USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Trabajo Fin de Grado

La distribución Gamma

Rita Troncoso de la Cuesta

2020/2021

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Trabajo Fin de Grado

La distribución Gamma

Rita Troncoso de la Cuesta

2020/2021

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación Operativa
Título: La distribución Gamma
Breve descripción del contenido
A pesar de no ser tan conocida como la Bernoulli o la normal, la distribución Gamma aparece en múltiples ocasiones como modelo para el cálculo de probabilidades y para el ajuste de datos reales. En la práctica, la distribución Gamma es el modelo de referencia para variables continuas y positivas, como pueden ser los flujos de agua, consumos de productos a granel, rentas, recogidas de residuos urbanos, y tantos otros. Sus dos parámetros, uno de forma y otro de escala, le dan una gran versatilidad, y de hecho contiene como casos particulares otros modelos de distribución tan famosos como la exponencial o la ji cuadrado. En este trabajo se revisarán las propiedades de la distribución Gamma, se estudiarán los métodos de inferencia para sus parámetros, y se considerarán modelos de regresión con respuesta de tipo Gamma. Todo ello se realizará apoyándose en ejemplos con datos reales o simulados, que permitirán entender la utilidad de la distribución Gamma y analizar las propiedades de los métodos considerados.

Índice general

Resumen

En este trabajo trataremos sobre la distribución Gamma. Esta es una distribución poco utilizada en la práctica porque presenta una mayor complejidad que otras distribuciones. Esta es una distribución biparamétrica y nos centraremos en obtener métodos sencillos para calcular las estimaciones de estos parámetros y sus intervalos de confianza. Asimismo, estudiaremos métodos para la validación de la distribución Gamma tanto gráfica como analíticamente. Finalmente, explicamos la regresión sobre una variable respuesta que sigue una distribución Gamma.

Abstract

In this projet we will talk about Gamma distribution. In practice, this distribution is not correntsly used because of its complexity. Gamma distribution has two parameters and we will focus on some simple methods in order to obtain estimators and their confidence intervals. In addition, we will explain goodness of fit test for Gamma. Finally, we will study the regression when the response variable is a Gamma.

Introducción

Este trabajo trata sobre la distribución Gamma. Aunque esta distribución es muy conocida no es la más habitual para trabajar. Sin embargo, esta distribución resalta por su importancia ya que algunas de las distribuciones más utilizadas en la práctica, en realidad, son un caso particular de la Gamma, como por ejemplo la exponencial o la χ^2.

Esta es una distribución estadística continua y biparamétrica (α, β) siendo α el parámetro de forma y β el parámetro de escala. La distribución Gamma se encarga de modelizar el comportamiento de variables aleatorias continuas con asimetría positiva. Esto quiere decir, que tiene una mayor densidad a la izquierda de la distribución respecto de la media. Veamos algunos casos donde esta distribución es aplicable:

- Tiempos de espera en la Teoría de Colas.
- En la meteorología para ajustar los datos de las precipitaciones.
- Análisis de supervivencia: el estudio estadístico del tiempo que pasa hasta que ocurre un evento.
- Ajusta la distribución de la renta.
- Modeliza los datos de consumo de productos a granel.

La distribución Gamma debe su nombre a la función Gamma. Ésta es muy conocida en diferentes ramas de las matemáticas desde la Teoría de Ecuaciones Diferenciales hasta la Estadística. Sin embargo, el origen de esta función se encuentra en la unión de un problema de interpolación con otro de cálculo integral. La función Gamma fue descubierta por Leonhard Euler en 1729 entre la correspondencia que mantenía con Goldbach.

En el *Capítulo 1* veremos la definición de la función y distribución Gamma y algunas propiedades interesantes sobre ellas. Al final de este capítulo estudiaremos dos conjuntos de datos y veremos si la distribución Gamma se ajusta adecuadamente a estos conjuntos de observaciones de manera gráfica.

En el *Capítulo 2* estudiaremos las estimaciones de los parámetros de forma y escala por diferentes métodos: el método de máxima verosimilitud y el método de los momentos. Veremos algún método iterativo para obtener estas estimaciones y para terminar este capítulo, proporcionaremos intervalos de confianza para las estimaciones de estos parámetros.

En el *Capítulo 3* estudiaremos el test Lilliefors. Ésta es una prueba que sirve para estudiar de forma analítica si un conjunto de datos sigue una distribución Gamma. Explicaremos brevemente este método y luego lo aplicaremos a los conjuntos de datos mencionados en el primer capítulo.

En el *Capítulo 4* explicaremos cómo realizar la regresión sobre la Gamma. Esta regresión es un caso particular de los modelos lineales generalizados, con lo cual, primeramente explicaremos este tipo de modelos y luego lo particularizaremos para la Gamma.

Capítulo 1

Definición y propiedades de la distribución Gamma

1.1. Función Gamma

En este apartado veremos la función Gamma, la cual da nombre a esta distribución.

Definición 1.1. *Se define la función Gamma como* $\Gamma : (0, \infty) \to \mathbb{R}$ *donde* $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ *para* $\alpha > 0$

Proposición 1.2. *Para cualquier* $\alpha > 1$ *tenemos que* $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$

Demostración. Partimos de la función Gamma dada por la integral anterior, es decir, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$. Ahora, resolvamos esta integral por partes tomando

$$u = x^{\alpha-1} \longrightarrow du = (\alpha - 1)x^{\alpha-2}dx$$
$$dv = e^{-x}dx \longrightarrow v = -e^{-x}$$

Con lo cual realizando estos cálculos obtenemos

$$\Gamma(\alpha) = \left|-x^{\alpha-1}e^{-x}\right|_0^\infty + \int_0^\infty (\alpha - 1)x^{\alpha-2}e^{-x}dx = (\alpha - 1)\int_0^\infty x^{\alpha-2}e^{-x}dx$$

Así, concluimos que $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ □

Corolario 1.3. *Para cualquier* $n \in \mathbb{N}$ *tenemos que* $\Gamma(n) = (n - 1)!$

Demostración. Si aplicamos reiteradamente la proposición anterior obtenemos que

$$\Gamma(\alpha) = (\alpha-1)(\alpha-2)\Gamma(\alpha-2) = \ldots = (\alpha-1)(\alpha-2)\ldots\Gamma(1)$$

Calculemos $\Gamma(1)$. Por definición

$$\Gamma(1) = \int_0^\infty e^{-x}dx = \lim_{M\to\infty}\int_0^M e^{-x}dx = \lim_{M\to\infty}\left|-e^{-x}\right|_0^M = \lim_{M\to\infty} -e^{-M}+e^0 = 1$$

Con lo cual, para $n \in \mathbb{N}$ se tiene que $\Gamma(n) = (n-1)!\Gamma(1) = (n-1)!$ □

Proposición 1.4. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Demostración. Tenemos que introducir el concepto de la función β para realizar esta demostración.

Tenemos $\beta(p,q) = \displaystyle\int_0^1 x^{p-1}(1-x)^{q-1}dx$ para p,q>0. Para resolver esta integral realizamos un cambio de variable de la siguiente forma

$$x = \sin^2 t \longrightarrow dx = 2\sin t\cos t dt$$
$$1-x = 1-\sin^2 t = \cos^2 t$$

Además, los límites de integración quedarían de la siguiente manera

$$x = \sin^2 t \longrightarrow t = \arcsin\sqrt{x} \longrightarrow \begin{cases} x = 0 \to t = 0 \\ x = 1 \to t = \frac{\pi}{2} \end{cases}$$

Realizando este cambio obtenemos

$$\beta(p,q) = \int_0^{\frac{\pi}{2}} (\sin^2 t)^{p-1}(\cos^2 t)^{q-1}2\sin t\cos t dt$$

Con lo cual obtenemos que la función β es

$$\beta(p,q) = 2\int_0^{\frac{\pi}{2}} (\sin t)^{2p-1}(\cos t)^{2q-1}dt$$

A continuación, calculemos $\beta\left(\frac{1}{2},\frac{1}{2}\right)$. Por un lado, calculamos la función beta por definición con esos valores y obtenemos

$$\beta\left(\frac{1}{2},\frac{1}{2}\right) = 2\int_0^{\frac{\pi}{2}} 1dt = 2\left(\frac{\pi}{2}\right) = \pi$$

Y además, por una propiedad de esta función sabemos que $\beta(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ y que $\Gamma(1) = 1$. Con lo cual

$$\beta\left(\frac{1}{2},\frac{1}{2}\right) = \frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(1)} = \Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right) = \pi$$

Así ya queda demostrado que $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ □

1.2. Función de densidad

En esta sección trataremos sobre la función de densidad de la distribución Gamma y sus parámetros α y β.

Definición 1.5. *Decimos que X es una variable aleatoria que sigue una distribución Gamma de parámetros $\alpha > 0$ y $\beta > 0$, si su función de densidad es*

$$f_X(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & \text{otro caso} \end{cases} \tag{1.1}$$

Ahora veamos el papel que desempeñan los parámetros α y β en esta distribución.

Comencemos analizando la influencia del parámetro α. Este se conoce como el "parámetro de forma". Cuando toma valores pequeños, $\alpha \leq 1$, podemos observar que $f_X(x)$ es decreciente en todo su soporte. Por otro lado, para $\alpha > 1$ el máximo valor de la función de densidad se consigue en $x = (\alpha - 1)\beta$. Conseguimos este máximo derivando $f_X(x)$ e igualándolo a 0. El valor máximo de la función de densidad es $\frac{(\alpha-1)^{\alpha-1}e^{-(\alpha-1)}}{\beta\Gamma(\alpha)}$.

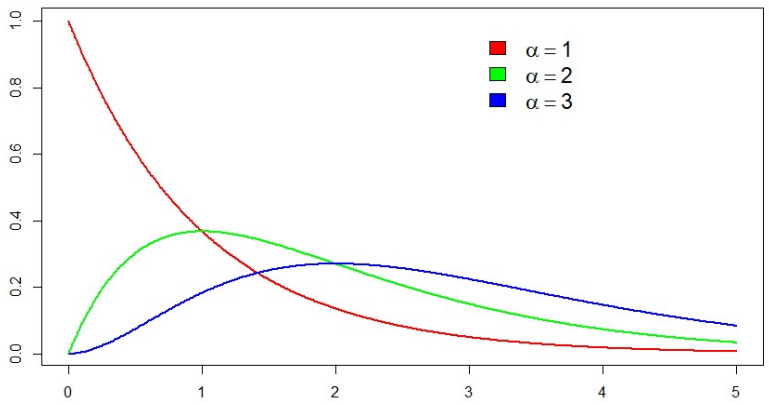

Figura 1.1: Funciones de densidad de la Gamma con $\beta = 1$ y distintos valores de α

Como podemos observar en la figura (1.1), cuando $\alpha = 1$ es una exponencial y a medida que aumentamos este parámetro el momento de máxima intensidad de probabilidad; es decir, la moda, se sitúa más hacia el centro consiguiendo una mayor simetría y que aparezca la campana de Gauss.

Ahora, veamos el parámetro β. Este es conocido como el "parámetro de escala", pues al multiplicar la variable X con distribución Gamma por una constante, la variable transformada sigue teniendo distribución Gamma con el mismo parámetro de forma y con parámetro de escala β multiplicado por esa constante. Demostremos lo que acabamos de afirmar.

Proposición 1.6. *Dada X una variable aleatoria tal que $X \sim Ga(\alpha, \beta)$ entonces $aX \sim Ga(\alpha, a\beta)$ siendo $a > 0$ un número real positivo.*

Demostración. Sea $a > 0$, F_X es la función de distribución de la variable X y F_{aX} es la función de distribución de aX. De esta forma, tenemos que

$$F_{aX}(x) = P(aX \leq x) = P\left(X \leq \frac{x}{a}\right) = F_X\left(\frac{x}{a}\right)$$

Entonces,

$$f_{aX} = \frac{d}{dx}F_{aX}(x) = \frac{d}{dx}F_X\left(\frac{x}{a}\right) = \frac{1}{a}f_X\left(\frac{x}{a}\right) = \frac{1}{(a\beta)^\alpha \Gamma(\alpha)}x^{\alpha-1}e^{-x/a\beta}$$

De esta forma, hemos llegado a la conclusión de que la función de densidad de la variable aX es la función de densidad de una Gamma de parámetros α y $a\beta$.

□

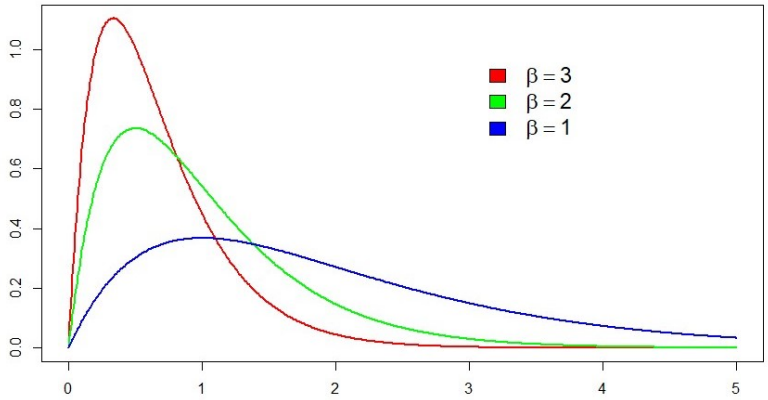

Figura 1.2: Funciones de densidad de la Gamma con $\alpha = 2$ y distintos valores de β

1.2.1. Función de distribución

La función de distribución indica la probabilidad de que la variable aleatoria tenga un valor menor o igual que un cierto x. Por ello, también es conocida como función de probabilidad acumulada. A continuación, veremos la función de probabilidad de la distribución Gamma que conseguimos integrando por partes (1.1).

Definición 1.7. *Sea $X \backsim Ga(\alpha, \beta)$ la función de distribución asociada a X es*

$$F_X(t) = P(X \leq t) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^t x^{\alpha-1} e^{-\frac{x}{\beta}} dx$$

1.3. Esperanza y varianza de la Gamma

Trataremos ahora sobre la esperanza y la varianza de la Gamma.

Proposición 1.8. *Sea X una variable aleatoria tal que $X \sim Ga(\alpha, \beta)$ entonces $E(X) = \alpha\beta$.*

Demostración.

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^{\infty} x^\alpha e^{-\frac{x}{\beta}} dx \quad (1.2)$$

Hagamos un cambio de variable, de manera que

$$u = \frac{x}{\beta} \longrightarrow x = \beta u$$
$$du = \frac{dx}{\beta} \longrightarrow dx = \beta du$$

Así, realizando este cambio en (1.2) tenemos que

$$E(X) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} (u\beta)^{\alpha} e^{-u} \beta du = \frac{\beta}{\Gamma(\alpha)} \int_0^{\infty} u^{\alpha} e^{-u} du = \frac{\beta\Gamma(\alpha+1)}{\Gamma(\alpha)} = \frac{\beta\alpha\Gamma(\alpha)}{\Gamma(\alpha)}$$

De esta forma concluimos que

$$E(X) = \alpha\beta$$

□

Proposición 1.9. *Sea X una variable aleatoria tal que $X \sim Ga(\alpha, \beta)$ entonces $Var(X) = \alpha\beta^2$.*

Demostración. Sabemos que $Var(X) = E(X^2) - E(X)^2$, con lo cual calculemos $E(X^2)$.

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^{\infty} x^2 \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} dx = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} x^{\alpha+1} e^{-\frac{x}{\beta}} dx \tag{1.3}$$

Hagamos un cambio de variable, de la forma

$$u = \frac{x}{\beta} \longrightarrow x = \beta u$$
$$du = \frac{dx}{\beta} \longrightarrow dx = \beta du$$

Así, realizando este cambio en (1.3) tenemos que

$$E(X^2) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} (u\beta)^{\alpha+1} e^{-u} \beta du = \frac{\beta^2}{\Gamma(\alpha)} \int_0^{\infty} u^{\alpha+1} e^{-u} du = \frac{\beta^2\alpha(\alpha+1)\Gamma(\alpha)}{\Gamma(\alpha)}$$

Entonces obtenemos que $E(X^2) = (\alpha+1)\alpha\beta^2$. Con lo cual, como $Var(X) = E(X^2) - E(X)^2$ podemos concluir que

$$Var(X) = (\alpha+1)\alpha\beta^2 - (\alpha\beta)^2 = \alpha\beta^2$$

□

1.4. Función generatriz de momentos de la Gamma

En esta sección calcularemos la función generatriz de momentos.

Sea X una variable aleatoria tal que $X \sim Ga(\alpha, \beta)$. Consideramos $t \in \mathbb{R}$

$$m_X(t) = E[e^{tX}] = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty e^{tx} x^{\alpha-1} e^{-\frac{x}{\beta}} dx = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{(t-\frac{1}{\beta})x} dx \quad (1.4)$$

Como podemos observar, la integral (1.4) vale infinito si $t - \frac{1}{\beta} \geq 0$, o equivalentemente, si $t \in \left[\frac{1}{\beta}, +\infty\right)$. Suponiendo que $t - \frac{1}{\beta} < 0$, calcularemos la integral haciendo el cambio de variable

$$u = (\tfrac{1}{\beta} - t)x \longrightarrow x = \tfrac{1}{\frac{1}{\beta}-t} u$$
$$du = (\tfrac{1}{\beta} - t)dx \longrightarrow dx = \tfrac{1}{\frac{1}{\beta}-t} du$$

Con lo cual, obtenemos que

$$m_X(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty \frac{\beta^{\alpha-1} u^{\alpha-1}}{(1-t\beta)^{\alpha-1}} e^{-u} \left(\frac{\beta}{1-t\beta}\right) du = \frac{1}{\Gamma(\alpha)(1-t\beta)^\alpha} \int_0^\infty u^{\alpha-1} e^{-u} du$$

Es decir, tenemos que la función generatriz de momentos es

$$m_X(t) = \frac{1}{(1-t\beta)^\alpha}$$

y está bien definida para todo $t \in \left(0, \frac{1}{\beta}\right)$.

1.5. Reproductividad de la Gamma

Proposición 1.10. *Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes tales que $X_i \sim Ga(\alpha_i, \beta)$ siendo $i = 1, \ldots, n$. Entonces,*

$$Y = \sum_{i=1}^n X_i \sim Ga\left(\sum_{i=1}^n \alpha_i, \beta\right)$$

Demostración. La haremos empleando la función generatriz de momentos.

$$m_Y(t) = E\left[e^{tY}\right] = E\left[e^{t(X_1+X_2+\ldots+X_n)}\right] = E\left[e^{tX_1}e^{tX_2}\ldots e^{tX_n}\right] =$$
$$= E\left[e^{tX_1}\right]E\left[e^{tX_2}\right]\ldots E\left[e^{tX_n}\right] = m_{X_1}(t)m_{X_2}(t)\ldots m_{X_n}(t) =$$
$$\frac{1}{(1-t\beta)^{\alpha_1}}\frac{1}{(1-t\beta)^{\alpha_2}}\cdots\frac{1}{(1-t\beta)^{\alpha_n}} = \frac{1}{(1-t\beta)^{\sum_{i=1}^n \alpha_i}} \text{ para } t < \frac{1}{\beta}$$

Como podemos ver, ésta es la función generatriz de momentos de $Ga(\sum_{i=1}^n \alpha_i, \beta)$ y por el Teorema de unicidad de la función generatriz de momentos tenemos que

$$Y = \sum_{i=1}^n X_i \sim Ga\left(\sum_{i=1}^n \alpha_i, \beta\right)$$

□

1.6. Relación con otras distribuciones

En esta sección vamos a estudiar la relación de la distribución Gamma con otras distribuciones. En particular, trataremos con la exponencial y la Erlang, que son interesantes cuando estudiamos tiempos de espera o tiempos en los que ocurre un determinado número de sucesos. Además, mencionamos la distribución χ^2 debido a su gran importancia en la Inferencia estadística.

1.6.1. Distribución exponencial

La distribución exponencial es un caso particular de una distribución Gamma con $\alpha = 1$, es decir, si $X \sim Exp(\beta)$ entonces $X \sim Ga(1, \beta)$. Esta distribución representa el tiempo de espera hasta que ocurre cierto evento.

A pesar de ser un caso particular de la Gamma, la distribución exponencial tiene una gran importancia por sí misma. Se usa para ver el tiempo de espera entre dos sucesos o para modelar los tiempos de supervivencia, por ejemplo, cuánto tarda en estropearse una bombilla, una farola, etc.

Definición 1.11. *Sea X una variable aleatoria, decimos que sigue una distribución exponencial si su función de densidad es de la forma*

$$f_X(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} \quad \text{con } x \in (0, +\infty)$$

Observación 1.12. *Es muy común expresar la función de densidad en función del parámetro $\lambda = \frac{1}{\beta}$. Así, tenemos que la función de densidad de la exponencial la podemos parametrizar como $f_X(x) = \lambda e^{-x\lambda}$, $x \in (0, +\infty)$.*

A continuación, veremos una propiedad interesante de la distribución exponencial.

Proposición 1.13. *La suma de exponenciales independientes es una distribución Gamma.*

Demostración. Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes tales que $X_i \sim Exp(\beta) = Ga(1, \beta)$ para todo $i = 1, \ldots, n$ con lo cual, por la reproductividad de la Gamma, tenemos que $\sum_{i=1}^{n} X_i \sim Ga(\sum_{i=1}^{n} 1, \beta) = Ga(n, \beta)$. □

1.6.2. Distribución de Erlang

La distribución de Erlang es otro caso particular de una distribución Gamma siendo el parámetro α un valor entero y positivo. Esta distribución representa el tiempo de espera hasta que ocurre el α-ésimo evento.

Definición 1.14. *Sea X una variable aleatoria, decimos que sigue una distribución de Erlang si su función de densidad es de la forma*

$$f_X(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & \text{otro caso} \end{cases} \tag{1.5}$$

siendo $\beta > 0$ y α un número natural.

Como veremos en la sección siguiente, tanto la distribución exponencial como la Erlang, juegan un papel importante para establecer la relación de la distribución Gamma con la distribución de Poisson. Por ello, igual que en la exponencial, es más común encontrarse esta distribución parametrizada en términos de λ, con $\lambda = \frac{1}{\beta}$.

$$f_X(x) = \begin{cases} \frac{\lambda}{(\alpha-1)!} (\lambda x)^{\alpha-1} e^{-\lambda x} & x > 0 \\ 0 & \text{otro caso} \end{cases}$$

donde $\lambda > 0$ y α es un número natural.

Además, también podemos establecer una relación clara entre la distribución Erlang y la exponencial ya que por la proposición (1.13) tenemos que la distribución Erlang es la suma de exponenciales.

1.6.3. Distribución χ^2

La distribución χ^2 también es un caso específico de la distribución Gamma tomando como $\alpha = \frac{n}{2}$ y $\beta = 2$. Con esos valores de los parámetros tenemos la distribución χ^2 con n grados de libertad, es decir, si $X \sim Ga\left(\frac{n}{2}, 2\right)$ entonces $X \sim \chi^2(n)$.

Esta distribución es muy importante para la Inferencia estadística ya que la prueba χ^2 se usa para la estimación de varianzas, como prueba de buen ajuste o como prueba de independencia, entre otras aplicaciones.

Definición 1.15. *Una variable aleatoria X sigue una distribución χ^2 con n grados de libertad si su función de densidad es de la siguiente forma*

$$f_X(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n-2}{2}} e^{-\frac{x}{2}} & x \geq 0 \\ 0 & \text{otro caso} \end{cases}$$

1.7. Relación con el proceso de Poisson

La distribución Gamma y la de Poisson están muy relacionadas entre sí. Esta relación la podemos establecer a partir de la conexión existente entre la Poisson, la Erlang y la exponencial.

En primer lugar, recordemos cómo es la distribución de Poisson.

Definición 1.16. *Sea X una variable aleatoria, decimos que sigue una distribución de Poisson de parámetro $\lambda \in (0, +\infty)$ si su función de probabilidad es de la siguiente forma*

$$p(x, \lambda) = P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{con } x \in \{0, 1, 2, 3, \ldots\}$$

La distribución de Poisson se encarga de ver cuántas veces ha ocurrido un suceso en un determinado intervalo de tiempo, donde el parámetro λ representa el número de veces que se espera que suceda el evento en dicho intervalo.

Con lo cual, la función de distribución de Poisson es

$$P[X \leq x] = \sum_{k=0}^{x} \frac{1}{k!} \left(\frac{1}{\beta}\right)^{k} e^{-\frac{1}{\beta}} \tag{1.6}$$

Seguidamente, definamos el proceso de Poisson.

Definición 1.17. *Un proceso de Poisson es la aparición aleatoria de sucesos a lo largo de un tiempo cumpliendo:*

- *El número de sucesos que ocurren en intervalos de tiempo disjuntos son variables aleatorias independientes.*
- *El número medio de sucesos por unidad de tiempo, λ, se mantiene constante a lo largo del tiempo.*
- *En un intervalo de tiempo de longitud diferencial $[t, \Delta + t]$ sólo se puede producir a lo sumo un suceso.*

La relación entre la distribución de Erlang y la de Poisson es de la siguiente manera.

Proposición 1.18. *El número de veces que ocurre un evento en un intervalo de tiempo $(0, t)$ sigue una distribución de Poisson de parámetro λt con $\lambda \in (0, +\infty)$ si, y solo si, el tiempo que transcurre hasta el n-ésimo evento sigue una distribución de Erlang de parámetros n y λ con $\lambda \in (0, +\infty)$.*

Demostración. Sea X_t la variable que denota el número de veces que ocurre un evento en un intervalo de tiempo $(0, t)$, y sea T_n la que contabiliza el tiempo que transcurre hasta el $n-$ésimo evento. A continuación, relacionaremos estas dos variables.

Como podemos razonar, si el tiempo hasta que ocurre el n-ésimo evento es menor o igual que $t > 0$ entonces el número de eventos que ocurren en el intervalo $(0, t)$ es mayor o igual que n. Entonces,

$$F_{T_n}(t) = P(T_n \leq t) = P(X_t \geq n) = 1 - P(X_t \leq n-1) = 1 - F_{X_t}(n-1) \tag{1.7}$$

A partir de la igualdad anterior (1.7) demostraremos la proposición.

$\Longrightarrow$ Supongamos que $X_t \sim P(\lambda t)$. A partir de la expresión (1.7) y la función de distribución de Poisson (1.6) podemos obtener la función de distribución de la variable T_n de la siguiente forma

$$F_{T_n}(t) = 1 - F_{X_t}(n-1) = 1 - e^{-\lambda t} \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} \tag{1.8}$$

Si derivamos la expresión (1.8), respecto de t, obtenemos la función de densidad de T_n

$$f_{T_n}(t) = \lambda e^{-\lambda t} \sum_{k=0}^{n-1} \left[\frac{(\lambda t)^k}{k!} - \frac{k(\lambda t)^{k-1}}{k!} \right] = \lambda e^{-\lambda t} \left[\sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} - \sum_{k=1}^{n-1} \frac{(\lambda t)^{k-1}}{(k-1)!} \right]$$

Ahora, hagamos un cambio de variable $h = k - 1$ en la segunda suma

$$f_{T_n}(t) = \lambda e^{-\lambda t} \left[\sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} - \sum_{h=0}^{n-2} \frac{(\lambda t)^h}{h!} \right] = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t}, \quad t > 0 \tag{1.9}$$

Como podemos observar, la expresión dada en (1.9) es la función de densidad de la distribución de Gamma. Con lo cual, deducimos que la variable T_n se distribuye como una Gamma de parámetros $n \in \mathbb{N}$ y $\lambda > 0$. Concluimos que T_n sigue una distribución de Erlang de parámetros n y λ.

$\Longleftarrow$ Sea $T_n \sim Erlang(n, \lambda)$. Razonemos de la misma manera que en la implicación anterior. Partimos de (1.7) y usamos la función de densidad de la distribución de Erlang (1.5). De esta manera obtenemos

$$F_{X_t}(n-1) = 1 - F_{T_n}(t) = \int_t^{+\infty} \frac{\lambda^n}{(n-1)!} s^{n-1} e^{-\lambda s} ds$$

Realizamos el cambio de variable $v = \lambda s \longrightarrow dv = \lambda ds$, con lo cual

$$F_{X_t}(n-1) = \int_{\lambda t}^{+\infty} \frac{v^{n-1}}{(n-1)!} e^{-v} dv, \quad \text{para todo } n \in \mathbb{N} \tag{1.10}$$

Como $n \in \mathbb{N}$, la integral anterior (1.10) puede expresarse en términos de una

suma sin más que ir integrando por partes sucesivamente. Si denotamos

$$I_{n,b} = \int_b^{+\infty} y^n e^{-y} dy \ \text{ con } \ b > 0. \tag{1.11}$$

Resolvemos la integral anterior (1.11) integrando por partes sucesivamente de la siguiente manera

$$\begin{cases} u = y^n \longrightarrow du = ny^{n-1}dy \\ dv = e^{-y}dy \longrightarrow v = -e^{-y} \end{cases}$$

Así, obtenemos

$$I_{n,b} = \left|-y^n e^{-y}\right|_b^{+\infty} + nI_{n-1,b} = b^n e^{-b} + nI_{n-1,b} =$$
$$= b^n e^{-b} + nb^{n-1}e^{-b} + n(n-1)I_{n-2,b} = \ldots = b^n e^{-b} + nb^{n-1}e^{-b} + n(n-1)b^{n-2}e^{-b} + \ldots + n!e^{-b}$$
$$= e^{-b}\left[b^n + nb^{n-1} + n(n-1)b^{n-2} + \ldots + n!\right] = e^{-b}\sum_{k=0}^{n} b^k \frac{n!}{k!}$$

Por lo tanto, tenemos que

$$F_{X_t}(n-1) = \int_{\lambda t}^{+\infty} \frac{v^{n-1}e^{-v}}{(n-1)!}dv = \frac{I_{n-1,\lambda t}}{(n-1)!} = e^{-\lambda t}\sum_{k=0}^{n-1}\frac{(\lambda t)^k}{k!}, \quad \text{para todo } \ n \in \mathbb{N}$$

De esta forma, podemos concluir que $X_t \sim P(\lambda t)$

□

La relación de la distribución de Poisson con la exponencial es muy sencilla ya que es un caso particular del anterior, tomando como parámetro de forma $n = 1$.

Veamos a continuación un ejemplo de la relación de la distribución Gamma con el proceso de Poisson.

Ejemplo 1.19. *A una centralita de teléfonos llegan 12 llamadas por minuto, siguiendo una distribución de Poisson. ¿Cuál es la probabilidad de que en menos de 1 minuto lleguen 8 llamadas?*

Tenemos que $\lambda = 12$, con lo cual $\beta = \frac{1}{12}$. Calculamos la probabilidad de que lleguen 8 llamadas en menos de 1 minuto. Tenemos que $\alpha = 8$ y $\beta = \frac{1}{12}$ como parámetros de la variable X que mide el tiempo transcurrido hasta que llega la octava llamada.

$$P(X < 1) = \frac{12^8}{7!} \int_0^1 x^7 e^{-12x} dx = 0.9104955$$

Podemos resolver esta problema integrando o usando R, "pgamma(1, 8, scale = 1/12, lower.tail = T)".

1.8. Ejemplos con datos reales

En esta sección veremos dos ejemplos de cantidades de lluvia recogida. En primer lugar, estudiaremos un conjunto de datos de lluvias diarias en la provincia de A Coruña y posteriormente trabajaremos con lluvias anuales en el estado de Nueva York. El objetivo de este apartado es, a partir de los datos empleados, ver qué distribución es la más adecuada, y en concreto si podría ser la distribución Gamma, que es un buen candidato por ser la cantidad de lluvia una variable continua y positiva.

Ejemplo 1.20. *Se han recogido los datos de la cantidad de precipitaciones diarias medidas en l/m^2 que han ocurrido a lo largo de los meses de enero y febrero de 2021 en la provincia de A Coruña. Estos datos han sido obtenidos a partir de Aemet, una base de datos meteorológica. En el apéndice, Comandos R: Capítulo 1, adjuntamos una tabla con los datos empleados, así como los comandos utilizados para obtener las gráficas que se muestran más adelante, ver [4.4].*

Para analizar qué distribuciones son buenas candidatas para ajustar estos datos usaremos el programa RStudio, en particular, empleando la librería fitdistrplus. Ésta es una librería que permite ajustar datos a distribuciones univariadas y comparar diferentes ajustes entre sí. En primer lugar, para escoger las distribuciones que son buenas candidatas para aproximar los datos usaremos el comando decdist. Éste nos proporciona un gráfico de asimetría y curtosis, como el propuesto por Cullen y Frey.

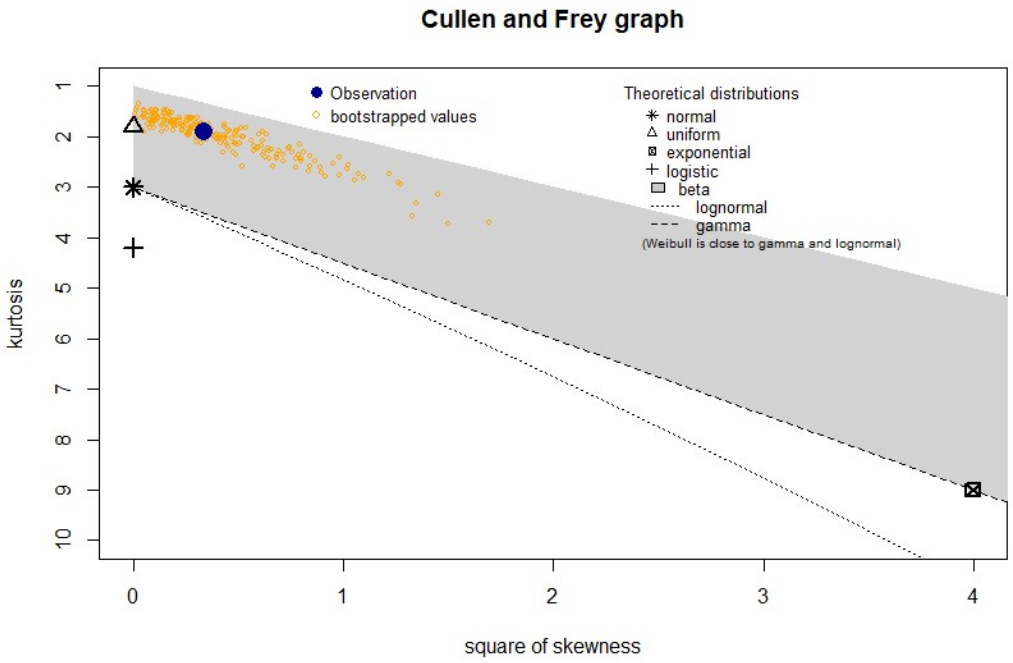

Figura 1.3: Gráfico Cullen and Frey para los datos del Ejemplo 1.20

Expliquemos el siguiente gráfico detalladamente.

Para obtener la figura (1.3) realizamos el comando "descdist(x, boots=200)", siendo x el conjunto de datos que queremos estudiar y n su tamaño muestral. A partir de este conjunto de observaciones, este comando genera con bootstrapped el número de muestras que le indiquemos, que en nuestro caso serán 200, con el mismo tamaño muestral que el conjunto inicial de datos. De estas muestras determinamos su curtosis y asimetría y las ubicamos en el gráfico. La asimetría y curtosis de estas muestras son los puntos amarillos que observamos en la figura. El punto azul es la ubicación de los datos iniciales, los que hemos denotado como x, según sus valores de asimetría y curtosis. Además, también se muestran estos valores para las distribuciones conocidas (normal, uniforme, exponencial, logística, beta, lognormal, Gamma). Hay algunas distribuciones cuyos valores de asimetría y curtosis son fijos, es decir, aunque su parámetro varíe estos valores no cambian. Son por ejemplo la normal, la uniforme, la logística y la exponencial. Sin embargo, para otras distribuciones, sí que varían estos valores en función de los parámetros. Este comando tiene en cuenta los posibles valores de asimetría y curtosis para todos los parámetros de la distribución. La distribución beta se representa con un área de posibles valores, ya que los coeficientes de asimetría y curtosis se modifican cuando varía cualquiera de

sus parámetros. En cambio, tanto la Gamma como la lognormal se representan con una línea ya que estos valores solo cambian cuando varía un parámetro, en particular, la Gamma varía cuando se modifica el parámetro de escala y la lognormal con la desviación típica.

De esta manera, podemos deducir que la que mejor aproxima estos datos es la distribución beta (la franja gris en el gráfico). Sin embargo, no es posible aproximar las observaciones con esta distribución ya que los datos no se encuentran entre 0 y 1. Así, entre las distribuciones que están más cercanas veamos cuál es la más adecuada. En este caso, estudiaremos la Gamma (la línea discontinua más cercana a la distribución β), la normal (el asterisco) y la uniforme (el triángulo).

A continuación, estudiaremos gráficamente las tres posibles distribuciones. Para ello usaremos el comando fitdist de la librería fitdistrplus que nos devuelve un objeto de clase "fitdist", es decir, una lista de diferentes componentes. Algunos de ellos son: estimadores de los parámetros, el método usado para estimarlos, el error estándar estimado, el criterio de información de Akaike o el criterio de información bayesiano. Para obtener las siguientes representaciones realizamos plot sobre el fitdistr. Así, para cada distribución producimos cuatro gráficos.

Densidad teórica y empírica : compara la densidad de los datos y la línea roja que es la densidad teórica de la distribución estudiada.

QQ-Plot : la línea continua son los cuantiles teóricos de la distribución y los puntos los cuantiles empíricos de los datos. Se puede observar la desviación que presentan los diferentes cuantiles empíricos respecto de los teóricos.

CDF teórico y empírico: compara la función de probabilidad empírica con la función de probabilidad teórica de la distribución.

PP-Plot: compara las probabilidades teóricas y las empíricas de los datos. Al igual que en el QQ-Plot también es interesante estudiar la desviación de las probabilidades empíricas (puntos) respecto de las teóricas (línea roja).

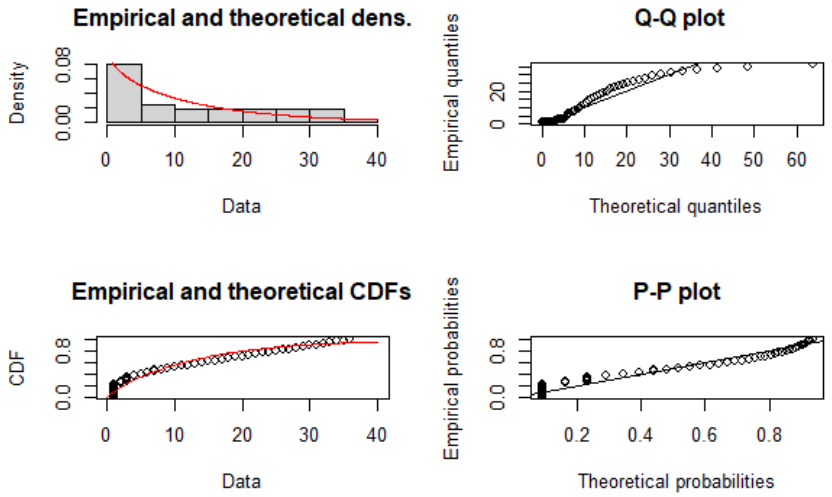

Figura 1.4: Aproximación de la distribución Gamma de los datos del Ejemplo 1.20

La distribución Gamma puede ser una buena aproximación ya que la función de densidad empírica y teórica son semejantes. Además, en la gráfica QQ-Plot vemos que los cuantiles de los datos presentan poca desviación con respecto a los cuantiles teóricos de la Gamma. Asimismo, tanto la gráfica CDF como la PP-Plot también son muy aceptables, se aproximan las probabilidades teóricas y empíricas.

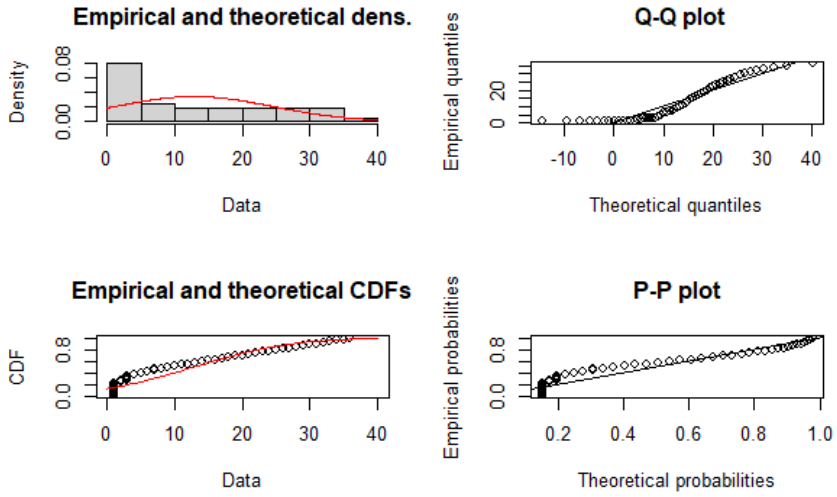

Figura 1.5: Aproximación de la distribución normal para los datos del Ejemplo 1.20

Por otro lado, en la distribución normal ocurre algo parecido que en la distribución anterior, aunque en este caso la densidad de los datos y la densidad teórica son ligeramente diferentes. Además, en la gráfica CDF y PP-Plot las probabilidades empíricas y teóricas se observa una mayor desviación que en el caso anterior. Sin embargo, aún no hay indicios para poder rechazar la hipótesis de que los datos sigan una distribución normal.

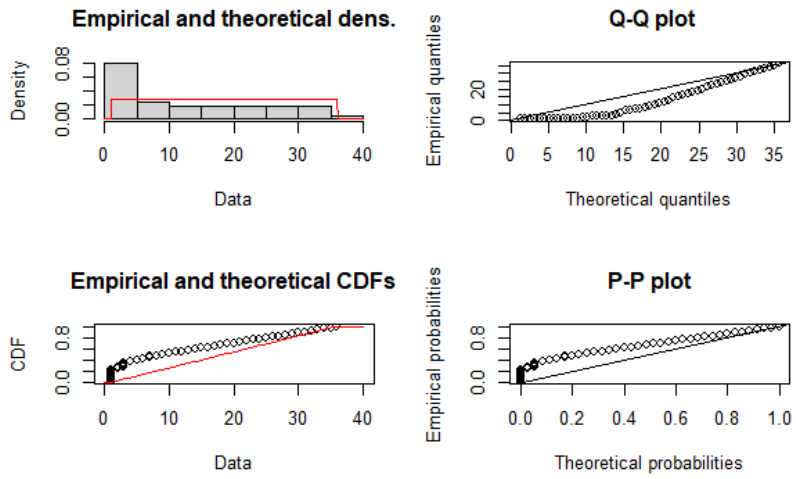

Figura 1.6: Aproximación de la distribución uniforme para los datos del Ejemplo 1.20

Para terminar, para la distribución uniforme vemos que tanto la densidad como los cuantiles de los datos se desvían mucho de los teóricos, al igual que las probabilidades, lo que nos hace pensar que esta distribución no sirve para trabajar con estas observaciones.

Como podemos observar, tenemos que tanto la distribución Gamma como la normal parecen aceptables.

Es importante destacar que estos datos son de lluvias diarias, con lo cual podrían presentar autocorrelación. Para ello, hagamos un test de Ljung-Box para estudiar esta correlación de los datos. Esto lo haremos a partir del comando "Box.test" de *R* cuya hipótesis nula es la independencia de los datos, mientras que la alternativa es la dependencia. Haciendo esto, obtenemos un p-valor de 0.085. Con lo cual, rechazamos la independencia de los datos con un nivel de significación del 10 %, pero no hay

motivos para rechazar la hipótesis nula a un nivel del 5 %. La correlación a un nivel de significación del 10 % podría afectar en los criterios de bondad de ajuste, como veremos en el Capítulo 3.

A continuación, presentaremos otro ejemplo que es de lluvias anuales.

Ejemplo 1.21. *Se han recogido los datos de lluvias anuales que sucedieron en Ithaca, una ciudad del estado de Nueva York, recogidas cada enero desde 1933 hasta 1982. Estos datos se pueden encontrar en el libro de Wilks [8]. Además, adjuntamos la tabla de datos en el apéndice, Comandos R: Capítulo 1, ver [4.4].*

A diferencia de los datos anteriores que mostraban autocorrelación a un significación del 10 % , éstos no presentan autocorrelación. Podemos afirmar que no tienen autocorrelación porque analizando el test de Ljung-Box para este conjunto de observaciones nos proporciona un nivel crítico de 0.2704, con lo cual no hay indicios para rechazar la independencia de los datos.

A continuación, haremos la misma validación que para el ejemplo anterior. Estudiaremos gráficamente a qué distribución pueden pertenecer estas observaciones. Para orientarnos cuáles de las distribuciones son buenas candidatas haremos el gráfico de Cullen and Frey.

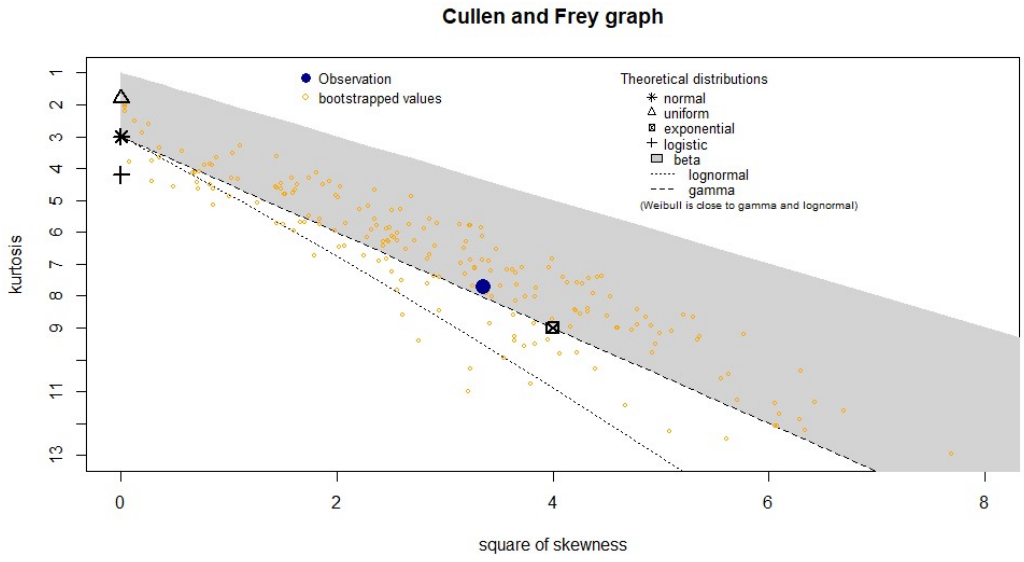

Figura 1.7: Gráfico Cullen and Frey para los datos del Ejemplo 1.21

Como podemos observar, estos datos están mejor repartidos a lo largo de la distribución Gamma. Según el gráfico algunas posibles distribuciones son la Gamma,

beta, exponencial, lognormal. Rechazamos la beta ya que los datos no se encuentran entre 0 y 1 y estudiaremos las demás.

En primer lugar, comencemos con la Gamma.

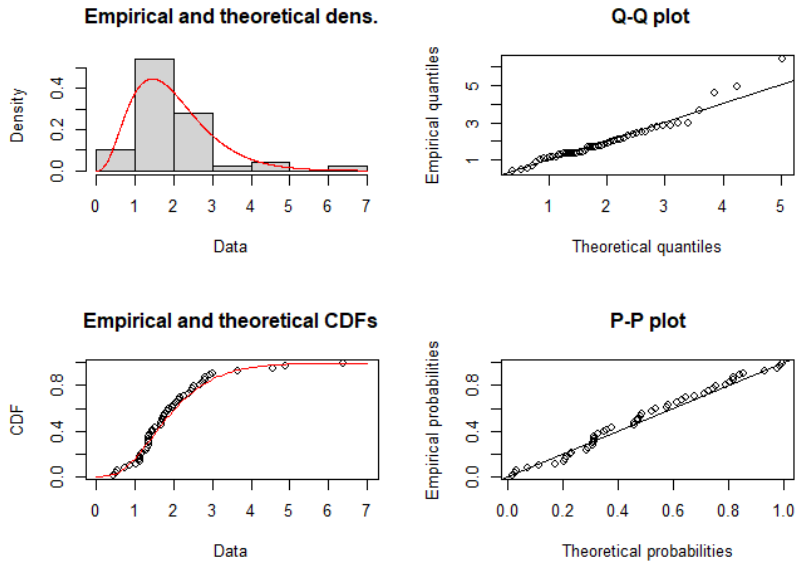

Figura 1.8: Aproximación de la distribución Gamma para los datos del Ejemplo 1.21

Como podemos observar, la función de densidad empírica y la teórica son similares. Por otro lado, hay tres cuantiles empíricos que presentan una desviación respecto de los cuantiles teóricos, sin embargo, los demás presentan muy poca desviación. Además, en las gráficas CFD y PP-Plot se percibe una gran semejanza con las probabilidades teóricas de la Gamma. Con lo cual, la Gamma es una muy buena candidata para afirmar que los datos siguen esta distribución.

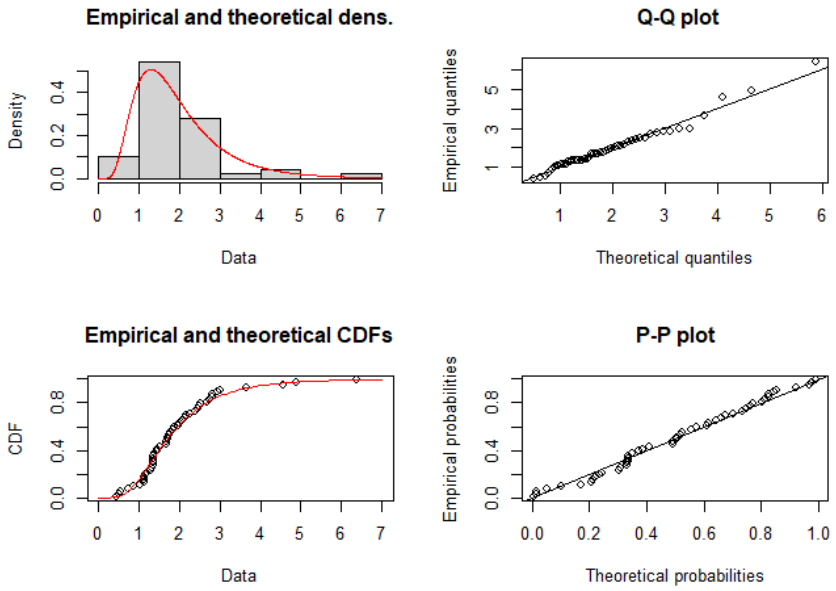

Figura 1.9: Aproximación de la distribución lognormal para los datos del Ejemplo 1.21

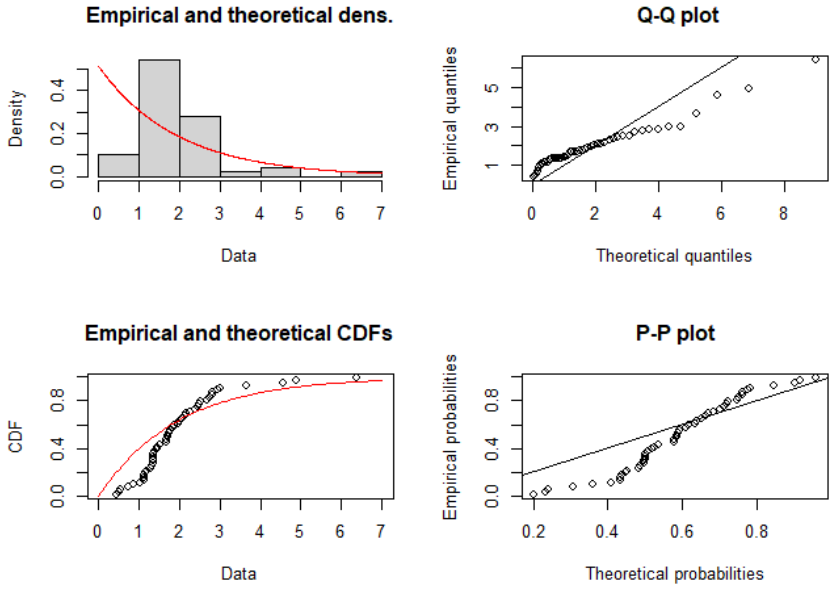

Figura 1.10: Aproximación de la distribución exponencial para los datos del Ejemplo 1.21

Comentemos ahora la figura (1.9), es decir, las gráficas de la distribución lognormal. En el gráfico de la densidad empírica y teórica observamos que la densidad teórica puede ser una buena aproximación de la densidad empírica. Además, los cuantiles y probabilidades empíricos presentan poca desviación respecto a los teóricos. Por lo tanto, podemos decir que la lognormal también parece ser una buena candidata para estos datos.

En cambio, en la figura (1.10) como podemos observar, rechazamos la idea de que los datos sigan una distribución exponencial. Tanto en los gráficos de densidad, cuantiles y probabilidad podemos observar que son muy diferentes los teóricos con los empíricos.

Capítulo 2

Inferencia sobre los parámetros

2.1. Introducción

En este capítulo trataremos sobre la inferencia respecto de los parámetros α y β de la distribución Gamma. Para ello, definamos primero una serie de conceptos previos.

Definición 2.1. ***Muestra aleatoria.*** *Decimos que* $\boldsymbol{X}$ *de tamaño* n *es una muestra aleatoria simple si es un conjunto de variables aleatorias* $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ *independientes e idénticamente distribuidas con la misma distribución que* $\boldsymbol{X}$*.*

Definición 2.2. ***Espacio paramétrico.*** *El espacio paramétrico* Θ *lo definimos como el conjunto de los posibles valores de los parámetros de una distribución.*

Definición 2.3. ***Estadístico.*** *Sea una función de la muestra* $T : \mathbb{R}^n \longrightarrow \mathbb{R}^k$ *y* $\boldsymbol{X}$ *una muestra aleatoria de tamaño* n*. Definimos el estadístico* $T = T(X_1, \ldots, X_n)$ *como una variable aleatoria que es función únicamente del resultado de la muestra. El estadístico tiene asociada una distribución de probabilidad, conocida como distribución en el muestreo.*

Definición 2.4. *Sea* Θ *el espacio paramétrico y* Ω *el espacio muestral. Un estimador* $t(\boldsymbol{X})$ *es cualquier función medible del espacio muestral en el espacio paramétrico,* $t : \Omega \longrightarrow \Theta$*. Una estimación es una realización del estimador.*

A continuación, realizaremos la inferencia sobre los parámetros α y β empleando diferentes métodos.

2.2. Método de momentos

El método de momentos es un método de estimación puntual que consiste en tomar los valores de los parámetros de la distribución que dan lugar a momentos iguales a sus valores muestrales.

A continuación, definiremos más formalmente este método y luego lo aplicaremos a la distribución Gamma.

Definición 2.5. *Sea $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ una muestra aleatoria que viene dada por una distribución con parámetro θ, siendo θ un parámetro k-dimensional con al menos k momentos finitos. Definimos el momento poblacional de orden j como*

$$\mu_j = E(X^j) \text{ con } j = 1, 2, \ldots, k$$

Como la distribución de $\boldsymbol{X}$ depende de θ, también lo harán sus momentos, lo cual conduce a la función $\mu(\theta) = (\mu_1(\theta), \mu_2(\theta), \ldots, \mu_k(\theta))$. Por otro lado definimos, $M(\mu_1, \mu_2, \ldots, \mu_k)$ como la función inversa tal que

$$\theta = M(\mu_1(\theta), \mu_2(\theta), \ldots, \mu_k(\theta))$$

Definimos los momentos muestrales como $m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$. El método de momentos de θ es $M(m_1, m_2, \ldots, m_k)$. Para ello igualamos los k momentos poblacionales con los k momentos muestrales, es decir, $m_j = \mu_j(\theta)$ y lo resolvemos para θ.

Seguidamente, buscaremos los estimadores por el método de momentos de una Gamma. Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria tal que cada $X_i \sim Ga(\alpha, \beta)$ con $i \in \{1, \ldots, n\}$. Como hay dos parámetros desconocidos trabajamos con los momentos de orden 1 y 2. Con lo cual

$$\mu_1 = E(X) = \alpha\beta$$

$$\mu_2 = E(X^2)$$

Como sabemos que $Var(X) = \alpha\beta^2$ y $Var(X) = E(X^2) - E(X)^2$ obtenemos que $E(X^2) = \alpha(\alpha + 1)\beta^2$. Por otro lado, tenemos los momentos muestrales

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Igualamos ambos momentos, poblacional y muestral, y obtenemos

$$m_1 = \hat{\alpha}\hat{\beta} \tag{2.1}$$

$$m_2 = \hat{\alpha}(\hat{\alpha}+1)\hat{\beta}^2 \tag{2.2}$$

Ahora, resolvemos este sistema de ecuaciones para calcular $\hat{\alpha}$ y $\hat{\beta}$. De la ecuación (2.1) tenemos que $\hat{\alpha} = \frac{m_1}{\hat{\beta}}$ y sustituyendo en (2.2) obtenemos

$$m_2 = \frac{m_1}{\hat{\beta}}\left(\frac{m_1}{\hat{\beta}}+1\right)\hat{\beta}^2 = m_1^2 + m_1\hat{\beta} \longrightarrow \hat{\beta} = \frac{m_2 - m_1^2}{m_1}$$

y así, deducimos que $\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2}$.

2.3. Método de máxima verosimilitud

Veamos ahora la estimación de los parámetros por el método de máxima verosimilitud, que es otro método de estimación puntual. Empezaremos recordando los conceptos generales de estimación por máxima verosimilitud.

2.3.1. Elementos generales de la estimación por máxima verosimilitud

Definición 2.6. ***Función de verosimilitud.*** *Asumiendo un modelo estadístico con un parámetro θ fijo pero desconocido, $\theta \in \Theta$, la función de verosimilitud es la probabilidad (o densidad) de la muestra observada x considerada como una función de θ.*

En particular, sea $\boldsymbol{X}$ una muestra aleatoria de tamaño n, $\boldsymbol{X} \sim f(x|\theta)$ y $\boldsymbol{x} = (x_1, \ldots, x_n)$ una realización de la muestra. Definimos la función de verosimilitud como

- *Si $f(x|\theta)$ es la función de probabilidad de una variable discreta*

$$L(\theta) = P[\boldsymbol{X} = \boldsymbol{x}|\theta] = \prod_{i=1}^{n} P(X_i = x_i|\theta)$$

- *Si $f(x|\theta)$ es la función de densidad de una variable continua*

$$L(\theta) = f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Definición 2.7. ***Función de log-verosimilitud.*** *Sea $L(\theta)$ la función de verosimilitud, llamamos función de log-verosimilitud a la función de θ definida como*

$$\ell(\theta) = logL(\theta)$$

Definición 2.8. ***Método de máxima verosimilitud.*** *El estado de máxima verosimilitud es el valor concreto de θ para el cual se alcanza el máximo de $L(\theta)$. El estimador de máxima verosimilitud es aquel que a cada muestra le asocia el correspondiente estado de máxima verosimilitud para el parámetro.*

Seguidamente hallaremos los estimadores de máxima verosimilitud de la distribución Gamma, denotados por $\hat{\alpha}_{VM}$ y $\hat{\beta}_{VM}$. La función de verosimilitud es de la forma

$$L(\alpha, \beta) = \prod_{i=1}^{n} f(x_i|\alpha, \beta) = \prod_{i=1}^{n} \left(\frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} \right) = \frac{1}{\beta^{n\alpha}\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} x_i \right)^{\alpha-1} e^{-\frac{\sum_{i=1}^{n} x_i}{\beta}}$$

Así, la función de log-verosimilitud es

$$\begin{aligned} \ell(\alpha, \beta) &= -n\alpha log(\beta) - nlog(\Gamma(\alpha)) + (\alpha - 1) log \left(\prod_{i=1}^{n} x_i \right) - \frac{\sum_{i=1}^{n} x_i}{\beta} \\ &= -n\alpha log(\beta) - nlog(\Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^{n} log(x_i) - \frac{\sum_{i=1}^{n} x_i}{\beta} \end{aligned} \tag{2.3}$$

2.3.2. Ecuaciones de máxima verosimilitud

Para obtener la máxima verosimilitud tenemos que derivar la función de log-verosimilitud (2.3) respecto de α y β e igualar a 0. Con ello planteamos el siguiente

sistema de ecuaciones

$$\begin{cases} \frac{\partial \ell(\alpha,\beta)}{\partial \alpha} = 0 \\ \frac{\partial \ell(\alpha,\beta)}{\partial \beta} = 0 \end{cases}$$

Primero calculamos $\frac{\partial \ell(\alpha,\beta)}{\partial \beta}$ y obtenemos

$$\frac{\partial \ell(\alpha,\beta)}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{\sum_{i=1}^{n} x_i}{\beta^2} = 0 \Longrightarrow -n\alpha\beta + \sum_{i=1}^{n} x_i = 0 \tag{2.4}$$

Con lo cual, el estimador máximo verosímil del parámetro β es

$$\hat{\beta}_{MV} = \frac{\sum_{i=1}^{n} x_i}{n\hat{\alpha}_{MV}} \tag{2.5}$$

Ahora, hallaremos $\frac{\partial \ell(\alpha,\beta)}{\partial \alpha}$ para obtener $\hat{\alpha}_{MV}$

$$\frac{\partial \ell(\alpha,\beta)}{\partial \alpha} = -nlog(\beta) - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} log(x_i) = 0 \tag{2.6}$$

Como sabemos que $\hat{\beta}_{MV} = \frac{\sum_{i=1}^{n} x_i}{n\hat{\alpha}_{MV}}$ sustituimos esta expresión en la ecuación (2.6). Así, para obtener $\hat{\alpha}_{MV}$ tenemos que resolver la siguiente ecuación

$$\begin{aligned} & -nlog\left(\frac{\sum_{i=1}^{n} x_i}{n\hat{\alpha}_{MV}}\right) - \frac{n\Gamma'(\hat{\alpha}_{MV})}{\Gamma(\hat{\alpha}_{MV})} + \sum_{i=1}^{n} log(x_i) = 0 \\ \Longrightarrow & -nlog\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) + nlog(\hat{\alpha}_{MV}) - \frac{n\Gamma'(\hat{\alpha}_{MV})}{\Gamma(\hat{\alpha}_{MV})} + \sum_{i=1}^{n} log(x_i) = 0 \\ \Longrightarrow & \; n\left(log(\hat{\alpha}_{MV}) - \frac{\Gamma'(\hat{\alpha}_{MV})}{\Gamma(\hat{\alpha}_{MV})}\right) = nlog\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) - \sum_{i=1}^{n} log(x_i) \end{aligned}$$

Con lo cual, el estimador máximo verosímil de α, es decir $\hat{\alpha}_{MV}$, se puede calcular de la siguiente forma

$$log(\hat{\alpha}_{MV}) - \frac{\Gamma'(\hat{\alpha}_{MV})}{\Gamma(\hat{\alpha}_{MV})} = log\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) - \frac{1}{n}\sum_{i=1}^{n} log(x_i)$$

Como podemos observar, esta ecuación es muy difícil de resolver explícitamente. Si conocemos $\hat{\alpha}_{MV}$ podemos obtener $\hat{\beta}_{MV}$ de manera inmediata. Es muy habitual

recurrir a métodos numéricos para aproximar este valor $\hat{\alpha}_{MV}$. Seguidamente trataremos sobre algunos métodos iterativos.

2.3.3. Método de Newton

En este apartado trataremos de calcular los estimadores de máxima verosimilitud a partir del método iterativo de Newton. Antes de explicar brevemente este método introduzcamos alguna notación que ayudará a seguir más claramente los procedimientos de este algoritmo.

En primer lugar, llamaremos digamma y trigamma a la primera y segunda derivadas de la función Gamma, es decir, $\frac{\partial}{\partial\alpha}\Gamma(\alpha)$ y $\frac{\partial^2}{\partial\alpha^2}\Gamma(\alpha)$ respectivamente.

Por otro lado, denotemos como

$$\Psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

$$\Upsilon = log\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) - \frac{1}{n}\sum_{i=1}^{n} log(x_i)$$

Algoritmo 2.9. *El método iterativo de Newton se basa en ir calculando α_k hasta encontrar un α_{k+1} y α_k que estén muy cercanos entre sí. Calculamos α_{k+1} a partir de α_k de la siguiente forma*

$$\alpha_{k+1} = \alpha_k - \frac{g(\alpha_k)}{g'(\alpha_k)}$$

siendo $g(\alpha) = log(\alpha) - \Psi(\alpha) - \Upsilon$

Este algoritmo proporciona estimadores con problemas de estabilidad numérica, con lo cual, los estimadores obtenidos a partir de este método pierden exactitud. Por ello, en el siguiente apartado estudiaremos otra forma para obtener los estimadores de máxima verosimilitud.

Ejemplo 2.10. *Implementando este método en R con los datos del Ejemplo 1.20 tenemos que*

$$\begin{cases} \hat{\alpha}_{MV} = 0.89940799 \\ \hat{\beta}_{MV} = 14.25179893 \end{cases}$$

Además, el algoritmo convergió en la iteración 6 con un error de 4.15e-14 entre α_{k+1} y α_k.

Los tres programas de R empleados para conseguir esta solución se encuentran en el apéndice, Comandos de R: Capítulo 2, ver [4.4].

2.3.4. Máxima verosimilitud en R

Los estimadores máximo verosímiles pueden ser calculados en R a través del comando *mle* que se encuentra en la librería *stats*4. Este comando usa el optimizador *optim* por defecto que se encarga de minimizar la función de log-verosimilitud negativa. Este mínimo se obtiene a partir de una matriz de covarianzas aproximada para los parámetros invirtiendo la matriz hessiana en el óptimo.

Ejemplo 2.11. *Recordamos los datos usados en el Ejemplo 1.20. Calculemos, con la ayuda de R, los estimadores de máxima verosimilitud con la función mle. Especificamos en el apéndice, Comandos en R: Capítulo 2, el script empleado para llegar a los resultados que se muestran a continuación, ver [4.4].*

De esta forma, R nos calcula explícitamente los estimadores de máxima verosimilitud. Es importante destacar que R nos muestra $\hat{\alpha}_{MV}$ *y* $\hat{\lambda}_{MV} = \frac{1}{\hat{\beta}_{MV}}$. *Así, obtenemos*

$$\begin{cases} \hat{\alpha}_{MV} = 0.89945600 \\ \hat{\lambda}_{MV} = 0.07017104 \Longrightarrow \hat{\beta}_{MV} = 14.25089324 \end{cases}$$

Como podemos observar los estimadores obtenidos coinciden en cuatro cifras significativas.

2.4. Estimación por intervalos de confianza

2.4.1. Intervalos de confianza (método pivotal)

Sea $X \sim Ga(\alpha, \beta)$. Veamos ahora cuál es el intervalo de confianza para β con α conocido. Si $X_i \sim Ga(\alpha, \beta)$ y son independientes entonces, por la reproductividad de la distribución Gamma respecto del parámetro de forma, $\sum_{i=1}^{n} X_i \sim Ga(n\alpha, \beta)$. Podemos observar que $\sum_{i=1}^{n} X_i$ es un estadístico suficiente para β.

Además, recordemos también que si $T \sim Ga(\alpha, \beta)$ entonces $aT \sim Ga(\alpha, a\beta)$. Con lo cual, en nuestro caso haciendo ciertas operaciones llegamos a una distribución ji-

cuadrado

$$\sum_{i=1}^{n} X_i \sim Ga(n\alpha, \beta) \Longrightarrow \frac{2}{\beta}\sum_{i=1}^{n} X_i \sim Ga(n\alpha, 2) = Ga\left(\frac{2n\alpha}{2}, 2\right) = \chi^2(2n\alpha)$$

De esta forma, los siguientes pasos conducen a un intervalo de confianza para β con un nivel de confianza $(1-\alpha)$

$$\begin{aligned} 1-\alpha &= P\left[\chi^2_{2n\alpha;\frac{\alpha}{2}} \leq \frac{2}{\beta}\sum_{i=1}^{n} X_i \leq \chi^2_{2n\alpha;1-\frac{\alpha}{2}}\right] \\ &= P\left[\frac{\chi^2_{2n\alpha;\frac{\alpha}{2}}}{2\sum_{i=1}^{n} X_i} \leq \frac{1}{\beta} \leq \frac{\chi^2_{2n\alpha;1-\frac{\alpha}{2}}}{2\sum_{i=1}^{n} X_i}\right] \\ &= P\left[\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{2n\alpha;1-\frac{\alpha}{2}}} \leq \beta \leq \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{2n\alpha;\frac{\alpha}{2}}}\right] \end{aligned}$$

siendo $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ los cuantiles que dejan una probabilidad $\frac{\alpha}{2}$ y $\left(1-\frac{\alpha}{2}\right)$ a la izquierda de la distribución ji-cuadrado con $2n\alpha$ grados de libertad.

Por tanto, el intervalo de confianza para β con un nivel de confianza $(1-\alpha)$ viene dado por

$$\left[\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{2n\alpha;1-\frac{\alpha}{2}}}, \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{2n\alpha;\frac{\alpha}{2}}}\right] \tag{2.7}$$

2.4.2. Intervalos de confianza asintóticos

Seguidamente, hallaremos la distribución asintótica del estimador máximo verosímil. Recordamos el resultado que establece la distribución asintótica de los estimadores de máxima verosimilitud, que después aplicaremos a la distribución Gamma.

Enunciamos la proposición para un parámetro θ univariante, y lo aplicaremos después para obtener un intervalo de confianza para β con α conocido.

Proposición 2.12. *Bajo ciertas condiciones de regularidad, cualquier sucesión $\hat{\theta}_n = \theta(X_1, \ldots, X_n)$, de raíces de la ecuación de verosimilitud consistentes para θ, se distribuye asintóticamente como una normal.*

$$\hat{\theta}_n \longrightarrow Z \sim N\left(\theta, \frac{1}{nI_X(\theta)}\right)$$

siendo $I_X(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} log f(X|\theta)\right]$ la información de Fisher de cada observación.

Como la información de Fisher es desconocida entonces debemos estimarla. Como sabemos que $\hat{\theta}_n$ converge puntualmente a θ y la información de Fisher es continua entonces $I(\hat{\theta}_n)$ converge puntualmente a $I(\theta)$. Así, podemos aproximar

$$\hat{\theta} \approx N\left(\theta, \frac{1}{nI_X(\hat{\theta})}\right)$$

Con lo cual, se puede considerar un intervalo de confianza asintótico de la siguiente manera

$$\left[\hat{\theta} - z_{1-\alpha/2}\left(nI_X(\hat{\theta})\right)^{-\frac{1}{2}}, \hat{\theta} + z_{1-\alpha/2}\left(nI_X(\hat{\theta})\right)^{-\frac{1}{2}}\right]$$

siendo $z_{1-\alpha/2}$ el cuantil que deja una probabilidad $\left(1-\frac{\alpha}{2}\right)$ a la izquierda en la distribución normal estándar.

Particularicemos estas definiciones para el caso de la distribución Gamma. Calculemos el intervalo de confianza asintótico para β cuando α es conocido.

Sea $X \sim Ga(\alpha, \beta)$ y calculemos su información de Fisher $I_X(\beta)$.

$$logf(X|\alpha,\beta) = -\alpha log(\beta) - log(\Gamma(\alpha)) + (\alpha-1)log(X) - \frac{X}{\beta}$$

A continuación, hagamos la primera y segunda derivadas

$$\frac{\partial}{\partial\beta}logf(X|\alpha,\beta) = -\frac{\alpha}{\beta} + \frac{X}{\beta^2}$$
$$\frac{\partial^2}{\partial\beta^2}logf(X|\alpha,\beta) = \frac{\alpha}{\beta^2} - \frac{2X}{\beta^3}$$

Así,

$$I_X(\beta) = -E\left[\frac{\alpha}{\beta^2} - \frac{2X}{\beta^3}\right] = \frac{\alpha}{\beta^2}$$

Resolviendo la ecuación de verosimilitud (2.4) con α conocido tenemos como estimador de β, $\hat{\beta} = \frac{\bar{X}}{\alpha}$. Empleando la proposición 2.12 se llega a que

$$\hat{\beta} = \frac{\bar{X}}{\alpha} \approx N\left(\beta, \frac{\hat{\beta}^2}{n\alpha}\right)$$

y el intervalo de confianza para β resulta

$$\left[\frac{\bar{X}}{\alpha} - z_{1-\alpha/2}\left(\frac{\bar{X}}{\sqrt{n\alpha^3}}\right), \frac{\bar{X}}{\alpha} + z_{1-\alpha/2}\left(\frac{\bar{X}}{\sqrt{n\alpha^3}}\right)\right] \tag{2.8}$$

En el siguiente ejemplo calcularemos los intervalos de confianza presentados anteriormente en las expresiones (2.7) y (2.8). Éstos los hallaremos a partir de datos simulados ya que para el cálculo de los intervalos de confianza partimos de la hipótesis de que α es conocido. Sin embargo, si usamos los ejemplos de datos reales del Capítulo 1 no conocemos α, solo conocemos una estimación de este parámetro.

Ejemplo 2.13. *A continuación, vamos a generar en R unos datos de tamaño muestral $n = 50$ que siguen una distribución Gamma de parámetros $\alpha = 2$ y $\beta = 3$. Calcularemos los intervalos de confianza obtenidos anteriormente, con un nivel de confianza del 95 %. Previamente, hallaremos los estimadores máximo verosímiles para los parámetros de forma y escala de esta simulación de datos. El script empleado para comprobar estos resultados se encuentra en el apéndice, Comandos de R: Capítulo 2, ver [4.4].*

Para comenzar, calculemos los estimadores máximo verosímiles para este conjunto de datos. Con el comando mle vemos que estos estimadores son

$$\begin{cases} \hat{\alpha}_{MV} = 1.6793506 \\ \hat{\lambda}_{MV} = 0.3658429 \Longrightarrow \hat{\beta}_{MV} = 2.7334137 \end{cases}$$

Ahora, calculemos el intervalo de confianza para β con el método pivotal (2.7). Calculándolo en R obtenemos que el intervalo de confianza a un nivel del 95 % para $\lambda = \frac{1}{\beta}$ es (0.3544987, 0.5251384). Con lo cual, el intervalo de confianza para β es (1.90426, 2.820885).

En el caso del intervalo asintótico de β, usamos el intervalo obtenido en (2.8). De esta forma, obtenemos los siguientes intervalos asintóticos. El intervalo de confianza asintótico a un nivel del 95 % para λ es (0.3642943, 0.5419064) y para β es (1.845337, 2.745033).

Como podemos observar, en ambos intervalos calculados el valor real de β, que es 3, no está dentro del intervalo de confianza. Sin embargo, el estimador máximo verosímil sí que pertenece a estos intervalos.

2.4.3. Intervalos de confianza en R

En este apartado vamos a calcular los intervalos de confianza de los parámetros α y β en R a partir del comando *confint* que proporciona los intervalos de confianza mediante un perfil de verosimilitudes. En primer lugar, definiremos el perfil de verosimilitud y los intervalos de confianza basados en el perfil de verosimilitud, y después lo aplicaremos a dos ejemplos.

Definición 2.14. *Se define el perfil de verosimilitud de un parámetros β_j como*

$$PL(\beta_j) = \underset{\beta_1,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p}{\text{máx}} L(\beta_1,\ldots,\beta_{j-1},\beta_j,\beta_{j+1},\ldots,\beta_p)$$

Se trata, por tanto, de fijar el valor β_j y maximizar respecto de los demás parámetros. Como se ha fijado el valor β_j esto produce que la verosimilitud empeore, es decir

$$PL(\beta_j) > PL(\hat{\beta}_j) = L(\hat{\beta})$$

Con lo cual, es más verosímil aquel valor de β_j que perjudica menos a la verosimilitud. Los valores de β_j que son más verosímiles, cuyo perfil de verosimilitud es más grande, son los que forman parte del intervalo de confianza para β_j.

Al nivel de confianza $(1-\alpha)$, se toma como intervalo de confianza el siguiente conjunto de valores

$$\left\{\beta_j : 2\left(PL(\hat{\beta}_j) - PL(\beta_j)\right) < \chi^2_{1,\alpha}\right\}$$

siendo $\chi^2_{1,\alpha}$ el cuantil que deja una probabilidad $(1-\alpha)$ a la izquierda de la distribución ji-cuadrado.

Los intervalos de confianza obtenidos por el perfil de verosimilitud son asimétricos en general, sobre todo para tamaños muestrales pequeños. Cuanto mayor sea el tamaño muestral más semejantes serán estos intervalos y los intervalos de confianza asintóticos.

A continuación, calcularemos los intervalos de confianza basados en el perfil de verosimilitud de los siguientes ejmplos,

Ejemplo 2.15. *Para el ejemplo anterior donde generamos datos de parámetros $\alpha = 2$ y $\lambda = 3$ y tamaño muestral $n = 50$. Usando el comando confint sobre los*

estimadores máximo verosímiles tenemos que el intervalo de confianza para α es (1.149967, 2.363811) y para $\lambda = \frac{1}{\beta}$ obtenemos (0.2322214, 0.5397209). Así, deducimos que el intervalo de confianza de β es (1.852809, 4.306235). Podemos observar que son ligeramente diferentes los intervalos de confianza asintóticos con los calculados en R pero los estimadores máximo verosímiles siempre pertenecen a estos intervalos.

Ejemplo 2.16. *Ahora usamos los datos reales del Ejemplo 1.20. Para calcular los intervalos de confianza en R empleamos otra vez el comando confint sobre los valores máximo verosímiles obtenidos en el apartado (2.3.3).*

De esta manera obtenemos que el intervalo de confianza para α es (0.63973523, 1.2285564) y el intervalo de confianza para λ es (0.04388603, 0.1042655), con lo cual para β es (9.5909, 22.78629).

Como podemos observar, los estimadores que hemos calculado en la sección anterior se encuentran dentro de estos intervalos de confianza.

Capítulo 3

Criterios de bondad del ajuste

3.1. Introducción

Cuando tenemos un conjunto de datos y queremos hacer inferencia sobre ellos, primero hay que verificar si esas observaciones siguen una distribución conocida. En este capítulo estudiaremos pruebas de bondad del ajuste que permiten validar un modelo de distribución para un cierto conjunto de datos. En nuestro caso, nos centraremos en pruebas para verificar una Gamma.

Previamente se puede realizar un análisis gráfico, como ya se mencionó en los ejemplos de la sección 1.8, a través de la librería de *R fitdistrplus*. De esta forma, podemos deducir qué distribuciones son adecuadas y cuáles no lo son. Aunque este método sirve como orientación, no es un test exacto. Por ello, estudiaremos la prueba de Lilliefors que nos permitirá conocer si un determinado conjunto de datos se ajusta a una distribución Gamma.

Este método, también llamado test de Kolmogorov-Smirnov/Lilliefors, se basa en el test de Kolmogorov-Smirnov original (KS). Por ello, vamos a explicar primero brevemente este procedimiento.

3.2. Test de Kolmogorov-Smirnov

Supongamos que tenemos una muestra aleatoria $\mathbf{X}$ de tamaño muestral n y F es su función de distribución, la cual no es conocida. El objetivo es aceptar la hipótesis

nula de este contraste

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$$

siendo F_0 la función de distribución conocida, en nuestro caso tendrá que ser la Gamma con unos parámetros concretos.

El test de Kolmogorov-Smirnov es una prueba de bondad de ajuste que se basa en la diferencia entre la distribución empírica y la teórica. Sin embargo, sólo se puede aplicar este test si los parámetros de la distribución teórica son conocidos completamente. Este test utiliza el estadístico de Kolmogorov-Smirnov:

$$D_n = sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \tag{3.1}$$

siendo $F_n(x)$ la probabilidad empírica acumulativa y $F(x)$ la probabilidad teórica acumulativa.

Rechazamos la hipótesis nula cuando D_n sea "grande" pues es un indicio en contra de la igualdad $F = F_0$.

Sin embargo, este proceso es poco práctico ya que, en general, queremos estudiar si unos datos siguen una distribución conocida pero sin conocer los parámetros. Éstos se estiman a partir de los mismos datos que queremos aproximar a una distribución. Por eso, no podemos aplicar el test de Kolmogorov-Smirnov original sino que usamos una modificación de éste, la prueba de Lilliefors.

3.3. Prueba de Lilliefors

Igual que en el caso anterior, la prueba de Lilliefors usa el estadístico de Lilliefors que se define como

$$L_n = sup_{x \in \mathbb{R}} |F_n(x) - F_{\hat{\theta}}(x)|$$

siendo $\hat{\theta}$ el estimador de los parámetros θ del modelo de distribución F_θ

La diferencia con el test anterior es que rechazamos la hipótesis nula si el valor del estadístico supera los valores críticos de la prueba de Lilliefors que son diferentes a los del test de Kolmogorov-Smirnov. Para el cálculo de estos valores críticos cuando se contrasta una distribución Gamma emplearemos las simulaciones de Monte Carlo.

Este proceso consiste en realizar muchas simulaciones, en particular, generamos $Ns = 5000$ muestras. Para cada parámetro de forma y cada tamaño muestral,

calculamos el estadístico L_n con 5000 muestras. Además, los parámetros α y β de la función de distribución Gamma, $F_{\alpha,\beta}(x)$, se estimaron a partir de los datos simulados. Para elaborar las tablas se empleó únicamente el valor $\beta = 40$, pues la distribución del estadístico de Lilliefors para la distribución Gamma no depende del valor de β.

Es importante resaltar que para estimar el parámetro α usaremos la aproximación introducida por Thom Herbert (1958-1968). Sea $A = log(\bar{X}) - \frac{\sum_{i=1}^{n} log(X_i)}{n}$ entonces estimamos α de la siguiente forma

$$\hat{\alpha} = \frac{1 + \sqrt{1 + \frac{4A}{3}}}{4A} \tag{3.2}$$

y $\hat{\beta}$ es de la forma $\frac{\bar{X}}{\hat{\alpha}}$.

Para mayor detalle sobre la estimación de α consultar [7].

3.3.1. Tabla de valores críticos

A continuación, calcularemos tres tablas que se presentan en los cuadros 3.1, 3.2 y 3.3. El cuadro 3.1 contiene los valores críticos de la prueba de Lilliefors a un nivel de significación del 1 %, el cuadro 3.2 tiene una significación del 5 % y el cuadro 3.3 del 10 %. Los comandos de *R* empleados para realizar estas tablas se encuentran en el apéndice, *Comandos R: Capítulo 3*, ver [4.4].

Cuadro 3.1: Valores críticos de la prueba de Lilliefors a un nivel de significación del 1 %

	Tamaño muestral					
Parámetro forma	40	45	50	55	60	70
0.5	0.171	0.165	0.155	0.150	0.140	0.132
1	0.163	0.155	0.148	0.137	0.132	0.124
1.5	0.161	0.153	0.146	0.139	0.132	0.121
2	0.160	0.151	0.143	0.137	0.132	0.121
2.5	0.157	0.150	0.141	0.134	0.132	0.121
3	0.157	0.150	0.141	0.134	0.132	0.121
4	0.157	0.147	0.141	0.134	0.129	0.121
8	0.157	0.149	0.140	0.134	0.128	0.120

Cuadro 3.2: Valores críticos de la prueba de Lilliefors a un nivel de significación del 5 %

	Tamaño muestral					
Parámetro forma	40	45	50	55	60	70
0.5	0.146	0.138	0.132	0.128	0.121	0.112
1	0.137	0.132	0.124	0.118	0.113	0.105
1.5	0.134	0.129	0.123	0.116	0.112	0.104
2	0.134	0.128	0.121	0.116	0.111	0.104
2.5	0.132	0.128	0.121	0.114	0.111	0.103
3	0.132	0.127	0.120	0.114	0.111	0.103
4	0.132	0.126	0.119	0.114	0.111	0.103
8	0.132	0.126	0.118	0.114	0.109	0.101

Cuadro 3.3: Valores críticos de la prueba de Lilliefors a un nivel de significación del 10 %

	Tamaño muestral					
Parámetro forma	40	45	50	55	60	70
0.5	0.132	0.125	0.120	0.115	0.111	0.102
1	0.124	0.118	0.112	0.108	0.103	0.0958
1.5	0.122	0.116	0.111	0.106	0.102	0.0944
2	0.122	0.116	0.109	0.106	0.102	0.0943
2.5	0.121	0.115	0.109	0.105	0.102	0.0934
3	0.121	0.115	0.108	0.104	0.101	0.0933
4	0.120	0.114	0.108	0.104	0.101	0.0931
8	0.120	0.114	0.108	0.104	0.0988	0.0931

Apliquemos esta prueba a nuestros ejemplos de la sección 1.8. Los comandos empleados para el cálculo de los estadísticos de Kolmogorov-Smirnov de los datos también se encuentran en el apéndice.

Empecemos con el Ejemplo 1.20 sobre las lluvias diarias, cuyo tamaño muestral es $n = 55$. Recordamos que estos datos tienen autocorrelación a un nivel de significación del 10 %, lo cual puede afectar a la prueba de Lilliefors. En primer lugar, calculamos en R las estimaciones de los parámetros de la distribución Gamma. Empleando la estimación dada en la expresión (3.2) obtenemos que $\hat{\alpha}$ =0.91 y $\hat{\beta} = \frac{\bar{X}}{\hat{\alpha}}$=14.09. Además, el estadístico L_n es 0.146.

Analizamos ahora los valores críticos de Lilliefors. Para el parámetro de forma tomamos 1, que es el más cercano a $\hat{\alpha}$, y tamaño muestral $n = 55$. Así, los valores

críticos son 0.137, 0.118 y 0.108 para los niveles de significación 1 %, 5 % y 10 % respectivamente. Así, como L_n es mayor que estos valores críticos entonces se rechaza la hipótesis nula de que estos datos sigan una distribución Gamma. Nótese que la autocorrelación también puede afectar también al test de Lilliefors, que está concebido para datos independientes.

Por otra parte, efectuaremos el test de Lilliefors para el Ejemplo 1.21 de las lluvias anuales de Ithaca, Nueva York, desde 1933 hasta 1982, cuyo tamaño muestral es $n = 50$. Primeramente, obtenemos las estimaciones de los parámetros de la distribución Gamma. Usando la expresión dada en (3.2) tenemos que la estimación de α es $\hat{\alpha}$=3.765. De esta forma, podemos obtener que $\hat{\beta} = \frac{\bar{X}}{\hat{\alpha}}$ es 0.521. Además, calculamos el estadístico de Lilliefors y obtenemos que L_n=0.0726

Con lo cual, tomaremos el valor crítico de Lilliefors con $n = 50$ y el parámetro de forma igual a 4, que es el que más se aproxima a $\hat{\alpha}$=3.765. Como podemos observar, estos valores críticos son 0.141, 0.119, 0.108 para los niveles de significación 1 %, 5 % y 10 % respectivamente. Como podemos deducir, L_n es menor que estos valores críticos. De esta forma, podemos confiar en que los datos de las lluvias anuales siguen una distribución Gamma.

Capítulo 4

Regresión Gamma

En esta sección trataremos sobre la regresión Gamma. Este tipo de regresión es un caso particular de modelos lineales generalizados, con lo cual trataremos con más detalle estos modelos.

Un modelo lineal generalizado (GLM) es una generalización flexible de una regresión lineal ordinaria cuya variable respuesta tiene errores con una distribución distinta de la normal.

4.1. Componentes de un modelo lineal generalizado

El modelo lineal generalizado está formado por:

- Componente aleatoria: consiste en una variable respuesta aleatoria Y. Estos modelos se pueden incluir dentro de la familia exponencial, que tiene la siguiente forma

$$f_Y(y) = exp\left(\frac{y\theta + b(\theta)}{a(\Phi)} + c(y, \Phi)\right)$$

  donde θ es el parámetro canónico y representa la localización y Φ es el de dispersión y representa la escala. En la familia exponencial las distribuciones tienenhttps://es.overleaf.com/project/60ad122ae559172b7b587075 media y varianza conocidos, y adoptan la siguiente forma

$$E[Y] = \mu = b'(\theta)$$
$$Var[Y] = b''(\theta)a(\Phi)$$

Algunos ejemplos de distribuciones dentro de la familia exponencial son la distribución normal, Gaussiana, Poisson, binomial o Gamma.

- Componente sistemática: son las variables explicativas del modelo lineal generalizado y se relacionan con la variable respuesta a través de

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

A esta combinación lineal de variables la llamamos predictor lineal. Además, esta combinación la podemos representar de manera matricial como $X^T\beta$.

- Función de enlace o link: describe cómo se relaciona la media de la variable respuesta con la combinación lineal de las variables explicativas, es decir, el predictor lineal, $\eta = X^T\beta$

Así, sea g la función link y sea $E[Y] = \mu$ entonces tenemos que $g(\mu) = \eta$. Para la familia exponencial es importante incidir en las funciones de enlace canónicas. Estas funciones de enlace, g, son de la forma $\eta = g(\mu) = \theta$, es decir, $g(b'(\theta)) = \theta$.

4.2. Proceso de estimación de los GLM

Los parámetros β de los GLM pueden ser estimados usando el método de máxima verosimilitud. Tomamos la función de log-verosimilitud para cada observación donde $a_i(\Phi) = \frac{\Phi}{w_i}$, es decir

$$logL(\theta_i, \Phi, y_i) = w_i \left[\frac{y_i\theta_i - b(\theta_i)}{\Phi}\right] + c(y_i, \Phi)$$

Con lo cual para estimar β maximizamos la función de log-versimilitud $\sum_{i=1}^{n} logL(\theta_i, \Phi, y_i)$. En general, no podemos encontrar la solución exacta, y tenemos que aplicar un método iterativo, como el método de Newton-Raphson o el método IRLS (iteratively reweighted least squares).

A continuación, veremos en qué consiste el método IRLS. La idea es hacer la regresión de $g(y)$ sobre X con los pesos que son inversamente proporcionales a

$Var(Y)$. Consideramos una linealización de la función g entorno a μ

$$g(y) \approx g(\mu) + (y - \mu)g'(\mu)$$

Recordando que $\eta = g(\mu)$ y $\mu = E[Y]$ y denotando $z = g(y)$ obtenemos

$$z = \eta + (y - \mu)\frac{\partial \eta}{\partial \mu}$$

y

$$\hat{Var}(z) = \left(\frac{\partial \eta}{\partial \mu}\right)^2 \ V(\hat{\mu}) = \frac{1}{w}$$

Introducida esta notación, explicamos brevemente los pasos del método IRLS:

1. Fijamos estimaciones iniciales de $\hat{\eta}_0$ y $\hat{\mu}_0$.
2. Ajustamos la variable respuesta z, $z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\frac{\partial \eta}{\partial \mu}|_{\hat{\eta}_0}$
3. Calculamos la inversa de los pesos $w_0^{-1} = \left(\frac{\partial \eta}{\partial \mu}\right)^2 |_{\hat{\eta}_0} V(\hat{\mu}_0)$
4. Realizamos la regresión de z_0 sobre las variables explicativas, $x_1, \ldots, x_n$ con los pesos w_0 para estimar $\hat{\beta}_1$ y posteriormente $\hat{\mu}_1$ a partir del predictor lineal.
5. Repetimos el proceso anterior hasta su convergencia, es decir, hasta que los cambios producidos sean lo suficientemente pequeños.

Asimismo, las estimaciones de la varianza pueden obtenerse a partir de la siguiente expresión

$$Var(\hat{\beta}) = (X^T W X)^{-1}\hat{\Phi}$$

siendo W la matriz de pesos y X la matriz de diseño.

4.3. Deviance del modelo

Introducimos la "deviance" o desviación del modelo, que es el análogo de la suma residual de cuadrados en los modelos lineales generalizados. Se emplea para realizar contrastes entre modelos, similares al test F.

Dadas n observaciones podemos definir el modelo completo o saturado, como un modelo con n parámetros, uno por cada dato, es decir, que estos parámetros coinciden con los valores de la variable respuesta Y.

Consideramos la diferencia entre la función de verosimilitud del modelo completo $L(y, \Phi|y)$ y la del modelo de p parámetros considerado $L(\hat{\mu}, \Phi|y)$. Se representa de la siguiente manera

$$2(L(y, \Phi|y) - L(\hat{\mu}, \Phi|y))$$

Sabiendo que las observaciones son independientes y para la familia exponencial cuando $a_i(\Phi) = \frac{\Phi}{w_i}$, esta desviación se expresa como

$$D(y; \hat{\mu})/\Phi = \sum_{i=1}^{n} 2w_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i))/\Phi$$

donde $\tilde{\theta}$ es el estimador para el modelo completo o saturado y $\hat{\theta}$ estima el modelo de interés, el modelo de p parámetros. Además, llamamos desviación a $D(y; \hat{\mu})$ y a $D(y; \hat{\mu})/\Phi$ desviación escalada.

4.4. Regresión Gamma

En esta sección particularizamos el modelo lineal generalizado para la distribución Gamma. Recordemos que la función de densidad de esta distribución, vista en el *Capítulo 1*, es de la forma

$$f_X(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & \text{otro caso} \end{cases}$$

con $\alpha > 0$ y $\beta > 0$.

No obstante, para los GLM reparametrizamos esta función. Denotaremos a ν como el parámetro de forma y $\beta = \frac{\mu}{\nu}$. Con lo cual, obtenemos

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1} e^{-\frac{y\nu}{\mu}} & y > 0 \\ 0 & \text{otro caso} \end{cases}$$

Con esta reparametrización obtenemos que $E[Y] = \alpha\beta = \nu\frac{\mu}{\nu} = \mu$ y $Var[Y] = \alpha\beta^2 = \nu(\frac{\mu}{\nu})^2 = \frac{\mu^2}{\nu} = \frac{(E[Y])^2}{\nu}$.

En el caso de la Gamma el parámetro canónico es $-\frac{1}{\mu}$. Así, tenemos que el enlace canónico es $\eta = -\frac{1}{\mu}$ aunque es típico representarlo sin el signo negativo, $\eta = \mu^{-1}$, siempre que se tenga en cuenta en las derivaciones. También tenemos que $b(\theta) = log\left(\frac{1}{\mu}\right) = -log(-\theta)$ y $b''(\theta) = \mu^2$.

La deviance de esta regresión es de la forma

$$D(y, \hat{\mu}) = \sum_{i=1}^{n} \left(log\left(\frac{y_i}{\hat{\mu}_i}\right) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)$$

Respecto a las funciones de enlace hay tres opciones:

1. El enlace canónico, como ya explicamos, este enlace es de la forma $\eta = \mu^{-1}$ y como $-\infty < \eta < \infty$ esta función link no garantiza la no negatividad de μ, lo cual requiere restricciones en β para evitar problemas. Este enlace es ideal en situaciones en las cuales sabemos que la media es acotada.

2. El enlace logarítmico es $\eta = log(\mu)$ y se usa cuando el efecto de las variables explicativas es multiplicativo. Cuando la varianza es pequeña este modelo se asemeja a un modelo gaussiano con respuesta logarítmica.

3. El enlace lineal $\eta = \mu$ se usa para modelar sumas de cuadrados o componentes de la varianza que son χ^2, distribución que es un caso particular de la Gamma.

Para estimar la dispersión, McCullagh y Nelder (1989), para más detalle ver [15], recomiendan usar

$$\hat{\Phi} = \frac{1}{\hat{\nu}} = \frac{X^2}{n - p}$$

Así, obtenemos un estimador consistente y este es empleado para estimar $Var(\hat{\beta})$ pues, como ya vimos en la sección anterior, $Var(\hat{\beta}) = (X^T W X)^{-1}\hat{\Phi}$.

El estimador máximo verosímil y el estimador habitual $\frac{D}{n-p}$ no son consistentes cuando la variable respuesta no tiene una distribución Gamma, con lo cual no es recomendable su uso.

La regresión Gamma es muy útil cuando tenemos la seguridad de que la respuesta sigue esta distribución. Alternativamente, se puede realizar una transformación a la variable respuesta para corregir la heterocedasticidad y aplicar un modelo lineal ordinario a la variable transformada.

Así, respecto de la varianza de la variable Y nos podemos encontrar tres casos:

1. La varianza sea constante con lo cual utilizamos la optimización por mínimos cuadrados.

2. La varianza es no constante pero se conoce con exactitud cómo evoluciona. En este caso, se emplea el método de los mínimos cuadrados ponderados.

3. La varianza es no constante pero tampoco se conoce cómo evoluciona. Este es un caso muy común en la práctica y se emplea una transformación de Y que conduce a un modelo con varianza constante. Esta transformación suele ser el logaritmo de Y o su raíz cuadrada. Si se quiere evitar realizar una transformación se puede utilizar un modelo lineal generalizado.

Si la distribución de Y es desconocida es complicado ver que es más adecuado: realizar la transformación logarítmica, que implica una distribución lognormal para la variable respuesta, o utilizar un GLM Gamma. La elección está condicionada por el objetivo y la interpretación deseada del modelo.

A continuación, veremos un ejemplo que se encuentran en la librería "faraway" de R. Este ejemplo ha sido tomado de [14].

Ejemplo 4.1. *Tomamos unos datos sobre el proceso de fabricación de semiconductores. Estos datos se encuentran en la base de datos "wafer" de la librería faraway de R. Se cree que cuatro factores influyen en la resistividad del agua. Estos cuatro factores pueden presentar dos estados + o -.*

Vamos a usar el enlace $\eta = log(\mu)$. Para realizar cualquier modelo lineal generalizado se usa el comando glm. Para usar este comando indicamos con qué distribución lo hacemos y la función link empleada. En este caso, lo hacemos sobre la Gamma y tenemos que especificar que usaremos la función link=log, ya que por defecto R tiene como función link el enlace canónico.

```
#Leemos los datos
>>library(faraway)
>>data(wafer)
>>names(wafer)
>>attach(wafer)

>>summary(wafer)
```

```
#Modelo lineal con transformación logarítmica
>>modelo=lm(log(resist)~.,wafer)
>>summary(modelo)
Call:
lm(formula = log(resist) ~ ., data = wafer)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17572 -0.06222  0.01749  0.08765  0.10841

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.44048    0.05982  90.948  < 2e-16 ***
x1+          0.12277    0.05350   2.295 0.042432 *
x2+         -0.29986    0.05350  -5.604 0.000159 ***
x3+          0.17844    0.05350   3.335 0.006652 **
x4+         -0.05615    0.05350  -1.049 0.316515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.107 on 11 degrees of freedom
Multiple R-squared:  0.8164,Adjusted R-squared:  0.7496
F-statistic: 12.22 on 4 and 11 DF,  p-value: 0.0004915

#Regresión Gamma
>>library(glmnet)
>>mod=glm(resist~.,family=Gamma(link=log),wafer)
>>summary(mod)
Call:
glm(formula = resist ~ ., family = Gamma(link = log), data = wafer)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
```

```
-0.17548  -0.06486   0.01423   0.08399   0.10898

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.44552    0.05856  92.983  < 2e-16 ***
x1+          0.12115    0.05238   2.313 0.041090 *
x2+         -0.30049    0.05238  -5.736 0.000131 ***
x3+          0.17979    0.05238   3.432 0.005601 **
x4+         -0.05757    0.05238  -1.099 0.295248
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.01097542)

    Null deviance: 0.69784  on 15  degrees of freedom
Residual deviance: 0.12418  on 11  degrees of freedom
AIC: 152.91

Number of Fisher Scoring iterations: 4
```

Como aparece en el script ajustamos este modelo de dos formas. La primera es un modelo lineal con interacciones con una transformación logarítmica de la variable respuesta, mientras que la segunda es un GLM Gamma.

Analizamos los coeficientes obtenidos por ambos modelos y vemos que son muy similares al igual que el nivel de significación de las variables explicativas. Además, el error estándar del modelo lineal con transformación logarítmica, 0.06735, es muy semejante con la raíz cuadrada de la dispersión de la regresión Gamma, $\sqrt{0.004524942} = 0.06726769$.

En este caso, debido a que el parámetro de forma ν *es grande entonces la distribución Gamma está bien aproximada por una normal. Además, este modelo lineal es una distribución lognormal con varianza pequeña, con lo cual también ajusta muy bien los datos. Por lo tanto, estos modelos son muy similares. La ventaja del modelo de regresión Gamma es que modela directamente la respuesta y el lognormal trabaja con datos transformados.*

Veamos esta similitud gráficamente

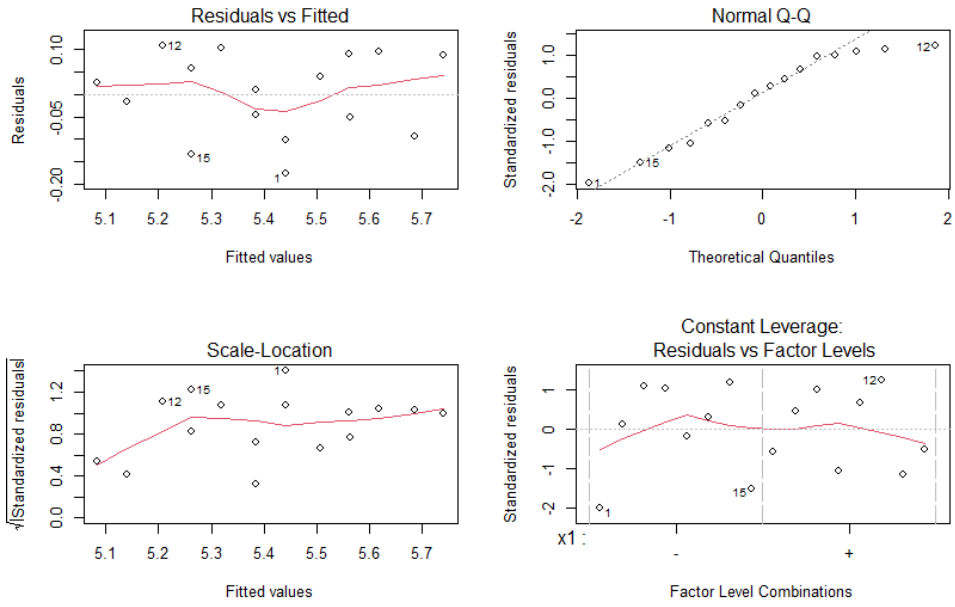

Figura 4.1: Validación del modelo con transformación logarítmica

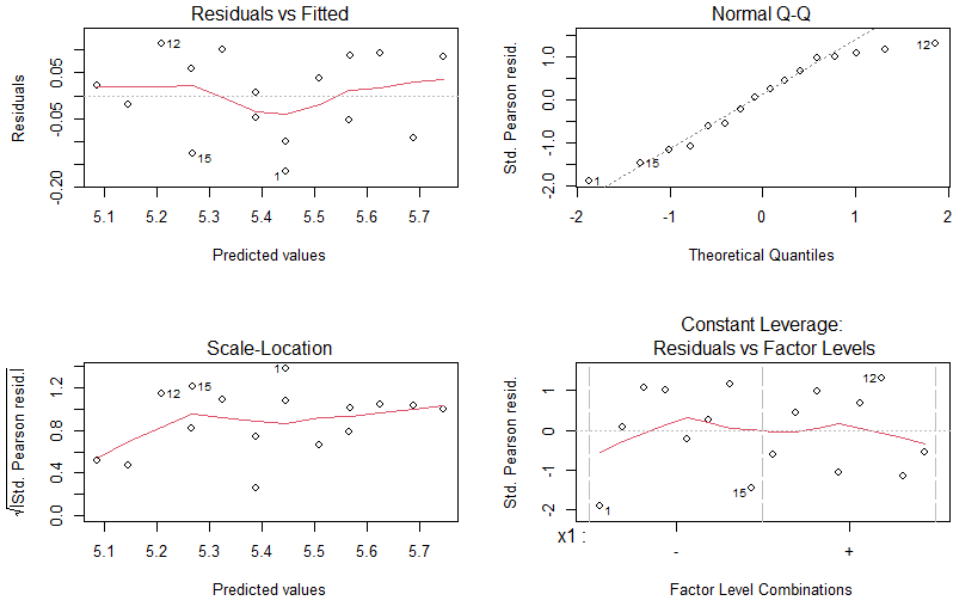

Figura 4.2: Validación del modelo de regresión Gamma

Como podemos observar estos dos modelos diferentes, son prácticamente idénticos gráficamente. Además, también percibimos que cualquiera de estos modelos son adecuados para el conjunto de datos estudiado ya que se cumplen las hipótesis de la validación del modelo. Percibimos que se cumple la homocedasticidad por cómo están distribuidas las observaciones en el gráfico de los residuos y los valores ajustados, la varianza prácticamente no cambia con respecto a los valores ajustados. Además, deducimos que todos los residuos menos los de algunos datos, que se desvían ligeramente, siguen una distribución normal, con lo cual la hipótesis de la normalidad también parece aceptada. Por último, la linealidad se cumple, como podemos observar, por las líneas rojas del gráfico de residuos y valores ajustados y escala y localización.

En conclusión, hemos obtenido dos modelos diferentes que se ajustan muy bien a los datos empleados.

Comandos R: Capítulo 1

En este anexo especificaremos los comandos utilizados en R en el capítulo 1, *Definición y propiedades de la Gamma*. A continuación, podemos observar los comandos empleados para la representación gráfica de las figuras (1.1) y (1.2):

```
#BETA=1
plot(seq(0,5,.01),dgamma(seq(0,5,.01),1,1),lwd=2,col="red",
type="l", xlab="",ylab="")
lines(seq(0,5,.01),dgamma(seq(0,5,.01),2,1),lwd=2,col="green",
type="l")
lines(seq(0,5,.01),dgamma(seq(0,5,.01),3,1),lwd=2,col="blue",
type="l")
legend(3,1,legend=c(expression(alpha==1,alpha==2,alpha==3)),
fill=c("red","green","blue"),bty = "n", cex=1.5)

#ALFA=2
plot(seq(0,5,.01),dgamma(seq(0,5,.01),2,3),lwd=2,col="red",
type="l",xlab="",ylab="")
lines(seq(0,5,.01),dgamma(seq(0,5,.01),2,2),lwd=2,col="green",
type="l")
lines(seq(0,5,.01),dgamma(seq(0,5,.01),2,1),lwd=2,col="blue",
type="l")
legend(3,1,legend=c(expression(beta==3,beta==2,beta==1)),
fill=c("red","green","blue"),bty="n", cex=1.5)
```

Ahora, veamos los comandos empleados para estudiar el ejemplo de lluvias diarias (1.20).

```
datos<-read.csv("datoslluvia.csv",sep=",",stringsAsFactors = T)
```

```
head(datos)
names(datos)
attach(datos)
datos1<-datos$Precipitación..l.m2
x<-c(datos1[1:5],datos1[7:15],datos1[17:33],datos1[35:58])

library(fitdistrplus)

#Veamos que distribución siguen nuestros datos
descdist(x,boot=200)

outy<-fitdist(x,"gamma")
summary(outy)
plot(outy)

outy2<-fitdist(x,"norm")
summary(outy2)
plot(outy2)

outy3<-fitdist(x,"unif")
summary(outy3)
plot(outy3)

#Autocorrelación
Box.test(x, lag = 1, type = c("Box-Pierce", "Ljung-Box"),
fitdf = 0)
```

Los comandos que se utilizan para los datos de lluvias anuales del ejemplo (1.21) son

```
datos<-read.csv("datoslluvia2.csv",sep=",",dec=".",
stringsAsFactors = T)
head(datos)
names(datos)
attach(datos)
```

```
x=datos$Precipitación..l.m2

library(fitdistrplus)

descdist(x,boot = 200)

outy<-fitdist(x,"gamma")
summary(outy)
plot(outy)

outy2<-fitdist(x,"lnorm")
summary(outy2)
plot(outy2)

outy3<-fitdist(x,"exp")
summary(outy3)
plot(outy3)

#Autocorrelación
Box.test(x, lag = 1, type = c("Box-Pierce", "Ljung-Box"),
fitdf = 0)
```

En último lugar, especificaremos los datos empleados en los ejemplos (1.20). En primer lugar, representamos la tabla con los datos de lluvias diarias y la segunda es con los datos de lluvias anuales.

Cuadro 1: Lluvias diarias, medidas en litros por metro cuadrado, en la provincia de A Coruña, en enero y febrero del 2021.

Fecha	Precipitación l/m^2	Fecha	Precipitación l/m^2
01-01-2021	7.80	01-02-2021	1.80
02-01-2021	13.00	02-02-2021	12.60
03-01-2021	6.80	03-02-2021	Ip
04-01-2021	4.20	04-02-2021	10.80
05-01-2021	3.60	05-02-2021	1.40
06-01-2021	Ip	06-02-2021	0.40
07-01-2021	0.20	07-02-2021	20.40
08-01-2021	1.60	08-02-2021	10.00
09-01-2021	2.80	09-02-2021	11.80
10-01-2021	0.00	10-02-2021	8.80
11-01-2021	0.00	11-02-2021	40.00
12-01-2021	0.00	12-02-2021	0.00
13-01-2021	0.00	13-02-2021	0.40
14-01-2021	1.60	14-02-2021	0.00
15-01-2021	0.00	15-02-2021	12.20
16-01-2021	Ip	16-02-2021	0.40
17-01-2021	1.00	17-02-2021	11.00
18-01-2021	0.00	18-02-2021	0.20
19-01-2021	1.40	19-02-2021	8.20
20-01-2021	13.40	20-02-2021	10.60
21-01-2021	14.20	21-02-2021	3.40
22-01-2021	5.60	22-02-2021	0.00
23-01-2021	16.00	23-02-2021	0.00
24-01-2021	20.80	24-02-2021	11.60
25-01-2021	4.80	25-02-2021	5.20
26-01-2021	1.20	26-02-2021	0.00
27-01-2021	0.00	27-02-2021	0.00
28-01-2021	0.40	28-02-2021	0.00
29-01-2021	5.00		
30-01-2021	17.40		
31-01-2021	0.40		

Cuadro 2: Lluvias anuales, medidas en pulgadas, en Ithaca, estado de Nueva York desde 1933 hasta 1882.

Fecha	Precipitación	Fecha	Precipitación
1933	0.44	1958	4.90
1934	1.18	1959	2.94
1935	2.69	1960	1.75
1936	2.08	1961	1.69
1937	3.66	1962	1.88
1938	1.72	1963	1.31
1939	2.82	1964	1.76
1940	0.72	1965	2.17
1941	1.46	1966	2.38
1942	1.30	1967	1.16
1943	1.35	1968	1.39
1944	0.54	1969	1.36
1945	2.74	1970	1.03
1946	1.13	1971	1.11
1947	2.50	1972	1.35
1948	1.72	1973	1.44
1949	2.27	1974	1.84
1950	2.82	1975	1.69
1951	1.98	1976	3.00
1952	2.44	1977	1.36
1053	2.53	1978	6.37
1954	2.00	1979	4.55
1955	1.12	1980	0.52
1956	2.13	1981	0.87
1957	1.36	1882	1.51

Comandos R: Capítulo 2

Este anexo trata sobre los comandos usados en R de capítulo 2, *Inferencia de los parámetros de Gamma*. Comencemos viendo los programas de R usados para el método iterativo de Newton. Estos programas son 3, *GammaNewton.R*, *emvalphaGamma.R* y *estimacionejemplo.R*.

GammaNewton.R

```
GammaNewton = function(muestra,alpha0){
    m = mean(muestra)
    mlog = mean(log(muestra))
    alpha1 = alpha0 *(1+(digamma(alpha0)+log(m/alpha0)-mlog)/
    (1-alpha0*trigamma(alpha0)))
}
```

emvalphaGamma.R

```
emvalphaGamma = function(muestra,iter=TRUE){
    # Si iter=TRUE, imprime en pantalla el error en cada iteracción.
    m = mean(muestra)
    v = var(muestra)
    # Inicio iteracción = estimador alpha(método de los momentos)
    alpha0 = m^2/v
    error = 1
    i = 0
    while(error>1E-8){
        i = i+1
        alpha1 = GammaNewton(muestra,alpha0)
        error = abs(alpha1-alpha0)
        alpha0 = alpha1
```

```
        if (iter==TRUE){writeLines(paste("Iteracion ",i,
        " Error =",error))}
  }
  alpha1
}
```

estimacionejemplo.R

```
#Leemos los datos
datos<-read.csv("datoslluvia.csv",sep=",",stringsAsFactors = T)
head(datos)
names(datos)
attach(datos)
datos1<-datos$Precipitación..l.m2
X<-c(datos1[1:5],datos1[7:15],datos1[17:33],datos1[35:58])

m = mean(X)
n=length(X)
emvalpha = emvalphaGamma(X)
emvalphavector = emvalpha
emvbetavector = m/emvalpha

writeLines(paste('Tamaño muestral n =',n))
writeLines(paste('e.m.v.(alpha) -> Media =',
round(emvalphavector,digits=8)))
writeLines(paste('e.m.v.(beta) -> Media =',
round(emvbetavector,digits=8)))
```

Ahora, sigamos viendo los comandos empleados para encontrar las estimaciones de máxima verosimilitud de α y β con los comandos *mle* que está en la librería *stats4*.

```
library(stats4)
datos<-read.csv("datoslluvia.csv",sep=",",stringsAsFactors = T)
head(datos)
names(datos)
```

```
attach(datos)
datos1<-datos$Precipitación..l.m2
y<-c(datos1[1:5],datos1[7:15],datos1[17:33],datos1[35:58])

#Construimos el negativo de la log-verosimilitud
NegLogLik2=function(alpha,beta)
{-sum(dgamma(y,alpha,beta,log=TRUE))}

#Usamos la función mle del paquete stats4
EMV2=mle(NegLogLik2,start=list(alpha=2,beta=3))
summary(EMV2)

#Intervalo confianza perfil de verosimilitud
confint(EMV2)
```

Ahora, veamos los comandos empleados para el Ejemplo 2.13. Primero generamos los datos y estudiamos los estimadores máximos verosímiles y los intervalos de confianza del método pivotal (2.7) y el asintótico (2.8).

```
library(stats4)
set.seed(123)
x=rgamma(n=50,shape=2,scale=3)

#Construimos el negativo de la log-verosimilitud
NegLogLik=function(alpha,beta){-sum(dgamma(x,alpha,beta,log=TRUE))}

#Estimadores máximo verosímiles
EMV=mle(NegLogLik,start=list(alpha=2,beta=3))
summary(EMV)

#Intervalo confianza perfil de verosimilitud
confint(EMV)

#INTERVALO CONFIANZA (método pivotal)
```

```
alpha=2
n=length(x)
xsuma=sum(x)
s=2*sum(x)

s/(qchisq(0.025,2*n*alpha))
s/(qchisq(0.975,2*n*alpha))

#INTERVALO CONFIANZA (asintótico)

alpha=2
n=length(x)
xmedia=mean(x)
p=((n*alpha^3)/(xmedia^2))^(-1/2)
err=p*qnorm(0.975,0,1)

(xmedia/alpha)-err
(xmedia/alpha)+err
```

Comandos R: Capítulo 3

En esta sección del apéndice, explicitamos los comandos usados para obtener L_n para los datos empleados y las tablas (3.1), (3.2) y (3.3). Como ya se comentó anteriormente, estas tablas se obtienen utilizando las simulaciones de Monte Carlo. Para conseguir los valores de la tabla hay que ir cambiando el parámetro de forma (shape) y el tamaño muestral n.

```
Ns=5000
n=55
beta=40
x=matrix(NA,n,1)
lilliefors=matrix(NA,Ns,1)
probpar=matrix(NA,n,1)
pos=matrix(1: n, n, 1)/n
shape=8
set.seed(123)

for (i in 1:Ns){
  x[,1]=rgamma(n,shape,1/beta)
  A=log(mean(x))-((sum(log(x)))/n)
  alfali=(1/(4*A))*(1+sqrt(1+(4*A/3)))
  betali=mean(x)/alfali
  probpar[,1]=pgamma(sort(x), alfali, 1/betali,lower.tail = TRUE)
  Dmax=max(abs(pos- probpar))
  lilliefors[i,1]=Dmax
}

NKScrit1=quantile(lilliefors, probs=0.99) # 1% nivel significación
```

```
NKScrit5=quantile(lilliefors, probs=0.95) # 5% nivel significación
NKScrit10=quantile(lilliefors, probs=0.90) # 10% nivel significación

format(NKScrit1,digits=3)
format(NKScrit5, digits=3)
format(NKScrit10, digits=3)
```

El estadístico L_n para el conjunto de datos de lluvias diarias se calcula con los siguientes comandos

```
datos<-read.csv("datoslluvia.csv",sep=",",
stringsAsFactors = T)
datos1<-datos$Precipitación..l.m2
x<-c(datos1[1:5],datos1[7:15],datos1[17:33],datos1[35:58])
n=length(x)

fn=matrix(1: n, n, 1)/n
A=log(mean(x))-((sum(log(x)))/n)
alfali=(1/(4*A))*(1+sqrt(1+(4*A/3)));alfali
betali=mean(x)/alfali;betali
f=pgamma(sort(x), alfali, 1/betali, lower.tail = TRUE,
log.p = FALSE)
D=max(abs(fn-f));D
```

El estadístico L_n para las lluvias anuales lo obtenemos de la siguiente forma

```
datos<-read.csv("datoslluvia2.csv",sep=",",
stringsAsFactors = T)
x<-datos$Precipitación..l.m2
n=length(x)

fn=matrix(1: n, n, 1)/n
A=log(mean(x))-((sum(log(x)))/n)
alfali=(1/(4*A))*(1+sqrt(1+(4*A/3)));alfali
betali=mean(x)/alfali;betali
```

```
f=pgamma(sort(x), alfali, 1/betali, lower.tail = TRUE,
log.p = FALSE)
D=max(abs(fn-f));D
```

Bibliografía

[1] I. Arroyo, L. C. Bravo, Dr. Ret. Nat. Humberto, Msc. F. L. Muñoz, Distribuciones Poisson y Gamma: Una Discreta y Continua Relación, *Prospectiva, 12(1)*, pp. 99-107, 2014.

[2] M. H. DeGroot, M. J. Schervish, *Probability and Statistics*, Fourth Edition, Pearson, pp. 381-432.

[3] M. J. García-Ligero, A. Hermoso, J. A. Maldonado, P. Román, F. Torres, *Relación entre la distribución Erlang y la de Poisson*, Grupo de innovación docente CPDYE-UGR, Universidad de Granada. *http* : $//www.ugr.es/\ cdpye/CursoProbabilidad/pdf/P_T05_ErlangPoisson.pdf$

[4] J. Chico, *Estimación de los parámetros de forma y escala de una distribución Gamma* (proyecto de fin de carrera), Universidad de Salamanca, 2010.

[5] *El proceso de Poisson y sus distribuciones asociadas*, Universidad de Valladolid, pp. 1. *http* : $//www.eio.uva.es/\ valentin/ging/Material\ \%20Grado\ \%20pdf\ \%202013_v1 /Tema0009_docum.pdf$.

[6] E. Barrios, *Las distribuciones Gamma y Beta*, pp. 3, 2019. *http* : $//allman.rhon.itam.mx/\ ebarrios/distribuciones/distribucion_GammaBeta.pdf$

[7] K. O. Bowman and L. R. Shenton, *Propierties of estimators for the gamma distribution*, Vol. 89, Marcel Dekker, INC, pp. 48, 1988

[8] D.S. Wilks, *Statistical Methods in the Atmospheric Sciences*, Second Edition, Elsevier, pp. 146-152, 2006.

[9] R. M. Crujeiras, Regresión logística e introducción a los GLM, *Modelos de regresión y análisis multivariante*,pp. 17, 2021.

[10] A. Corberán, Estadística Matemática, *Estadística Matemática*, Universidad de Valencia, 2020.

[11] A. Redchuk, D.G. Soria, Estimación Máximo Verosímil del Parámetro de Forma de la Distribución Gamma, *Revista de la Escuela de Perfeccionamiento en Investigación Operativa*, EPIO, No 18, pp. 28-35, 2000.

[12] G.C. Blain, Revisiting the critical values of the Lilliefors test: towards the correct agrometeorological use of the Kolmogorov-Smirnov framework, *Bragantía*, Vol 73, pp. 192-202, 2014.

[13] H.W. Lilliefors, On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown, *Journal of the American Statistical Association*, Vol 64, No 325, pp. 387-389, 1969.

[14] J.J. Faraway, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Taylor& Francis, pp. 126-155, 2006.

[15] P. McCullagh, J.A. Nelder, *Generalized Linear Models*, Second Edition, Chapman and Hall, pp. 21-41 y 295-296, 1989.