(\exp(-i/k_c))] - \sum_{j=1}^N k_i \ln[C_1(\gamma; k_c)] \\
&= \sum_{j=1}^N [k_i \ln(i^{-\gamma}) + k_i \ln(\exp(-i/k_c))] - \ln[C_1(\gamma; k_c)] \sum_{j=1}^N k_i \\
&= -\gamma \sum_{j=1}^N k_i \ln(i) + \sum_{j=1}^N k_i (-i/k_c) - M \ln[C_1(\gamma; k_c)] \\
&= -\gamma \sum_{j=1}^N k_i \ln(i) - \frac{1}{k_c} \sum_{j=1}^N i k_i - M \ln[C_1(\gamma; k_c)].
\end{aligned}$$

Por tanto la función de verosimilitud queda de la siguiente forma

$$l(x_1, x_2, \dots, x_M/\gamma, k_c) = -\gamma \sum_{j=1}^N k_i \ln(i) - \frac{1}{k_c} \sum_{j=1}^N i k_i - M \ln[C_1(\gamma; k_c)].$$

Para obtener el máximo de la función de verosimilitud $l(x_1, x_2, \dots, x_M/\gamma, k_c)$, calculamos la primera derivada parcial de γ y k_c respectivamente e igualamos a cero y mediante el método de Newton-Raphson conseguimos los estimadores de estos parámetros.

Es importante destacar que debemos calcular la segunda derivada parcial de la función de verosimilitud con respecto a γ y k_c ya que al utilizar el método de Newton-Raphson la función con que estamos trabajando son las primera derivada de γ y k_j de la función de verosimilitud.

Primera derivada con respecto a γ

$$\frac{\partial(l(x_1, x_2, \dots, x_M/\gamma, k_c))}{\partial \gamma} = - \left[\sum_{j=1}^N k_j \ln(j) \right] - M \frac{\left[\sum_{j=1}^N (-j)^{-\gamma} \ln(j) \exp(-j/k_c) \right]}{\left[\sum_{j=1}^N (-j)^{-\gamma} \exp(-j/k_c) \right]}.$$

Segunda derivada con respecto a γ

$$\frac{\partial^2(l(x_1, x_2, \dots, x_M/\gamma, k_c))}{\partial \gamma^2} = -M \frac{\left[\sum_{j=1}^N j^{-\gamma} \ln(j)^2 \exp(-j/k_c) \right]}{\left[\sum_{j=1}^N j^{-\gamma} \exp(-j/k_c) \right]} + M \frac{\left[\sum_{j=1}^N (-j)^{-\gamma} \ln(j) \exp(-j/k_c) \right]^2}{\left[\sum_{j=1}^N j^{-\gamma} \exp(-j/k_c) \right]^2}.$$

Primera derivada con respecto a k_c

$$\frac{\partial(l(x_1, x_2, \dots, x_M/\gamma, k_c))}{\partial k_c} = \frac{\left[\sum_{j=1}^N j k_j \right]}{k_c^2} - M \frac{\left[\sum_{j=1}^N j^{-\gamma} j \exp(-j/k_c) \right]}{\left[\sum_{j=1}^N j^{-\gamma} \exp(-j/k_c) \right]}.$$

Segunda derivada con respecto a k_c

$$\begin{aligned} \frac{\partial^2(l(x_1, x_2, \dots, x_M/\gamma, k_c))}{\partial k_c^2} &= \frac{-2 \left[\sum_{j=1}^N j k_j \right]}{k_c^3} - M \frac{\left[\sum_{j=1}^N \frac{[-2j^{-\gamma+1} \exp(-j/k_c)]}{k_c^3} + j^{-\gamma+2} \exp(-j/k_c) \right]}{\sum_{j=1}^N j^{-\gamma} \exp(-j/k_c)} \\ &\quad + M \frac{\left[\frac{\sum_{j=1}^N j^{-\gamma+1} \exp(-j/k_c)}{k_c^2} \right]}{\left[\sum_{j=1}^N j^{-\gamma} \exp(-j/k_c) \right]}. \end{aligned}$$

4. Resultados

Grado del Exponente

En la siguiente tabla se presenta el estimador γ (o grado del exponente) de cada data utilizando la distribución de ley de potencia, calculado mediante el método de máxima verosimilitud. De los cuales tres de los exponentes se aproximan al grado 2 cuyo grado es característico de las redes libres de escala.

Data	γ
Rain	1.6061
Guelzim	1.5552
Uetz	2.1817
Li-data	1.8480
Lee	2.0312
Ito	2.0331

CUADRO 1. Estimadores de la distribución de ley de potencia.

A continuación se muestran los resultados de la prueba χ^2 para cada una de las redes.

Data	# Nodos	γ	Test	p-valor
Rain	732	1.6061	49.74228	1.5505×10^{-6}
Guelzim	478	1.5552	68.6956	2.7424×10^{-11}
Uetz	1044	2.1817	16.91389	0.0310
Li-data	3216	1.8480	84.8783	2.511×10^{-9}
Lee	2398	2.0312	93.93423	1.1676×10^{-12}
Ito	3278	2.0331	390.52632	0

CUADRO 2. Resultados de la Prueba χ^2 para cada red.

El análisis estadístico que se muestra en la tabla 2 establece que los datos de la red de interacción biológica difieren significativamente de una distribución de ley de potencia. Estos resultados están basados en la prueba estadística *Chi-Cuadrado* y demuestran claramente

que las datas provenientes de estas redes de interacción biológica no siguen una distribución de ley de potencia contrario a lo que se había supuesto al graficar la recta en la escala log-log.

Recordemos que en las redes libres de escala la probabilidad $\mathbb{P}(k)$ de que un nodo de la red esté conectado con k nodos es proporcional a $k^{-\gamma}$, es decir, sigue una ley de potencias y el exponente γ se encuentra entre el rango $2 < \gamma \leq 3$.

En ciertas oportunidades la distribución de conectividad en algunas redes biológicas pueden ser mejor explicadas por una Distribución de Ley de Potencia Exponencial Truncada, es decir la ley de potencia de la forma $\mathbb{P}(k) = \frac{k^{-\gamma} \exp(-k/k_c)}{C_1(\gamma; k_c)}$ donde $C_1(\gamma; k_c) = \sum_{j=1}^{\infty} j^{-\gamma} \exp(-j/k_c)$ es la función que normaliza la distribución, $k \geq 1$, k_c es el parámetro tal que el número de conexiones es menor que el esperado por la red libre de escala.

Grado del Exponente

En la siguiente tabla se presentan los estimadores γ y k_c (grados de los exponentes) de cada data utilizando la Distribución de Ley de Potencia Exponencial Truncada, calculados mediante el método de máxima verosimilitud. En esta oportunidad el estimador γ no está próximo al grado 2. Sin embargo no es un argumento suficiente para decir si la red es o no libre de escala. Por otro lado sabemos que k_c funciona como un umbral donde la distribución se trunca con lo que valores de conectividad por encima de dicho umbral son más improbables de lo que predeciría una ley de potencia.

Data	γ	k_c
Rain	1.11374	11.3485
Guelzim	1.03909	9.69371
Uetz	1.65974	5.98149
Li-data	1.58981	26.9452
Lee	2.03	294
Ito	1.80779	26.3855

CUADRO 3. Estimadores de la Distribución de Ley de Potencia Exponencial Truncada.

Al aplicar la prueba estadística a la la Distribución de Ley de Potencia Exponencial Truncada para las datas de interacción (tabla 4) se obtuvo que varias datas son consistentes

con los resultados obtenidos al utilizar la Distribución de Ley de Potencia truncada (tabla 2). Asimismo se encontró que para las redes Uetz y Li- Data se obtuvo un p valor es grande por lo que no podemos rechazar que estemos en presencia de un comportamiento libre de escala. Sin embargo, entre estas datas (Uetz y Li- Data) que son consideradas provenientes de una distribución de ley de potencia truncada, los estimadores del parámetro k_c tienen valores pequeños comparado con el tamaño de la red; esto significa que esas redes muestran un comportamiento libre de escala solo en un rango muy pequeño de conectividad: $1 \leq k < k_c$.

Data	# Nodos	γ	k_c	Test	p-valor
Rain	732	1.11374	11.3485	15.29939	0.162
Guelzim	478	1.03909	9.69371	41.87589	1.4288×10^{-6}
Uetz	1044	1.65974	5.98149	2.34771	0.9381
Li-data	3216	1.58981	26.9452	17.10063	0.7050
Lee	2398	2.03	294	93.93287	4.68×10^{-13}
Ito	3278	1.80779	26.3855	281.64961	0

CUADRO 4. Resultados de la Prueba χ^2 con la Distribución Exponencial Truncada.

5. Conclusión

En este trabajo encontramos que el número de conexiones en una red biológica difiere significativamente de la distribución de ley de potencia y en consecuencia esas redes no son libres de escala contrario a lo supuesto al ajustar una recta a los datos experimentales. Por otro lado al utilizar la Distribución de Ley de Potencia Exponencial Truncada se obtuvo igual resultado excepto en dos redes (Uetz y Li- Data) cuyo p valor es grande por lo que no podemos rechazar que estemos en presencia de un comportamiento libre de escala. Este es un resultado importante para el estudio de las propiedades de las redes biológicas ya que si bien comparten ciertas propiedades no se puede suponer que tengan un comportamiento libre de escala y por tanto no se puede predecir el comportamiento de la interacción de genes, proteína y metabolismo, por esta razón se deben considerar con más cuidado las conclusiones al momento de trabajar con el comportamiento de estas redes.

Por otro lado con los resultados obtenidos al calcular el diámetro de la red obtuvimos que a pesar de ser redes numerosas el camino mas largo entre los nodos mas alejados de la red es muy corto comparado con el tamaño de la red por lo que podemos suponer que estas redes sean de mundo pequeño. Este resultado es de gran importancia ya que si bien las redes biológicas no son libres de escala sería de gran avance para la Biología saber si de verdad estas redes son de mundo pequeño. Asimismo se debería matematizar más las redes y sistemas complejos en general y así se puedan trabajar estos sistemas de mejor forma para obtener resultados con mayor base teórica del comportamiento de las redes.

Bibliografía

- [1] D. SANCHEZ, Estadística Modelos y métodos, *Alianza Editorial*. (1991), 360-364.
- [2] H. CRAMER, Métodos Matemáticos de Estadística, *Aguilar, S.A. De Ediciones*. (1960), 447-498.
- [3] M.SPIEGEL, Estadística, *Mac Graw Hill*. (1991), 268-278.
- [4] M.E.J. NEWMAN *The structure and function of complex networks*. SIAM Review 45, 167–256, 2003.
- [5] R.L. BURDEN, J.D. FAIRES *Análisis Numérico*. Editorial beroamericana. 2002.
- [6] OLIVARES M.M., *Estadística*. Guía para estudiantes de Matemática, Facultad de Ciencias UCV, 2004.
- [7] *Redes Biológicas*. Material elaborado en el Seminario de posgrado en Modelos Aleatorios, Facultad de Ciencias UCV, 2006.
- [8] R. KHAIN, E. WIT *How Scale Free Are Biological Networks*. Journal of Computational Biology 3, 810-818, 2006.

Apéndice

1. Apéndice A

A continuación se presentan las tablas de frecuencias de cada red de interacción empleadas en la prueba estadística.

Cluster	Freq	Freq %	CumFreq	CumFreq %	Representative
1	281	38.3880	289	39.4809	HP0064
2	140	19.1257	429	58.6066	HP0045
3	92	12.5683	521	71.1749	HP0002
4	51	6.9672	572	78.1421	HP0001
5	25	3.4153	597	81.5574	HP0073
6	26	3.5519	623	85.1093	HP0031
7	22	3.0055	645	88.1148	HP0066
8	8	1.0929	653	89.2077	HP0184
9	11	1.5027	664	90.7104	HP0145
10	9	1.2295	673	91.9399	HP0042
11	9	1.2295	682	93.1694	HP0010
12	6	0.8197	688	93.9891	HP0149
13	6	0.8197	694	94.8087	HP0712
14	9	1.2295	703	96.0383	HP0016
15	4	0.5464	707	96.5847	HP0887
16	5	0.6831	712	97.2678	HP0005
17	2	0.2732	714	97.5410	HP0088
18	1	0.1366	715	97.6776	HP1489
19	2	0.2732	717	97.9508	HP0249
21	2	0.2732	719	98.2240	HP0100
22	2	0.2732	721	98.4973	HP0775

23	1	0.1366	722	98.6339	HP0339
24	1	0.1366	723	98.7705	HP0962
25	1	0.1366	724	98.9071	HP0137
27	2	0.2732	726	99.1803	HP0705
31	1	0.1366	727	99.3169	HP1198
33	1	0.1366	728	99.4536	HP0879
43	1	0.1366	729	99.5902	HP0849
47	1	0.1366	730	99.7268	HP1559
51	1	0.1366	731	99.8634	HP0336
55	1	0.1366	732	100.0000	HP1409

Sum 724 100.0000

CUADRO 5. Frecuencias de Rain. Tamaño de la red $n = 732$.

Cluster	Freq	Freq %	CumFreq	CumFreq %	Representative
1	1699	51.8304	1699	51.8304	YAL011W
2	806	24.5882	2505	76.4185	YAL012W
3	350	10.6772	2855	87.0958	YAL054C
4	154	4.6980	3009	91.7938	YAL036C
5	84	2.5625	3093	94.3563	YAL041W
6	49	1.4948	3142	95.8511	YBL016W
7	31	0.9457	3173	96.7968	YAL032C
8	17	0.5186	3190	97.3154	YDL044C
9	11	0.3356	3201	97.6510	YCR086W
10	10	0.3051	3211	97.9561	YBR217W
11	9	0.2746	3220	98.2306	YBR101C
12	3	0.0915	3223	98.3221	YEL017W
13	2	0.0610	3225	98.3832	YGR113W
14	5	0.1525	3230	98.5357	YDL130W
15	4	0.1220	3234	98.6577	YIL034C
16	2	0.0610	3236	98.7187	YGL044C
17	2	0.0610	3238	98.7797	YBR270C
18	2	0.0610	3240	98.8408	YGR119C
19	2	0.0610	3242	98.9018	YDL116W
20	2	0.0610	3244	98.9628	YBR061C
21	1	0.0305	3245	98.9933	YIR038C
22	1	0.0305	3246	99.0238	YLR291C
23	4	0.1220	3250	99.1458	YDR259C
25	2	0.0610	3252	99.2068	YDL100C
27	1	0.0305	3253	99.2373	YLR347C
28	2	0.0610	3255	99.2984	YNL092W
30	2	0.0610	3257	99.3594	YDL239C

Cluster	Freq	Freq %	CumFreq	CumFreq %	Representative
31	1	0.0305	3258	99.3899	YLR373C
32	1	0.0305	3259	99.4204	YDL153C
34	2	0.0610	3261	99.4814	YDR318W
40	1	0.0305	3262	99.5119	YKL002W
44	1	0.0305	3263	99.5424	YGL070C
54	1	0.0305	3264	99.5729	YML064C
60	1	0.0305	3265	99.6034	YLR423C
63	1	0.0305	3266	99.6339	YDR034C
68	1	0.0305	3267	99.6644	YGL127C
78	1	0.0305	3268	99.6949	YLR288C
79	1	0.0305	3269	99.7254	YLR453C
88	1	0.0305	3270	99.7559	YLR447C
91	1	0.0305	3271	99.7865	YHR114W
95	1	0.0305	3272	99.8170	YER081W
98	1	0.0305	3273	99.8475	YER022W
99	1	0.0305	3274	99.8780	YPR086W
122	1	0.0305	3275	99.9085	YNL189W
124	1	0.0305	3276	99.9390	YLR295C
125	1	0.0305	3277	99.9695	YMR047C
285	1	0.0305	3278	100.0000	YJR091C

Sum 3278 100.0000

CUADRO 6. Frecuencias de Ito. Tamaño de la red $n = 3278$.

Cluster	Freq	Freq %	CumFreq	CumFreq %	Representative
1	172	35.9833	172	35.9833	YBL093C
2	105	21.9665	277	57.9498	YAL021C
3	79	16.5272	356	74.4770	YBR083W
4	28	5.8577	384	80.3347	YCR065W
5	12	2.5105	396	82.8452	YFL031W
6	18	3.7657	414	86.6109	YBR289W
7	8	1.6736	422	88.2845	YBL103C
8	2	0.4184	424	88.7029	YBR019C
9	8	1.6736	432	90.3766	YBL005W
10	6	1.2552	438	91.6318	YGL207W
11	4	0.8368	442	92.4686	YDR146C
12	4	0.8368	446	93.3054	YCL0C
13	7	1.4644	453	94.7699	YBR112C
14	3	0.6276	456	95.3975	YCL066W
15	2	0.4184	458	95.8159	YJR094C
17	4	0.8368	462	96.6527	YDL106C
18	3	0.6276	465	97.2803	YBR049C
19	2	0.4184	467	97.6987	YBL021C
20	2	0.4184	469	98.1172	YER040W
22	4	0.8368	473	98.9540	YHR119W
25	1	0.2092	474	99.1632	YCR084C
26	1	0.2092	475	99.3724	YGL209W
28	1	0.2092	476	99.5816	YGL035C
29	1	0.2092	477	99.7908	YMR043W
38	1	0.2092	478	100.0000	YDR207C
Sum	478	100.0000			

CUADRO 7. Frecuencias de Guelzim. Tamaño de la red $n = 478$.

Cluster	Freq	Freq %	CumFreq	CumFreq %	Representative
1	672	64.3678	672	64.3678	YAL016W
2	181	17.3372	853	81.7050	YAL032C
3	76	7.2797	929	88.9847	YAL034WA
4	45	4.3103	974	93.2950	YBL101WA
5	26	2.4904	1000	95.7854	YBR217W
6	13	1.2452	1013	97.0307	YAL036C
7	8	0.7663	1021	97.7969	YDL012C
8	5	0.4789	1026	98.2759	YGL115W
9	4	0.3831	1030	98.6590	YAL028W
10	3	0.2874	1033	98.9464	YCR086W
11	5	0.4789	1038	99.4253	YBL026W
12	1	0.0958	1039	99.5211	YER133W
13	1	0.0958	1040	99.6169	YJR022W
14	1	0.0958	1041	99.7126	YDR328C
15	1	0.0958	1042	99.8084	YGL212W
17	1	0.0958	1043	99.9042	YNL189W
24	1	0.0958	1044	100.0000	YML064C
Sum	1044	100.0000			

CUADRO 8. Frecuencias de Uetz. Tamaño de la red $n = 1044$.

Cluster	Freq	Freq %	CumFreq	CumFreq %	Representative
0	12	0.3717	12	0.3717	B0041.6
1	1663	51.5180	1675	51.8897	AC8.1
2	520	16.1090	2195	67.9988	B0024.1
3	270	8.3643	2465	76.3631	B0272.2
4	167	5.1735	2632	81.5366	AC7.2
5	116	3.5936	2748	85.1301	AC3.3
6	89	2.7571	2837	87.8872	B0393.1
7	69	2.1375	2906	90.0248	C05C8.6
8	52	1.6109	2958	91.6357	C03A7.14
9	42	1.3011	3000	92.9368	C04C3.3
10	31	0.9603	3031	93.8971	B0207.4
11	29	0.8984	3060	94.7955	C23G10.4
12	30	0.9294	3090	95.7249	B0507.1
13	18	0.5576	3108	96.2825	B0336.10
14	14	0.4337	3122	96.7162	C02F5.9
15	14	0.4337	3136	97.1499	C12D8.1
16	10	0.3098	3146	97.4597	C04F12.3
17	5	0.1549	3151	97.6146	C09D4.5
18	11	0.3408	3162	97.9554	B0547.1
19	10	0.3098	3172	98.2652	B0041.4
20	6	0.1859	3178	98.4511	C18C4.10
21	4	0.1239	3182	98.5750	C01B10.8
22	2	0.0620	3184	98.6369	R02F11.1
23	3	0.0929	3187	98.7299	B0218.3
24	2	0.0620	3189	98.7918	F54C9.5
25	2	0.0620	3191	98.8538	F10C5.1
27	2	0.0620	3193	98.9157	F08G2.5
28	2	0.0620	3195	98.9777	C18H9.7
29	3	0.0929	3198	99.0706	F35G12.9

Cluster	Freq	Freq %	CumFreq	CumFreq %	Representative
31	1	0.0310	3199	99.1016	K01G5.4
32	2	0.0620	3201	99.1636	R05F9.10
33	1	0.0310	3202	99.1945	R12B2.1
34	1	0.0310	3203	99.2255	T01D1.6
35	2	0.0620	3205	99.2875	F47G6.1
36	4	0.1239	3209	99.4114	B0205.3
40	1	0.0310	3210	99.4424	W09H1.6
42	1	0.0310	3211	99.4734	B0024.14
44	2	0.0620	3213	99.5353	C03A7.4
46	1	0.0310	3214	99.5663	R09B5.5
50	1	0.0310	3215	99.5973	W02G9.2
52	1	0.0310	3216	99.6283	ZK849.2
56	1	0.0310	3217	99.6592	F43C1.2
61	1	0.0310	3218	99.6902	C38D4.6
66	1	0.0310	3219	99.7212	F44G3.9
73	2	0.0620	3221	99.7831	W04D2.1
75	1	0.0310	3222	99.8141	C49A1.4
79	1	0.0310	3223	99.8451	T05C12.6
83	1	0.0310	3224	99.8761	K09B11.9
90	1	0.0310	3225	99.9071	T04H1.2
94	1	0.0310	3226	99.9380	T11B7.4
109	1	0.0310	3227	99.9690	T17H7.4
187	1	0.0310	3228	100.0000	W07B3.2
Sum	3216	100.0000			

CUADRO 9. Frecuencias de li-data. Tamaño de la red $n = 3228$.

Cluster	Freq	Freq %	CumFreq	CumFreq %	Representative
1	1390	57.9650	1390	57.9650	YMR037C
2	465	19.3912	1855	77.3561	YGR288W
3	210	8.7573	2065	86.1134	YPL230W
4	104	4.3369	2169	90.4504	YKL185W
5	51	2.1268	2220	92.5771	YBR240C
6	20	0.8340	2240	93.4112	YHL027W
7	30	1.2510	2270	94.6622	YDR216W
8	17	0.7089	2287	95.3711	YLR098C
9	7	0.2919	2294	95.6631	YDR123C
10	9	0.3753	2303	96.0384	YGL035C
11	11	0.4587	2314	96.4971	YCR018C
12	4	0.1668	2318	96.6639	YGL181W
13	2	0.0834	2320	96.7473	YNL314W
14	1	0.0417	2321	96.7890	YJL148W
15	4	0.1668	2325	96.9558	YDR277C
16	2	0.0834	2327	97.0392	YBL103C
17	1	0.0417	2328	97.0809	YMR042W
18	6	0.2502	2334	97.3311	YDR463W
19	4	0.1668	2338	97.4979	YGL237C
21	2	0.0834	2340	97.5813	YMR164C
23	4	0.1668	2344	97.7481	YBL021C
24	1	0.0417	2345	97.7898	YPL177C
26	1	0.0417	2346	97.8315	YFL044C
27	1	0.0417	2347	97.8732	YJL206C
28	1	0.0417	2348	97.9149	YLR176C
29	5	0.2085	2353	98.1234	YDL106C
31	2	0.0834	2355	98.2068	YBL008W
32	3	0.1251	2358	98.3319	YGL071W

Cluster	Freq	Freq %	CumFreq	CumFreq %	Representative
35	1	0.0417	2359	98.3736	YKL062W
36	2	0.0834	2361	98.4570	YMR021C
37	1	0.0417	2362	98.4987	YGL192W
38	1	0.0417	2363	98.5405	YPL049C
39	1	0.0417	2364	98.5822	YDR423C
40	3	0.1251	2367	98.7073	YIR023W
44	2	0.0834	2369	98.7907	YJR060W
47	1	0.0417	2370	98.8324	YML007W
56	2	0.0834	2372	98.9158	YGL073W
61	1	0.0417	2373	98.9575	YIL131C
62	2	0.0834	2375	99.0409	YDR310C
67	1	0.0417	2376	99.0826	YLR013W
68	2	0.0834	2378	99.1660	YBR182C
70	1	0.0417	2379	99.2077	YKL109W
75	1	0.0417	2380	99.2494	YLR131C
81	1	0.0417	2381	99.2911	YDR043C
82	1	0.0417	2382	99.3328	YEL009C
88	1	0.0417	2383	99.3745	YMR043W
89	1	0.0417	2384	99.4162	YDR259C
91	1	0.0417	2385	99.4579	YKL043W
96	1	0.0417	2386	99.4996	YLR182W
100	1	0.0417	2387	99.5413	YOR372C
102	1	0.0417	2388	99.5830	YDR146C
105	1	0.0417	2389	99.6247	YHR206W
106	1	0.0417	2390	99.6664	YIR018W
109	1	0.0417	2391	99.7081	YNL068C
112	1	0.0417	2392	99.7498	YDL056W
138	1	0.0417	2393	99.7915	YER111C

Cluster	Freq	Freq %	CumFreq	CumFreq %	Representative
147	1	0.0417	2394	99.8332	YOR028C
148	1	0.0417	2395	99.8749	YBR049C
195	1	0.0417	2396	99.9166	YPR104C
205	1	0.0417	2397	99.9583	YNL216W
257	1	0.0417	2398	100.0000	YKL112W

Sum 2398 100.0000

CUADRO 10. Frecuencias de Lee. Tamaño de la red $n = 2398$.

2. Apéndice B

Algoritmo 1 Prueba Chi-Cuadrado y Gráficos

```

■ function []=libre de escala;
■ frecuencias;
■ clases;
■     fprintf('*****Análisis para Rain*****');
■     ff=@versimar1;
■     Analisis_red(freq1,ff,clase1);
■     fprintf('*****Análisis para Guelzim*****');
■     ff=@versimar2;
■     Analisis_red(freq2,ff,clase2);
■     fprintf('*****Análisis para Uetz*****');
■     ff=@versimar3;
■     Analisis_red(freq3,ff,clase3);
■     fprintf('*****Análisis para Li-data*****');
■     ff=@versimar4;
■     Analisis_red(freq4,ff,clase4);
■     fprintf('*****Análisis para Lee*****');
■     ff=@versimar5;
■     Analisis_red(freq5,ff,clase5);
■     fprintf('*****Análisis para Ito*****');
■     ff=@versimar6;
■     Analisis_red(freq6,ff,clase6);

```

Algoritmo 2 Análisis para cada Red

```
▪ function []=Analisis_red(freq,ff,clase);  
▪ est=fminsearch(ff,1)  
▪ cc= 1;  
▪ for i = 1:length(freq)  
▪     if freq(i) = 0  
▪         freqq(cc)=freq(i);  
▪         cc=cc+1;  
▪     end  
▪ end  
▪ sum(freqq)  
▪ cum=0;
```

Funcion Zeta-Riemann

```
▪ for i=1:length(freqq)  
▪     cum = cum+ $i^{-est}$ ;  
▪     end
```

Calculo de la probabilidades

```
▪ for i=1:length(freqq)  
▪     x(i)=i;  
▪      $p(i)=\frac{freqq(i)}{n}$ ;  
▪      $pa(i)=i^{-est}/cum$ ;  
▪     end  
▪ figure  
▪ plot(log(x),log(p),'o',log(x),log(pa) )
```

Prueba Chi-Cuadrado

```

■ freqq ;
■ pa;
■ clase;
■ i=1;
■     while  $i < \text{length}(\text{clase})$ 
■          $w(i) = (\text{clase}(i) - n * pa(i))^2 / (n * pa(i));$ 
■          $i = i + 1;$ 
■     end
■ if  $i == \text{length}(\text{clase})$ 
■     prob = 0;
■     j = i;
■     while  $j \leq \text{length}(\text{freqq})$ 
■         prob = prob + pa(j);
■          $j = j + 1;$ 
■     end
■  $w(\text{length}(\text{clase})) = (\text{clase}(\text{length}(\text{clase})) - n * (\text{sum}(\text{prob})))^2 / (n * \text{sum}(\text{prob}));$ 
■ end
■ test = sum(w);
■ fprintf('El resultado de la prueba Chi-Cuadrado es:,test); p-valor
■ pvalor = 1 - chi2cdf(test,length(clase)-2)
■ fprintf('p-valor es:, pvalor);

```

Algoritmo 3 Función de verosimilitud para cada red

▪ **function** sal=versimar1(r)

▪

$freq = [281 \quad 140 \quad 92 \quad 51 \quad 25 \quad 26 \quad 22 \quad 8 \quad 11 \quad 9 \quad 9 \quad 6 \quad 6 \quad 9 \quad 4$
 $\quad \quad 5 \quad 2 \quad 1 \quad 2 \quad 0 \quad 2 \quad 2 \quad 1 \quad 1 \quad 10 \quad 2 \quad 0 \quad 0 \quad 0 \quad 1$
 $\quad \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0$
 $\quad \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1]$

▪ sal1=0;

▪ sal2=0;

▪ for i=1:length(freq)

▪ sal1=sal1+freq(i)*log(i);

▪ sal2=sal2+ exp(-r*log(i));

▪ sal=-(-r*sal1-sum(freq)*log(sal2));

▪ return

```
■ function sal=versimar2(r)
```

```
■
```

```
freq = [172  105  79  28  12  18  8  2  8  6  4  4  7  3  2  0  4  3
        2    2   0   4   0   0  1  1  0  1  1  0  0  0  0  0  0  0
        0    1
```

```
■ sal1=0;
```

```
■ sal2=0;
```

```
■ for i=1:length(freq)
```

```
■     sal1=sal1+freq(i)*log(i);
```

```
■     sal2=sal2+exp(-r*log(i));
```

```
■     sal=-(-r*sal1-sum(freq)*log(sal2));
```

```
■ return
```

```
■ function sal=versimar3(r)
```

```
■
```

```
freq = [672  181  76  45  26  13  8  5  4  3  5  1  1  1  1
        0    1   0   0   0   0  0  0  0  1]
```

```
■ sal1=0;
```

```
■ sal2=0;
```

```
■ for i=1:length(freq)
```

```
■     sal1=sal1+freq(i)*log(i);
```

```
■     sal2=sal2+exp(-r*log(i));
```

```
■     sal=-(-r*sal1-sum(freq)*log(sal2));
```

```
■ return
```

```
■ function sal=versimar4(r)
```

```
■
```

```
freq = [1663  520  270  167  116  89  69  52  42  31  29  30  18  14  14  10  5
        11   10   6   4   2   3   2   2   0   2   2   3   0   1   2   1   1
         2    4    0    0    0   1   0   1   0   2   0   1   0   0   0   1   0
         1    0    0    0    1   0   0   0   0   1   0   0   0   0   0   1   0
         0    0    0    0    2   0   1   0   0   0   1   0   0   0   0   1   0
         0    0    0    0    1   0   0   0   1   0   0   0   0   0   0   0   0
         0    0    0    0    0   0   1   0   0   0   0   0   0   0   0   0   0
         0    0    0    0    0   0   0   0   0   0   0   0   0   0   0   0   0
         0    0    0    0    0   0   0   0   0   0   0   0   0   0   0   0   0
         0    0    0    0    0   0   0   0   0   0   0   0   0   0   0   0   0
         0    0    0    0    0   0   0   0   0   0   0   0   0   0   0   0   1]
```

```
■ sal1=0;
```

```
■ sal2=0;
```

```
■ for i=1:length(freq)
```

```
■     sal1=sal1+freq(i)*log(i);
```

```
■     sal2=sal2+exp(-r*log(i));
```

```
■     sal=-(-r*sal1-sum(freq)*log(sal2));
```

```
■ return
```

```
■ function sal=versimar5(r)
```

```
■
```

```
freq = [1390  465  210  104  51  20  30  17  7  9  11  4  2  1  4  2  1  6  4  0  2
        0    4    1    0    1    1    1    5  0  2    3  0  0  1  2  1  1  1  3  0  0
        0    2    0    0    1    0    0    0  0  0    0  0  0  2  0  0  0  0  1  2  0
        0    0    0    1    2    0    1    0  0  0    0  1  0  0  0  0  0  1  1  0  0
        0    0    0    1    1    0    1    0  0  0    0  1  0  0  0  1  0  1  0  0  1
        1    0    0    1    0    0    1    0  0  0    0  0  0  0  0  0  0  0  0  0  0
        0    0    0    0    0    0    0    0  0  0    0  1  0  0  0  0  0  0  0  0  1
        1    0    0    0    0    0    0    0  0  0    0  0  0  0  0  0  0  0  0  0  0
        0    0    0    0    0    0    0    0  0  0    0  0  0  0  0  0  0  0  0  0  0
        0    0    0    0    0    1    0    0  0  0    0  0  0  0  0  1  0  0  0  0  0
        0    0    0    0    0    0    0    0  0  0    0  0  0  0  0  0  0  0  0  0  0
        0    0    0    0    0    0    0    0  0  0    0  0  0  0  0  0  0  0  0  0  0
        0    0    0    0    0  1]
```

```
■ sal1=0;
```

```
■ sal2=0;
```

```
■ for i=1:length(freq)
```

```
■     sal1=sal1+freq(i)*log(i);
```

```
■     sal2=sal2+exp(-r*log(i));
```

```
■
```

```
■ sal=-(-r*sal1-sum(freq)*log(sal2));
```

```
■ return
```

```
■ function sal=versimar6(r)
```

```
■
```

```
freq = [1699  806  350  154  84  49  31  17  11  10  9  3  2  5  4  2  2  2  2  2  1
        1   4   0   2   0   1   2   0   2   1  1  0  2  0  0  0  0  0  1  0  0
        0   1   0   0   0   0   0   0   0   0  0  1  0  0  0  0  0  1  0  0  1
        0   0   0   0   1   0   0   0   0   0  0  0  0  0  1  1  0  0  0  0  0
        0   0   0   1   0   0   1   0   0   0  1  0  0  1  1  0  0  0  0  0  0
        0   0   0   0   0   0   0   0   0   0  0  0  0  0  0  0  1  0  1  1  0
        0   0   0   0   0   0   0   0   0   0  0  0  0  0  0  0  0  0  0  0  0
        0   0   0   0   0   0   0   0   0   0  0  0  0  0  0  0  0  0  0  0  0
        0   0   0   0   0   0   0   0   0   0  0  0  0  0  0  0  0  0  0  0  0
        0   0   0   0   0   0   0   0   0   0  0  0  0  0  0  0  0  0  0  0  0
        0   0   0   0   0   0   0   0   0   0  0  0  0  0  0  0  0  0  0  0  0
        0   0   0   0   0   0   0   0   0   0  0  0  0  0  0  0  0  0  0  0  0
        0   0   0   0   0   0   0   0   0   0  0  0  0  0  0  0  0  0  0  0  0
        0   0   0   0   0   0   0   0   0   0  0  0  1]
```

```
■ sal1=0;
```

```
■ sal2=0;
```

```
■ for i=1:length(freq)
```

```
■     sal1=sal1+freq(i)*log(i);
```

```
■     sal2=sal2+exp(-r*log(i));
```

```
■
```

```
■ sal=-(-r*sal1-sum(freq)*log(sal2));
```

```
■ return
```
