FACULTAD DE MATEMÁTICAS

MÁSTER UNIVERSITARIO EN MATEMÁTICAS

TRABAJO FIN DE MÁSTER

Test de bondad de ajuste de la distribución Poisson

Realizado por:
Manuel Méndez Hurtado

Dirigido por:
Dña. María Dolores Jiménez Gamero

Departamento:
Estadística e Investigación Operativa

Sevilla, Junio 2021

Índice general

Agradecimientos

> ”Lo más terrible se aprende en seguida y lo hermoso nos cuesta la vida”
>
> *Silvio Rodríguez*

Quisiera comenzar transmitiendo mi más sincero agradecimiento a todas las personas que me han ayudado a lo largo de esta etapa.

Primeramente, a mi tutora María Dolores Jiménez por su inestimable ayuda a la hora de planificar y redactar este trabajo, además de la gran cantidad de información de la que me ha dotado para la realización del mismo.

En segundo lugar, quisiera agradecer a mi familia, a mis padres, a mis hermanos y a mi abuela el apoyo constante que me han dado para lograr estos objetivos.

Finalmente, a todos y cada uno de mis amigos, compañeros y, en definitiva, a todas aquellas personas que me han ayudado y apoyado durante este proceso.

A todos ellos, muchas gracias.

Abstract

Unlike the normality tests, the goodness of-fit tests for the Poisson distribution are not present in commonly used software environments for statistical computing, such as R or Python. In this work, we review tests for the Poisson distribution, study the underlying theory for some of them, and numerically compare their powers against a wide range of alternatives. Our main contribution is the development of a package in R to check the poissonity of the data, that calculates the previously analyzed tests, using a parametric bootstrap approximation method to determine the critical points. When a test can be calculated in several ways, it is implemeted the more efficient one, in the sense of requiring less computing time. Finally, we applied these tests to a real data set.

Resumen

A diferencia de los test de normalidad, los test de bondad de ajuste para la distribución Poisson no se encuentran en los principales entornos y/o lenguajes de programación estadísticos como R o Python. En este trabajo, nos centraremos en una profunda revisión teórica de algunos de estos test y compararemos numéricamente su bondad de ajuste a la hipótesis nula y su potencia frente a alternativas. La principal contribución de esta trabajo es el desarrollo de un paquete en el lenguaje R para determinar la Poissonidad de un conjunto de datos, la cual es calculada por los test previamente analizados, utilizando el método de aproximación bootstrap para determinar los puntos críticos. Cuando un test puede ser programado de varias maneras, se implementa la más eficiente, es decir, la forma que requiera un menor coste computacional. Finalmente, aplicamos estos test a un conjunto de datos.

Capítulo 1

Introducción

Los test de bondad de ajuste tienen como objetivo contrastar si un conjunto de datos sigue o no un modelo probabilístico determinado. Este es un aspecto transcendental en cualquier análisis de datos que haga uso de hipótesis distribucionales. Multitud de autores han tratado este área de la Estadística cuando se les presupone a los datos una distribución continua. Sin embargo, la literatura sobre test de bondad de ajuste para familias de distribuciones discretas no es tan abundante. No obstante, diversos autores han propuesto test para la distribución de Poisson.

La principal dificultad que presenta este problema es que los test de Poissonidad (llamaremos test de Poissonidad a los test cuya hipótesis nula es que un conjunto determinado de datos sigue una distribución de Poisson) normalmente no siguen bajo la hipótesis nula una distribución asintótica conocida, por tanto a la hora de calcular el p-valor de un conjunto de datos se suele recurrir a una estimación usando el método bootstrap de la distribución nula de estadístico del contraste.

Sea $X_1, \ldots, X_n$ una muestra aleatoria de X, es decir, n variables aleatorias independientes e idénticamente distribuidas (i.i.d.) como X, que toma valores en los enteros no negativos $X \in \mathbb{N}_0$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. A estas variables se les suele denominar variables *de conteo*. A partir de la muestra se desea contrastar si X se distribuye según una ley de Poisson, esto es, contrastar.

$$\begin{aligned} H_0 : &\quad X \sim Pois(\lambda), \quad \text{para algún } \lambda > 0, \\ H_1 : &\quad X \nsim Pois(\lambda), \quad \forall \lambda > 0. \end{aligned} \tag{1.1}$$

La memoria está organizada como sigue. En el Capítulo 2 recopilamos una serie de resultados previos que serán importante conocer por su aplicación en capítulos posteriores. En el Capílutlo 3 recopilamos una serie de test para el contraste (1.1). En el Capítulo 4 estudiamos en profundidad el test propuesto por Baringhaus y Henze (1992) [2]. En el Capítulo 5 estudiamos en profundidad varios test propuestos en Puig y Weiß (2020) [16]. En el Capítulo 6 estudiaremos en profundidad el test propuesto en

Székely y Rizzo (2004) [20].

En el Capítulo 7 presentamos los resultados de un estudio de simulación donde comparamos numéricamente, en términos de nivel y potencia, los test listados en el Capítulo 3. En el Capítulo 8 presentamos un estudio computacional para determinar qué alternativa de programación es más eficiente en términos de tiempo requerido para su cálculo en varios test. En el Capítulo 9 aplicamos los test vistos en el Capítulo 3 a un conjunto de datos reales. En el Capítulo 10 se presentan varios test creado a colación de este trabajo en el programa R el cual contiene la mayoría de los test vistos en los Capítulos anteriores.

La memoria concluye presentado un resumen y apuntando posibles líneas de investigación en el Capítulo 11.

La bibliografía consultada para la elaboración de esta memoria se muestra al final.

Capítulo 2

Preliminares

Las siguientes definiciones y resultados han sido tomados de los siguientes libros: *'Inference and Predictions in Large Dimension'*, (Dennis Bosq, 2007) [5]; *'An introduction to probability and statistics'* (Rohatgi y Ehsanes Saleh, 2015) [17]; *'Inference for Functional Data with Applications'* (Kokoszka y Lajos, 2012) [11]; *'Approximation Theorems of Mathematical Statistic'* (Selfing, 1980) [19] y *'Convergence of Probability Measures'* (Bilingsley, 1968) [3].

En adelante supondremos que todas las variables aleatorias a las que se hace referencia en las definiciones y resultados están definidas sobre un mismo espacio de probabilidad $(\Omega, \mathbb{A}, P)$.

Definiciones y enunciados para variables aleatorias

Definición 1. Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias. Decimos que X_n converge en probabilidad a X si

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \ \forall \epsilon > 0$$

Usaremos la notación $X_n \xrightarrow{P} X$.

Definición 2. Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias y X una variable aleatoria (que no se encuentran necesariamente en un mismo espacio de probabilidad) con funciones de distribución $\{F_n(\cdot)\}_{n\geq 1}$ y $F(\cdot)$ respectivamente. Decimos que X_n converge en distribución a X si

$$lim_{n\to\infty} F_n(t) = F(t),$$

para todo punto t de continuidad de F.

Usaremos la notación $X_n \xrightarrow{\mathcal{L}} X$.

Definición 3. Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias. Decimos que X_n converge casi seguro a X si

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

Usaremos la notación $X_n \stackrel{c.s}{\to} X$.

Teorema 1. ***Teorema de la aplicación continua:*** *Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorios con valores en un espacio métrico S. Sea $f : S \to S'$ una función continua en los valores de X, entonces:*

- $X_n \stackrel{P}{\to} X \Rightarrow g(X_n) \stackrel{P}{\to} g(X)$
- $X_n \stackrel{c.s}{\to} X \Rightarrow g(X_n) \stackrel{c.s}{\to} g(X)$
- $X_n \xrightarrow{\mathcal{L}} X \Rightarrow g(X_n) \xrightarrow{\mathcal{L}} g(X)$

Teorema 2. ***Teorema Central del Límite:*** *Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias independientes idénticamente distribuidas con media μ y varianza $\sigma^2 < \infty$, entonces:*

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{L}} N(0,1)$$

Definición 4. Función generatriz de probabilidad: Sea X una variable aleatoria que toma valores enteros no negativos. Sea $p_k = P\{X = k\}$. Llamaremos función generatriz de probabilidad de X a la función definida por

$$g(t) = \sum_{k=0}^{\infty} p_k t^k, \;\; k \in \mathbb{N}_0, \;\; t \in [-1,1].$$

Teorema 3. *Sea $g(t)$ una función generatriz de probabilidad, se cumple:*

1. $g(1) = 1$
2. *La serie determinada por $\sum_{k=0}^{\infty} p_k t^k$ es uniforme y absolutamente convergente en $|t| \leq 1$*
3. *g es una función continua en t.*
4. *$g(t)$ puede ser representada solamente de una forma como serie de potencia. (Unicidad.)*
5. *Sea X una distribución aleatoria Poisson de parámetro λ, con $P\{X = k\} = e^{-\lambda}\frac{\lambda^k}{k!}$, $k = 0,1,2,\ldots$, entonces:*

$$g(t) = \sum_{k=0}^{\infty} (t\lambda)^k \frac{e^{-\lambda}}{k!} = e^{-\lambda(1-t)}$$

6. *Sean X e Y dos variables aleatorias independientes con función generatriz de probabilidad g_X y g_Y respectivamente, se tiene que*

$$g_{aXbY}(s) = g_X(as)g_Y(bs), \;\; a,b \in \mathbb{R}.$$

Teorema 4. ***Teorema de Slutsky:*** *Sean $\{X_n\}_{n\geq 1}$, $\{Y_n\}_{n\geq 1}$ dos sucesiones de variables aleatorias independientes tales que $X_n \xrightarrow{\mathcal{L}} X$ y $Y_n \xrightarrow{P} c$, siendo c una constante finita, entonces:*

- $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$.
- $X_n Y_n \xrightarrow{\mathcal{L}} cX$.
- $X_n / Y_n \xrightarrow{\mathcal{L}} X/c$ *si* $c \neq 0$.

Definición 5. Test consistente: Se dice que un test es consistente para una alternativa determinada si la potencia del test (probabilidad de que la hipótesis nula se rechace cuando la hipótesis alternativa es verdadera) tiende a uno cuando el tamaño de la muestra tiende a infinito.

Teorema 5. ***Método Delta:*** *Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias independientes idénticamente distribuidas. Si $\sqrt{n}(X_n - \theta) \xrightarrow{\mathcal{L}} N_p(0, \Sigma)$. Sea $g : \mathbb{R}^p \to \mathbb{R}$ una función continua en θ que cumple que $g'(\theta) \neq 0$, entonces se cumple que*

$$\sqrt{n}(g(X_n - g(\theta))) \xrightarrow{\mathcal{L}} N_1(0, g'(\theta)^t \Sigma g'(\theta))$$

Definiciones y enunciados para funciones aleatorias

Definición 6. El σ álgebra de Borel en un espacio métrico D es el menor σ-álgebra que contiene a todos los conjuntos abiertos (y por tanto también los cerrados). Una función definida en un espacio métrico es Borel-medible si es medible en todos los σ-álgebras de Borel. Una función Borel-medible $X : \Omega \to D$ definida en un espacio de probabilidad (Ω, U, P) se dice que es un elemento aleatorio con valores en D. Una función aleatoria es un tipo de elemento aleatorio que consiste en una función elegida aleatoriamente de una familia de funciones.

Definición 7. Espacio de funciones $\mathbf{L}^2$: el espacio $L^2 = \mathbf{L}^2([0,1])$ es el conjunto de funciones x reales y medibles definidas en $[0,1]$ tales que: $\int_0^1 x^2(t)dt < \infty$. El espacio $\mathbf{L}^2$ es un espacio de Hilbert separable con producto escalar:

$$\langle x, y \rangle = \int_0^1 x(t)y(t)dt.$$

Definición 8. Operadores en L^2:

- $\mu(t) = E[X(t)]$ (función media)
- $c(t,s) = E[(X(t) - \mu(t))\,(X(s) - \mu(s))]$ (función covarianza)
- $C = E[\langle (X - \mu), \cdot \rangle (X - \mu)]$ (operador covarianza)

.

Definición 9. Función Gaussiana: Una función aleatoria X se dice gaussiana si para cada $y \in L^2$ la función aleatoria $\langle X, y\rangle$ se distribuye según una normal cuyos parámetros son el la función media, $\mu(t)$ y el operador covarianza (C).

Definición 10. Convergencia en Ley de elementos aleatorios. Sea X un elemento aleatorio en Ω con valores en S y siendo P su distribución, se dice que una sucesión $\{X_n\}_{n\geq 1}$ de elementos aleatorios convergen en ley al elemento aleatorio X ($X_n \xrightarrow{\mathcal{L}} X$), si las distribuciones P_n de X_n convergen débilmente a la distribución P de X.

Teorema 6. ***Teorema de equivalencias en la convergencia en ley.*** *Los siguientes cinco enunciados son equivalentes:*

- $X_n \xrightarrow{\mathcal{L}} X.$
- $lim_n E(f(X_n)) = E(f(X))$, *para toda f acotada, uniformemente continua y real.*
- $\limsup_n P(X_n \in V) \leq P(X \in V)$, *para todo V conjunto cerrado.*
- $\liminf_n P(X_n \in G) \geq P(X \in G)$, *para todo conjunto G abierto.*
- $\lim_n P(X_n \in A) = P(X \in A)$, *para todo conjunto A continuo en X.*

Teorema 7. ***Teorema central del límite en espacios de Hilbert:*** *Sea $\{X_n\}_{n\geq 1}$ una sucesión de elementos aleatorios que toman valores en un espacio de Hilbert separable, si $E||X_1||^2 < \infty$, $EX_1 = m$ y $C_{x1} = C$, siendo C la función de covarianzas, entonces:*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - m) \xrightarrow{\mathcal{L}} N(0, C)$$

Teorema 8. ***Teorema de Slutsky para datos funcionales:*** *Sean Sea $\{X_n\}_{n\geq 1}$, $\{Y_n\}_{n\geq 1}$ sucesiones de funciones aleatorias. Si X_n converge en ley a una función aleatoria X e Y_n converge en probabilidad a c, siendo c una constante, entonces:*

- $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$
- $X_n Y_n \xrightarrow{\mathcal{L}} cX$
- $X_n / Y_n \xrightarrow{\mathcal{L}} X/c$

Órdenes

Definición 11. O_P**(1) y** o_P**(1):**

- Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables, y sea $\{a_n\}_{n\geq 1}$ una sucesión de constantes reales, $X_n = o_P(a_n)$ indica que el conjunto de valores X_n/a_n convergen a 0 en probabilidad. Nótese que $X_n = o_P(a_n)$ implica que $X_n/a_n = o_P(1)$.

- Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias, siendo $\{F_n\}_{n\geq 1}$ sus respectivas funciones de distribución. Se dice que $\{X_n\}_{n\geq 1}$ está acotada en probabilidad si para todo $\epsilon > 0$, existe un M_ϵ y un N_ϵ tales que

  $$F_n(M_\epsilon) - F_n(-M_\epsilon) > 1 - \epsilon, \forall n > N_\epsilon.$$

  Para denotarlo usaremos la notación $X_n = O_P(1)$. Es fácilmente observable que $X_n \xrightarrow{\mathcal{L}} X \Rightarrow X_n = O_P(1)$.

- Generalizando los casos anteriores, para dos secuencias de variables aleatorias $\{U_n\}_{n\geq 1}$ y $\{V_n\}_{n\geq 1}$, la notación $U_n = O_P(V_n)$ indica que $\{U_n/V_n\}$ es $O_P(1)$.

  La notación $U_n = o_P(V_n)$ indica que $\{U_n/V_n\} \xrightarrow{P} 0$.

  Por otro lado, $U_n = o_P(V_n) \rightarrow U_n = O_P(V_n)$.

Cálculos previos

Lema 1. *El momento de orden 2 de la distribución Poisson es* $E[X^2] = \frac{\sum_{i=1}^{\infty} x_i^2}{n}$*, siendo* $X_1, \ldots, X_n$ *una muestra de una variable aleatoria de conteo* X *que sigue una distribución* $Pois(\lambda)$*. A continuación vamos a probar que su valor es* $E[X^2] = \lambda^2 + \lambda$*.*

Demostración *Tenemos que:*

$$E[X^2] = E[X(X-1) + X] = E[X(X-1)] + E[X] \tag{2.1}$$

Sabemos, por definición de la Poisson, que $E[X] = \lambda$*, por tanto vamos nos queda calcular el valor de la primera componente de 2.1.*

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1)f(x),$$

siendo $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ *la función de probabilidad de la distribución Poisson.*

Por tanto tenemos:

$$\sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty} x(x-1)\frac{\lambda^x}{x!} \tag{2.2}$$

Sea $t > 2$ un entero, $t! = t(t-1)(t-2)!$. Aplicando esto en 2.2 y multiplicando y dividiendo dicha expresión por λ^2 nos queda que:

$$e^{-\lambda}\sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!}\frac{\lambda^2}{\lambda^2} = e^{-\lambda}\lambda^2\sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} \tag{2.3}$$

Sabemos que $\sum_{i=0}^{\infty} \frac{x^i}{i!} = e^x$. Aplicando esto y haciendo el cambio de variable $y = x-2$ en 2.3 nos queda:

$$\lambda^2 e^{-\lambda}\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \lambda^2 e^{-\lambda}e^{\lambda} = \lambda^2 \tag{2.4}$$

Por tanto nos queda que

$$E[X^2] = E[X(X-1)+X] = E[X(X-1)] + E[X] = \lambda^2 + \lambda.$$

□

Lema 2. *El momento de orden 3 de la distribución Poisson es $E[X^3] = \frac{\sum_{i=1}^{\infty} x_i^3}{n}$, siendo $X_1, \ldots, X_n$ una muestra de una variable aleatoria de conteo X que sigue una distribución $Pois(\lambda)$. A continuación vamos a probar que su valor es $E[X^3] = \lambda^3 + 3\lambda^2 + \lambda$.*

Demostración *Tenemos que:*

$$E[X^3] = E[X(X-1)(X-2)+3X(X-1)+X] = E[X(X-1)(X-2)]+3E[X(X-1)]+E[X] \tag{2.5}$$

Sabemos, por definición de la Poisson, que $E[X] = \lambda$ y vimos en el Lema 1 que $E[X(X-1)] = \lambda^2$, por tanto vamos nos queda calcular el valor de la primera componente de 2.5.

$$E[X(X-1)(X-2)] = \sum_{x=0}^{\infty} x(x-1)(x-2)f(x),$$

siendo $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ la función de probabilidad de la distribución Poisson.

Por tanto tenemos:

$$\sum_{x=0}^{\infty} x(x-1)(x-2)\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty} x(x-1)(x-2)\frac{\lambda^x}{x!} \tag{2.6}$$

Sea $t > 3$ un entero, $t! = t(t-1)(t-2)(t-3)!$. Aplicando esto en 2.6 y multiplicando y dividiendo dicha expresión por λ^3 nos queda que:

$$e^{-\lambda}\sum_{x=3}^{\infty}\frac{\lambda^x}{(x-3)!}\frac{\lambda^3}{\lambda^3}=e^{-\lambda}\lambda^3\sum_{x=3}^{\infty}\frac{\lambda^{x-3}}{(x-3)!} \tag{2.7}$$

Sabemos que $\sum_{i=0}^{\infty}\frac{x^i}{i!}=e^x$. *Aplicando esto y haciendo el cambio de variable* $y=x-3$ *en* 2.7 *nos queda:*

$$\lambda^3 e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!}=\lambda^3 e^{-\lambda}e^{\lambda}=\lambda^3 \tag{2.8}$$

Por tanto nos quedaría

$$E[X^3]=E[X(X-1)(X-2)+3X(X-1)+X]=E[X(X-1)(X-2)]+3E[X(X-1)]+E[X]$$
$$=\lambda^3+3\lambda^2+\lambda.$$

□

Lema 3. *El momento de orden 4 de la distribución Poisson es* $E[X^4]=\frac{\sum_{i=1}^{\infty}x_i^4}{n}$, *siendo* $X_1,\ldots,X_n$ *una muestra de una variable aleatoria de conteo* X *que sigue una distribución* $Pois(\lambda)$. *A continuación vamos a probar que su valor es* $E[X^4]=\lambda^4+6\lambda^3+7\lambda^2+\lambda$.

Demostración *Tenemos que:*

$$\begin{aligned}E[X^4]&=E[X(X-1)(X-2)(X-3)+6X(X-1)(X-2)+7X(X-1)+X]\\&=E[X(X-1)(X-2)(X-3)]+6E[X(X-1)(X-2)]+7E[X(X-1)]+E[X]\end{aligned} \tag{2.9}$$

Sabemos por la definición de la Poisson que $E[X]=\lambda$, *vimos en el Lema* 1 *que* $E[X(X-1)]=\lambda^2$ *y vimos en el Lema* 2 *que* $E[X(X-1)(X-2)]=\lambda^3$, *por tanto nos queda calcular el valor de la primera componente de* 2.9.

$$E[X(X-1)(X-2)(X-3)]=\sum_{x=0}^{\infty}x(x-1)(x-2)(x-3)f(x),$$

siendo $f(x)=\frac{e^{-\lambda}\lambda^x}{x!}$ *la función de probabilidad de la distribución Poisson.*

Por tanto tenemos:

$$\sum_{x=0}^{\infty}x(x-1)(x-2)(x-3)\frac{e^{-\lambda}\lambda^x}{x!}=e^{-\lambda}\sum_{x=0}^{\infty}x(x-1)(x-2)(x-3)\frac{\lambda^x}{x!} \tag{2.10}$$

Sea $t>4$ *un entero cualquiera,* $t!=t(t-1)(t-2)(t-3)(t-4)!$. *Aplicando esto en* 2.10 *y multiplicando y dividiendo dicha expresión por* λ^4 *nos queda que:*

$$e^{-\lambda}\sum_{x=4}^{\infty}\frac{\lambda^x}{(x-4)!}\frac{\lambda^4}{\lambda^4}=e^{-\lambda}\lambda^4\sum_{x=4}^{\infty}\frac{\lambda^{x-4}}{(x-4)!} \tag{2.11}$$

Sabemos que $\sum_{i=0}^{\infty}\frac{x^i}{i!}=e^x$*. Aplicando esto y haciendo el cambio de variable* $y=x-4$ *en* 2.11 *nos queda:*

$$\lambda^4 e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!}=\lambda^4 e^{-\lambda}e^{\lambda}=\lambda^4 \tag{2.12}$$

Por tanto nos quedaría

$$E[X^4]=E[X(X-1)(X-2)(X-3)+6X(X-1)(X-2)+7X(X-1)+X]$$
$$=E[X(X-1)(X-2)(X-3)]+6E[X(X-1)(X-2)]+7E[X(X-1)]+E[X]$$
$$=\lambda^4+6\lambda^3+7\lambda^2+\lambda$$

□

Capítulo 3

Test para la ley Poisson

3.1. Introducción

En este capítulo se muestra una recopilación de test para el contraste (1.1). Específicamente, consideraremos tres tipos de test: test basados en los momentos, test basados en la función de distribución y test basados en la función generatriz de probabilidad.

Una forma común de construir test de bondad de ajuste es utilizando una caracterización de la familia de distribuciones considerada. En este capítulo veremos algunas caracterizaciones de la distribución Poisson que nos llevarán a describir varios test de bondad de ajuste para la misma.

3.2. Test basado en los momentos

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable de conteo X que sigue una distribución Pois(λ) se tiene que

$$E(\bar{X}_n) = \lambda$$

$$E(S_c^2) = \lambda$$

siendo $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$

siendo $S_c^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$.

Como λ es un estimador consistente de la media y de la varianza de una distribución Poisson, tenemos que el índice de dispersión,

$$D = S_c^2 / \bar{X}_n,$$

para una distribución Poisson debe estar próximo a 1.

Teorema 9. *Sean $X_1, \dots, X_n$ i.i.d. de $X \sim Pois(\lambda)$, para algún $\lambda > 0$, entonces*

$$U = \left(\frac{S_c^2}{\bar{X}_n} - 1\right)\sqrt{\frac{n-1}{2(1-1/n\bar{X}_n)}} \xrightarrow{\mathcal{L}} Z, \tag{3.1}$$

donde $Z \sim N(0,1)$.

Demostración La demostración tiene dos partes:

(a) En primer lugar vamos a demostrar que

$$W_n = \sqrt{n}\left(\frac{S^2}{\bar{X}_n} - 1\right) \xrightarrow{\mathcal{L}} Z_1, \tag{3.2}$$

donde $Z_1 \sim N(0, \sigma_1^2)$, $\sigma_1^2 = 2$ y $S^2 = \frac{1}{n}\sum_{j=1}^n X_j = \frac{n-1}{n}S_c^2$.

(b) Nótese que $U = A_n \frac{1}{\sqrt{2}} W_n$, con

$$A_n^2 = \frac{n}{n-1}\frac{\bar{X}_n}{\bar{X}_n - 1/n}.$$

Posteriormente veremos que

$$A_n \xrightarrow{P} 1. \tag{3.3}$$

La convergencia en (3.1) se sigue de (3.2), (3.3) y el Teorema de Slutsky.

Parte (a) Sean

- $m_1 = \frac{\sum_{i=1}^n X_i}{n}$,
- $m_2 = \frac{\sum_{i=1}^n X_i^2}{n}$.

Se tiene que

- $E[m_1] = E[X_i] = \lambda$,
- $E[m_2] = Var(X_i) + E[X_i^2] = \lambda + \lambda^2$.

Por el Teorema central del límite se tiene que

$$\sqrt{n}\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix} - \begin{bmatrix} \lambda \\ \lambda + \lambda^2 \end{bmatrix}\right) \xrightarrow{\mathcal{L}} N_2(0, \Sigma) \tag{3.4}$$

siendo

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix}$$

una matriz formada por los siguientes componentes (en el Capítulo 2 en la Sección 2 vemos cómo proceden los cálculos de cada uno de los momentos de la distribución Poisson):

- $\sigma_{1,1} = E[X^2] - E^2[X] = \lambda^2 + \lambda - \lambda^2 = \lambda$,
- $\sigma_{2,2} = E[X^4] - E^2[X^2] = (\lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda) - (\lambda^4 + 2\lambda^3 + \lambda^2) = 4\lambda^3 + 6\lambda^2 + \lambda$,
- $\sigma_{1,2} = \sigma_{2,1} = E[X^3] - E[X]E[X^2] = (\lambda^3 + 3\lambda^2 + \lambda) - (\lambda^2 + \lambda^3) = 2\lambda^2 + \lambda$.

Sea

$$\begin{aligned} g : (0, +\infty)^2 &\rightarrow \mathbb{R} \\ (x, y) &\rightarrow g(x, y) = \frac{y - x^2}{x}. \end{aligned}$$

Nótese que

$$\begin{aligned} g(m_1, m_2) &= \frac{S^2}{\bar{X}_n}, \\ g(\theta) &= 1, \end{aligned}$$

siendo

$$\theta = \begin{bmatrix} E[X] \\ E[X^2] \end{bmatrix} = \begin{bmatrix} \lambda \\ \lambda^2 + \lambda \end{bmatrix}.$$

Aplicando el método Delta se sigue que

$$W_n = \sqrt{n}\,(g(m_1, m_2) - g(\theta)) \xrightarrow{\mathcal{L}} N(0, \sigma^2),$$

con

$$\sigma^2 = g'(\theta)^\top \Sigma g'(\theta). \tag{3.5}$$

La derivada de g es:

$$g'(x, y) = \begin{pmatrix} \frac{\partial}{\partial x} g(x, y) \\ \frac{\partial}{\partial y} g(x, y) \end{pmatrix} = \begin{pmatrix} \frac{-x^2 - y}{x^2} \\ \frac{1}{x} \end{pmatrix}$$

Evaluando este resultado en θ queda,

$$g'(\lambda, \lambda^2 + \lambda) = \begin{bmatrix} (-2\lambda^2 - \lambda)/\lambda^2 \\ 1/\lambda \end{bmatrix}.$$

Sustituyendo en 3.5 se obtiene que $\sigma^2 = 2$. Esto concluye la demostración de la parte (a).

Parte (b) Como $\frac{n}{n-1} \rightarrow 1$, $\frac{1}{n} \rightarrow 0$ y, por la ley fuerte de los grandes números, $\bar{X}_n \rightarrow E(X) = \lambda$ casi seguro, por el teorema de la función continua se sigue que (3.3) es cierto. Esto concluye la demostración de la parte (b).

□

A diferencia de la mayoría de los test que veremos a continuación, este puede aproximarse, como acabamos de ver, mediante una distribución $N(0,1)$. La región crítica para un nivel de significación bilateral $\alpha = 0.05$ es:

$$RC_{\alpha=0.05} = |u| > 1.96.$$

Ya que el cálculo de de puntos críticos puede hacerse a partir de la distribución normal, usaremos para el cálculo de los mismos en la Sección 10.3 tanto este método a partir de la normal como un bootstrap paramétrico propuesto por Henze (1996) [8]

3.3. Test basados en la función de distribución empírica

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable de conteo X que toma valores enteros no negativos. Llamaremos F a la función de distribución, desconocida, de X y $F(k;\lambda) = \exp^{-\lambda} \sum_{j=0}^{k} \frac{\lambda^j}{j!}, k = 0, 1, 2, \ldots$ a la función de distribución de una Poisson de media λ.

Sea $\hat{\lambda}_n = \frac{1}{n} \sum_{j=1}^{n} X_j = \bar{X}_n$ el estimador de máxima verosimilitud de la media de la distribución y $f(j;\lambda) = \exp^{-\lambda} \frac{\lambda^j}{j!}$ la función de probabilidad de una Poisson de media λ.

Sea $F_n(k) = \frac{1}{n} \sum_{j=1}^{n} 1\{X_j \leq k\}, k = 0, 1, 2, \ldots$ la función de distribución empírica asociada a $X_1, \ldots, X_n$. $F_n(k)$ es un estimador consistente de la función de distribución de X, pues sabemos por la ley de los grandes números, que al ser $F_n(k)$ una media aritmética se cumple:

$$F_n(k) \xrightarrow{c.s.} F(k), \text{para cada } \ k \in \mathbb{N}_0$$

Como $F_n(k,\lambda)$ es una función continua de λ y, por la ley fuerte de los grandes números, $\hat{\lambda}_n \xrightarrow{c.s.} \lambda = E(X)$, se tiene que

$$F(k;\hat{\lambda}_n) \xrightarrow{c.s.} F(k;\lambda), \text{para cada } \ k \in \mathbb{N}_0.$$

Por tanto, si X sigue una distribución Poisson de parámetro λ, $F_n(k)$ y $F(k,\hat{\lambda})$ convergen al mismo límite, o lo que es lo mismo

$$F_n(k) - F(k,\hat{\lambda}) \xrightarrow{c.s.} 0, \text{para cada } \ k \in \mathbb{N}_0$$

- Henze (1996) [8] propone, basándose en la proximidad que habría, caso que X siguiese una ley Poisson de parámetro λ, entre el estimador de la función de distribución de X, $F_n(k)$ y la función de distribución de una Poisson de media $\hat{\lambda}_n$, rechazar la hipótesis nula para valores grandes de los siguientes estadísticos:

$$K_n = \sup_{k \geq 0} \sqrt{n} |F_n(k) - F(k,\hat{\lambda}_n)|, \tag{3.6}$$

$$C_n = n \sum_{k=0}^{\infty} [F_n(k) - F(k;\hat{\lambda}_n)]^2 f(k,\hat{\lambda}_n),$$

$$C_n^{\star} = n \sum_{k=0}^{\infty} [F_n(k) - F(k;\hat{\lambda}_n)]^2 f_n(k),$$

siendo $f_n(k) = F_n(k) - F_n(k-1) = \frac{1}{n}\sum_{j=1}^{n} 1\{X_j = k\}$.

- Trabajando en el espacio l_1 de las series que satisfacen $\sum_{k\geq 0} |x_k| < \infty$, Klar (2000) [9] propone rechazar la hipótesis nula para valores grandes del siguiente estadístico:

$$L_n = \sum_{k\geq 0} \sqrt{n}|F_n(k) - F(k,\hat{\lambda}_n)|.$$

- Sea $\eta(t) = E(X-t)^+ = \int_t^{\infty}(1-F(x))dx$ la función de distribución integrada, el siguiente estadístico está basado en la función de distribución empírica integrada, definida como

$$\eta_n(t) = \int_t^{\infty}(1-F_n(x))dx = \frac{1}{n}\sum_{j=1}^{n}(X_j - t)1\{X_j > t\}.$$

Sea $\hat{\eta}_n(t) = \int_t^{\infty}(1-F(x;\hat{\lambda}_n))dx$ la función de distribución integrada de una Poisson de parámetro $\hat{\lambda}_n$. Klar (2000) [9] propone, de forma semejante a K_n pero utilizando las funciones integradas rechazar la hipótesis nula para valores grandes del siguiente estadístico:

$$I_n = \sup_{t\geq 0} \sqrt{n}|\eta_n(t) - \hat{\eta}_{n(t)}|. \tag{3.7}$$

- Sea $\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} |k - x_i|$. Sea $\hat{F}_X(0) = \hat{f}_X(0) = (\hat{m}_1 + 1 - \hat{\lambda})/2$. Sean $\hat{f}_X$ y $\hat{F}_X$ determinados por la siguiente fórmula recursiva:

$$\hat{f}_X(k) = \frac{\hat{m}_{k+1} - (k+1-\hat{\lambda})(2\hat{F}_X(k-1)-1)}{2(k+1)}.$$

Sean X e Y dos variables aleatorias discretas que toman valores no negativos enteros, siendo $E[X] < \infty$ y $E[Y] < \infty$. Entonces X e Y están idénticamente distribuidas si y solo si

$$E_X[k-X] = E_Y[k-Y]$$

para todo k entero no negativo.

Székely y Rizzo (2004) [20] se basan en esto a la hora de proponer rechazar la hipótesis nula para valores grandes del siguiente test:

$$M_n = n \sum_{k=0}^{\infty} [\hat{F}_X(k) - F(k; \hat{\lambda}_n)]^2 f(k, \hat{\lambda}_n) \tag{3.8}$$

Al no ser ni la distribución nula ni la distribución asintótica de los test de esta sección tratables, Henze (1996) [8] propone el cálculo de los puntos críticos de estos test mediante un bootstrap paramétrico en la Sección 10.3.

3.4. Test basados en la función generatriz de probabilidad

Vamos a dividir esta sección en dos conjuntos de test, ambos basados en la función generatriz de probabilidad. El primer conjunto se basa en la caracterización de la Poisson a partir de la proximidad entre la función generatriz de probabilidad empírica y la función generatriz de probabilidad poblacional. A este conjunto de test los llamaremos ”Test propuestos por Baringhaus y otros”, pues es este autor junto a otros quienes los proponen. El segundo conjunto se basa en una caracterización de la Poisson a partir del operador α-thinning. A este conjunto de test los llamaremos ”Test propuestos por Puig y Weiß”, pues son estos dos autores quienes lo proponen.

3.4.1. Test propuestos por Baringhaus y otros

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable de conteo X que toma valores enteros no negativos con función generatriz de probabilidad $g(s)$.

$$g(s) = E[s^X] = \sum_{k=0}^{\infty} P(X = k) s^k,$$

con $|s| \leq 1$.

Sea $g_n(s)$ la función generatriz de probabilidad empírica asociada a $X_1, \ldots, X_n$, definida como

$$g_n(s) = \frac{1}{n} \sum_{j=1}^{n} s^{X_j}.$$

Llamamos $g(s; \lambda) = \sum_{k=0}^{\infty} f(k; \lambda) s^k = \exp^{\lambda(s-1)}$ a la función generatriz de probabilidad de una Poisson de parámetro λ.

Por la ley fuerte de los grandes números, al ser $g_n(s)$ una media aritmética:

$$g_n(s) \xrightarrow{c.s.} g(s), \forall s \in [-1, 1].$$

Como $g(s;\lambda)$ es una función continua en λ y por la ley de fuerte de los grandes números $\hat{\lambda}_n \xrightarrow{c.s.} \lambda$, se tiene que

$$g(s,\lambda) \xrightarrow{c.s.} g(s,\hat{\lambda}_n), \forall s \in [-1,1].$$

Por tanto, si X sigue una distribución Poisson, $g_n(s)$ y $g(s,\hat{\lambda}_n)$ convergen al mismo límite, es decir,

$$g_n(s) - g(s,\hat{\lambda}_n) \xrightarrow{c.s.} 0, \forall s \in [-1,1].$$

Los siguientes estadísticos del test están basados en

$$G_n(s) = \sqrt{n}(g_n(s) - g(s;\hat{\lambda}_n)), \ \ s \in [0,1],$$

que se basa a su vez en la proximidad que habría caso de que X siguiese una distribución Poisson entre el estimador de la función generatriz de probabilidad de X y la función de distribución de una ley Poisson de media $\hat{\lambda}_n$.

- Rueda et al. (1991) [18] proponen, basándose en $G_n(s)$, rechazar la hipótesis nula para valores grandes del siguiente estadístico:

$$R_n = \int_0^1 G_n^2(s)ds. \tag{3.9}$$

- Baringhaus et al. (2000) [1] realizan una modificación del test anterior, proponiendo rechazar la hipótesis nula para valores grandes del siguiente estadístico:

$$R_{n,a} = \int_0^1 G_n^2(s)s^a ds, \tag{3.10}$$

  donde $a \geq 0$.

- Baringhaus y Henze (1992) [2], basándose en que $g(s,\hat{\lambda})$ es la única función de probabilidad que resuelve la ecuación diferencial $g'(s) = \lambda g(s)$, proponen rechazar la hipótesis nula para valores grandes del siguiente estadístico:

$$T_n = \int_0^1 [\bar{X}_n g_n(s) - g_n'(s)]^2 ds.$$

- Modificando T_n, Treutler (1995) [21] propone rechazar la hipótesis nula para valores grandes del siguiente estadístico:

$$T_{n,a} = \int_0^1 [\bar{X}_n g_n(s) - g_n'(s)]^2 s^a ds,$$

  donde $a \geq 0$.

- Nakamura y Pérez Abreu (1993) [13], basándose en que $\frac{\partial^2}{\partial s^2} \log g(s;\lambda) \equiv 0$, proponen rechazar la hipótesis nula para valores grandes de la suma de coeficientes cuadrados del polinomio $g_n^2(s)\frac{\partial^2}{\partial} u^2 \log g_n(s)$. El estadístico puede también ser representado del siguiente modo:

$$V_n = \frac{1}{n^3} \sum_{i,j,k,l=1}^{n} X_i(X_i - X_j - 1)X_k(X_k - X_l - 1)1\{X_i + X_j = X_k + X_l\}$$

  Estos autores proponen, además, utilizar la distribución asintótica del siguiente estadístico:

$$V_n^\star = V_n/(\bar{X}_n)^{1.45} \tag{3.11}$$

  y utilizar los puntos críticos de dicha distribución asintótica que, computacionalmente comprueban, que no dependen de λ.

Al no ser ni la distribución nula ni la distribución asintótica de los test de esta sección tratables, Henze (1996) [8] propone el cálculo de los puntos críticos de estos test mediante un bootstrap paramétrico que veremos en la Sección 10.3.

Esto no así para el estadístico $V_n^\star$ por lo que se ha explicado anteriormente.

Relación de los estadísticos del test ponderados con el índice de dispersión de Fischer.

Baringhaus et al.(2000) [1] demuestran unas relaciones particulares entre D_n y los estadísticos $R_{n,a}$ y $T_{n,a}$, siendo $D_n = \sum_{j=1}^{n} \frac{(X_j - \bar{X})^2}{\bar{X}}$. Éstas son:

$$\lim_{a\to\infty} a^5 R_{n,a} = 6\bar{X}_n^2(D_n - n^2)/n$$

$$\lim_{a\to\infty} a^3 T_{n,a} = 2\bar{X}_n^2(D_n - n)^2/n$$

3.4.2. Test propuestos por Puig y Weiß

- Basándose en caracterizaciones de la distribución Poisson basadas, a su vez, en el operador α - thinning de la distribución binomial, Puig y Weiß (2020) [16] proponen rechazar la hipótesis nula para valores grandes de los siguientes estadísticos, siendo la hipótesis alternativa distribuciones pertenecientes a la clase LC (variables aleatorias de conteo tales que el logaritmo de su función generatriz de probabilidad en $[0,1]$ es convexo):

$$\hat{\Delta}_1 = \int_0^1 g_n(t) - [g_n(t)(\frac{t+1}{2})]^2)dt$$

$$\hat{\Delta}_\infty = \max_{t\in[0,1]}(g_n(t) - [g_n(t)(\frac{t+1}{2})]^2)$$

- Para alternativas más generales, proponen rechazar la hipótesis nula para valores grandes de los siguientes estadísticos:

$$\hat{\Delta}_1^\star = \int_0^1 |g_n(t) - [g_n(t)(\frac{t+1}{2})]^2|dt$$

$$\hat{\Delta}_2 = \int_0^1 (g_n(t) - [g_n(t)(\frac{t+1}{2})]^2)^2 dt$$

$$\hat{\Delta}_\infty^\star = \max_{t\in[0,1]}(|g_n(t) - [g_n(t)(\frac{t+1}{2})]^2|)$$

- Se propone también rechazar la hipótesis nula para las siguientes ponderaciones de algunos de los estadísticos anteriores:

$$\hat{\Delta}_{1,a} = \int_0^1 g_n(s) - [g_n(s)(\frac{t+1}{2})]^2)s^a ds,$$

$$\hat{\Delta}_{1,a}^\star = \int_0^1 |g_n(s) - [g_n(s)(\frac{t+1}{2})]^2|s^a ds,$$

$$\hat{\Delta}_{2,a} = \int_0^1 (g_n(s) - [g_n(s)(\frac{t+1}{2})]^2)^2 dss^a$$

Al no ser ni la distribución nula ni la distribución asintótica nula de los test de esta sección tratables, Henze (1996) [8] propone el cálculo de los puntos críticos de estos test mediante un bootstrap paramétrico que veremos en en la Sección 10.3.

Capítulo 4

El test de Baringhaus y Henze

4.1. Introducción

Kocherlakota y Kocherlakota (1986) [10] indican que la inferencia estadística para las distribuciones de conteo podría estar basada en la función generatriz de probabilidad empírica. Como ejemplo propone un nuevo test de bondad de ajuste para el contraste (1.1). Sin embargo, debido a que el test está basado en comparar la función generatriz de probabilidad empírica asociada a muestra con la función generatriz de probabilidad de la ley Poisson en un número finito de puntos, el test resultante no es consistente frente a toda alternativa.

Baringhaus y Henze (1992) [2] proponen un test basado en la función generatriz de probabilidad empírica que es consistente, al menos frente las distribuciones alternativas con momento de primer orden finito.

4.2. Una caracterización de la ley Poisson y estadístico propuesto

Sea X una variable aleatoria que toma valores enteros no negativos. Ya visto que la distribución de X está determinada por su función generatriz (o generadora) de probabilidad en el intervalo [0,1], definida como sigue

$$g(t) = E(t^X), \quad t \in [0,1].$$

Si $X \sim Pois(\lambda)$, entonces

$$g(t; \lambda) = \exp\{\lambda(t-1)\}.$$

Proposición 1. *$g(t; \lambda)$ es la única función generatriz de probabilidad de una variable de conteo*

X con $E(X) < \infty$ que satisface la siguiente ecuación diferencial,

$$\lambda g(t) = g'(t), \quad t \in [0,1], \tag{4.1}$$

para algún $\lambda > 0$.

Demostración Si $g(t) = \lambda g'(t)$, $t \in [0,1]$, entonces

$$\frac{g'(t)}{g(t)} = \lambda, \quad t \in [0,1],$$

de donde,

$$\ln(g(t)) = \lambda t + C, \quad t \in [0,1],$$

siendo C una constante, y por tanto,

$$g(t) = e^{\lambda t + C}, \quad t \in [0,1].$$

Para determinar C imponemos que toda función generatriz de probabilidad cumple que $g(1) = 1$

$$g(1) = e^{\lambda + C} = 1 \Longleftrightarrow C = -\lambda,$$

en consecuencia $g(t) = \exp\{\lambda(t-1)\}$. □

Nótese que si $g(t)$ satisface la ecuación (4.1), como $g(1) = 1$ y $g'(1) = E(X)$, se sigue que $\lambda = E(X)$.

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable de conteo X. Con el objetivo de utilizar la caracterización en la Proposición 1 para construir un test para la ley Poisson, reemplazamos $g(t)$, $g'(t)$ y λ por estimadores. Específicamente, reemplazamos $g(t)$, $g'(t)$ y λ por $g_n(t)$, $g'_n(t)$ y $\bar{X}_n$, donde

$$\begin{aligned}
\bar{X}_n &= \frac{1}{n}\sum_{j=1}^{n} X_j, \\
g_n(t) &= \int t^x dF_n(x) = \frac{1}{n}\sum_{j=1}^{n} t^{X_j} = \sum_{j\geq 0} \hat{p}_j t^j, \\
g'_n(t) &= \frac{d}{dt} g_n(t) = \frac{1}{n}\sum_{j=1}^{n} X_j I(X_j \geq 1) t^{X_j - 1} = \sum_{j\geq 1} \hat{p}_j t^{j-1},
\end{aligned}$$

donde F_n es la función de distribución empírica asociada a la muestra,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), \quad x \in \mathbb{R},$$

y
$$\hat{p}_j = \frac{1}{n}\sum_{k=1}^{n} I(X_k = j) = F_n(j) - F_n(j-1)$$
es la frecuencia relativa de j, $j = 0, 1, 2, \ldots$.

Como $\bar{X}_n$ es un estimador consistente de $E(X)$, $g_n(t)$ es un estimador consistente de $g(t)$, y $g_n'(t)$ es un estimador consistente de $g'(t)$, si $X \sim Pois(\lambda)$, para algún $\lambda > 0$, entonces $\bar{X}_n g_n(t) - g_n'(t)$ debe estar cerca de 0, $\forall t \in [0,1]$. Esta cercanía puede ser interpretada de varias maneras. Baringhaus y Henze (1992) [2] proponen considerar el estadístico
$$T_n = n\int_0^1 \left\{\bar{X}_n g_n(t) - g_n'(t)\right\}^2 dt, \tag{4.2}$$
y rechazar H_0 para valores *grandes* de T_n.

El siguiente resultado proporciona una expresión del estadístico T_n que resulta útil para su cálculo práctico.

Proposición 2. *Usando el convenio $0/0 = 0$, T_n puede ser escrito de la siguiente manera,*
$$T_n = \frac{1}{n}\sum_{i,j=1}^{n}\left(\frac{\bar{X}_n^2}{X_i + X_j + 1} + \frac{X_i + X_j}{X_i + X_j - 1}\right) - n\bar{X}_n(1 - \bar{Z}_n^2), \tag{4.3}$$
donde $\bar{Z}_n = \frac{1}{n}\sum_{j=1}^{n} Z_j$, $Z_j = I(X_j = 0)$, $1 \le j \le n$.

Demostración Se tiene que
$$\begin{aligned}\left\{\bar{X}_n g_n(t) - g_n'(t)\right\}^2 &= \bar{X}_n^2 \frac{1}{n^2}\sum_{i,j=1}^{n} t^{X_i+X_j} - 2\bar{X}_n \sum_{i,j=1}^{n} X_i I(X_i \ge 1) t^{X_i+X_j-1} \\ &\quad + \frac{1}{n^2}\sum_{i,j=1}^{n} X_i X_j I(X_i \ge 1) I(X_j \ge 1) t^{X_i+X_j-2}.\end{aligned}$$
Teniendo en cuenta que para todo $x \ge 0$
$$\int_0^1 t^x dt = \frac{1}{x+1},$$
y usando el convenio $0/0 = 0$, se sigue que
$$\begin{aligned}\int_0^1 \frac{1}{n^2}\sum_{i,j=1}^{n} t^{X_i+X_j} dt &= \frac{1}{n^2}\sum_{i,j=1}^{n}\frac{1}{X_i + X_j + 1}, \\ \int_0^1 \frac{1}{n^2}\sum_{i,j=1}^{n} X_i x_j I(X_i \ge 1) I(X_j \ge 1) t^{X_i+X_j-2} &= \frac{1}{n^2}\sum_{i,j=1}^{n}\frac{X_i X_j}{X_i + X_j - 1} I(X_i \ge 1) I(X_j \ge 1) \\ &= \frac{1}{n^2}\sum_{i,j=1}^{n}\frac{X_i X_j}{X_i + X_j - 1}, \\ \int_0^1 \frac{1}{n^2}\sum_{i,j=1}^{n} X_i I(X_i \ge 1) t^{X_i+X_j-1} &= \frac{1}{n^2}\sum_{i,j=1}^{n}\frac{X_i}{X_i + X_j} I(X_i \ge 1).\end{aligned}$$

Nótese que

$$
\begin{aligned}
2\frac{1}{n^2}\sum_{i,j=1}^{n}\frac{X_i}{X_i+X_j}I(X_i\geq 1) &= \frac{1}{n^2}\sum_{i,j=1}^{n}\frac{X_i}{X_i+X_j}I(X_i\geq 1)+\frac{1}{n^2}\sum_{i,j=1}^{n}\frac{X_j}{X_i+X_j}I(X_j\geq 1)\\
&= \frac{1}{n^2}\sum_{i,j=1}^{n}\frac{X_iI(X_i\geq 1)+X_jI(X_j\geq 1)}{X_i+X_j}.
\end{aligned}
$$

Ahora bien

- si $X_i \geq 1$ y $X_j \geq 1$, entonces $\frac{X_iI(X_i\geq 1)+X_jI(X_j\geq 1)}{X_i+X_j}=1$,
- si $X_i \geq 1$ y $X_j = 0$ (ó $X_i = 0$ y $X_j \geq 0$) entonces $\frac{X_iI(X_i\geq 1)+X_jI(X_j\geq 1)}{X_i+X_j}=1$,

por tanto,

$$
\begin{aligned}
2\frac{1}{n^2}\sum_{i,j=1}^{n}\frac{X_i}{X_i+X_j}I(X_i\geq 1) &= \frac{1}{n^2}\sum_{i,j=1}^{n} I(X_i\geq 1 \text{ ó } X_j\geq 1)\\
&= \frac{1}{n^2}\sum_{i,j=1}^{n}\{1-I(X_i=0,X_j=0)\}\\
&= \frac{1}{n^2}\sum_{i,j=1}^{n}\{1-I(X_i=0)I(X_j=0)\}\\
&= (1-\bar{Z}_n),
\end{aligned}
$$

lo que demuestra el resultado. □

4.3. Límite del estadístico

Teorema 10. *Sea X una variable de conteo con $\mu = E(X) < \infty$ y función generatriz de probabilidad $g(t)$, entonces*

$$\frac{1}{n}T_n \xrightarrow{c.s} \tau = \int_0^1 \{\mu g(t) - g'(t)\}^2 dt.$$

Demostración Aplicando la Proposición 1 en Novoa Muñoz y Jiménez-Gamero [14], se sigue que

$$\sup_{u\in[0,1]} |g_n(t) - g(t)| \xrightarrow{c.s} 0,$$

y

$$\sup_{u\in[0,1]} |g_n'(t) - g'(t)| \xrightarrow{c.s} 0.$$

Además, sabemos por la ley fuerte de los grandes números que

$$\bar{X} \xrightarrow{c.s} \mu < \infty.$$

Teniendo en cuenta que $0 \leq g(t) \leq 1$, $\forall t \in [0,1]$, se tiene que

$$\begin{aligned} \left|\bar{X}_n g_n(t) - g_n'(t) - \mu g(t) + g'(t)\right| &= \left|\bar{X}_n \left\{g_n(t) - g(t)\right\} + (\bar{X}_n - \mu)g(t) + \left\{g'(t) - g_n'(t)\right\}\right| \\ &\leq |\bar{X}_n| \left|g_n(t) - g(t)\right| + |\bar{X}_n - \mu| + \left|g'(t) - g_n'(t)\right|, \end{aligned}$$

y por tanto,

$$\sup_{u\in[0,1]} \left|\bar{X}_n g_n(t) - g_n'(t) - \mu g(t) + g'(t)\right| \xrightarrow{c.s} 0,$$

lo que implica que

$$\int_0^1 \left\{\bar{X}_n g_n(t) - g_n'(t) - \mu g(t) + g'(t)\right\}^2 dt \leq \left\{\sup_{u\in[0,1]} \left|\bar{X}_n g_n(t) - g_n'(t) - \mu g(t) + g'(t)\right|\right\}^2 \xrightarrow{c.s} 0. \tag{4.4}$$

Se tiene que

$$\begin{aligned} \frac{1}{n}T_n &= \int_0^1 \left\{\bar{X}_n g_n(t) - g_n'(t)\right\}^2 dt \\ &= \int_0^1 \left\{\bar{X}_n g_n(t) - g_n'(t) - \mu g(t) + g'(t) + \mu g(t) - g'(t)\right\}^2 dt \\ &= \int_0^1 \{\mu g(t) - g'(t)\}^2 dt + \int_0^1 \left\{\bar{X}_n g_n(t) - g_n'(t) - \mu g(t) + g'(t)\right\}^2 dt \\ &\quad +2\int_0^1 \left\{\mu g(t) - g'(t)\right\}\left\{\bar{X}_n g_n(t) - g_n'(t) - \mu g(t) + g'(t)\right\} dt. \end{aligned}$$

Teniendo en cuenta que $0 \leq g(t) \leq 1$, $\forall t \in [0,1]$ y que $0 \leq g'(t) \leq E(X) < \infty$, $\forall t \in [0,1]$, se sigue que

$$\int_0^1 \{\mu g(t) - g'(t)\}^2 dt < \infty. \tag{4.5}$$

Por la desigualdad de Cauchy-Schwarz,

$$\begin{aligned} &\int_0^1 \left\{\mu g(t) - g'(t)\right\}\left\{\bar{X}_n g_n(t) - g_n'(t) - \mu g(t) + g'(t)\right\} dt \\ &\leq \left[\int_0^1 \{\mu g(t) - g'(t)\}^2 dt \int_0^1 \left\{\bar{X}_n g_n(t) - g_n'(t) - \mu g(t) + g'(t)\right\}^2 dt\right]^{1/2}. \end{aligned} \tag{4.6}$$

De (4.4), (4.5) y (4.6), se sigue que

$$\int_0^1 \left\{\mu g(t) - g'(t)\right\}\left\{\bar{X}_n g_n(t) - g_n'(t) - \mu g(t) + g'(t)\right\} dt \xrightarrow{c.s} 0,$$

y por tanto $\frac{1}{n}T_n \xrightarrow{c.s} \tau$. □

Nótese que $\tau \geq 0$ con $\tau = 0$ si y sólo si H_0 es cierta. Por tanto, es razonable considerar un test que rechace para valores grandes de T_n.

4.4. Distribución nula asintótica del estadístico

Como vimos antes, el test propuesto rechaza H_0 para valores *grandes* de T_n. Para determinar los puntos críticos del test, o equivalentemente, los p-valores, necesitamos averiguar cuál la distribución nula del estadístico, que es claramente desconocida. Por lo que tendremos aproximarla de algún modo. La forma clásica de aproximarla es mediante la su distribución nula asintótica. Por este motivo, en este apartado determinaremos distribución nula asintótica de T_n.

Con este objetivo observamos que

$$T_n = \|Z_n\|_{L^2}^2$$

donde

$$Z_n = \sqrt{n}\bar{X}_n \left\{g_n(t) - g_n'(t)\right\}.$$

El siguiente resultado da la distribución asintótica nula de Z_n.

Teorema 11. *Sean $X_1, \ldots, X_n$ i.i.d. de $X \sim Pois(\lambda)$, para algún $\lambda > 0$, entonces*

$$Z_n \xrightarrow{\mathcal{L}} Z,$$

en L^2, donde Z es un elemento Gaussiano de L^2 con media cero y función de covarianza

$$cov(Z(t), Z(s)) = \{\lambda^2(1-t)(1-s) + \lambda\}g(ts) - \lambda g(t)g(s).$$

Demostración Recordemos que $Z_n(t) = \bar{X}_n g_n(t) - g_n'(t)$. Se tiene que

$$\bar{X}_n g_n(t) = (\bar{X}_n \pm \lambda)\left\{g_t \pm g(t)\right\} =$$

$$(\bar{X} - \lambda)\left\{g_n(t) - g(t)\right\} + \lambda\left\{g_n(t) - g(t))\right\} + (\bar{X} - \lambda)g(t) + \lambda g(t).$$

Sustituyendo en Z_n,

$$Z_n(t) = L_n(t) + R_n(t),$$

donde

$$\begin{aligned}
L_n(t) &= \lambda\left\{g_n(t) - g(t)\right\} + (\bar{X} - \lambda)g(t) - \left\{g_n'(t) - \lambda g(t)\right\} \\
&= \lambda\left\{g_n(t) - g(t)\right\} + (\bar{X} - \lambda)g(t) - \left\{g_n'(t) - g'(t)\right\}, \\
R_n(t) &= (\bar{X} - \lambda)\left\{g_n(t) - g(t)\right\}.
\end{aligned}$$

Nótese que

$$L_n = \frac{1}{n}Y_i(t),$$

con $Y_1, \ldots, Y_n$ i.i.d. definidas como

$$Y_i(t) = \lambda\{t^{X_i} - g(t)\} + g(t)(X_i - \lambda) - I(X_i \geq 1)X_i t^{X_i - 1} + g'(t),$$

$1 \leq i \leq n$. Se tiene que

$$E\{Y_1(t)\} = 0, \quad \forall t \in [0,1],$$

Vamos ahora a calcular la función de covarianza de $Y = Y_1$. Se tiene que

$$cov\{Y(t), Y(s)\} = E\{Y(t)Y(s)\} - E\{Y(t)\}E(Y(s)\} = E\{Y(t)Y(s)\}.$$

Ahora,

$$\begin{aligned} E\{Y(t)Y(s)\} &= \lambda^2 E[\{t^X - g(t)\}\{s^X - g(s)\}] \\ &\quad +\lambda g(s)E[\{t^X - g(t)\}(X-\lambda)] \\ &\quad -\lambda E[\{t^X - g(t)\}\{I(X \geq 1)Xs^{X-1} - g'(s)\}] \\ &\quad +\lambda g(t)E[(X-\lambda)\{s^X - g(s)\}] \\ &\quad +g(t)g(s)E\{(X_1-\lambda)^2\} \\ &\quad -g(t)E[(X-\lambda)\{I(X \geq 1)Xs^{X-1} - g'(s)\}] \\ &\quad -\lambda E[\{I(X \geq 1)Xt^{X-1} - g'(t)\}\{s^X - g(s)\}] \\ &\quad -\lambda g(s)E[\{I(X \geq 1)Xt^{X-1} - g'(t)\}(X-\lambda)] \\ &\quad +E[\{I(X \geq 1)Xt^{X-1} - g'(t)\}\{I(X \geq 1)Xs^{X-1} - g'(s)\}]. \end{aligned} \tag{4.7}$$

A continuación calculamos por separado cada término en la expresión de $E\{Y(t)Y(s)\}$, utilizando que $g'(t) = \lambda g(t)$ y $g''(t) = \lambda^2 g(t)$.

$$\begin{aligned} E[\{t^X - g(t)\}\{s^X - g(s)\}] &= E(st)^X - g(t)E(s^X) - g(s)E(t^X) + g(t)g(s) \\ &= g(ts) - g(t)g(s) - g(s)g(t) + g(t)g(s) \\ &= g(ts) - g(t)g(s). \end{aligned} \tag{4.8}$$

$$E[\{t^X - g(t)\}(X-\lambda)] = E[Xt^X] - \lambda E[t^X] - g(t)E(X) + \lambda g(t).$$

Como

$$E[Xt^X] = e^{-\lambda} \sum_{k\geq 1} kt^k \frac{\lambda^k}{k!} = te^{-\lambda} \sum_{k \geq 0} \frac{(t\lambda)^k}{k!} = \lambda t g(t),$$

se sigue que

$$\begin{aligned} E[\{t^X - g(t)\}(X-\lambda)] &= \lambda t g(t) - \lambda g(t) - \lambda g(t) + \lambda g(t) \\ &= \lambda t g(t) - \lambda g(t) \\ &= \lambda(t-1)g(t). \end{aligned} \tag{4.9}$$

$$\begin{aligned} &E[\{t^X - g(t)\}\{I(X \geq 1)Xs^{X-1} - g'(s)\}] \\ &\quad = E[I\{X \geq 1\}Xs^{X-1}t^X] - g'(s)E[t^X] - g(t)E[I\{X \geq 1\}Xs^{X-1}] + g'(s)g(t) \\ &\quad = E[I\{X \geq 1\}Xs^{X-1}t^X] - g'(s)g(t) - g(t)g'(s) + g'(s)g(t) \\ &\quad = tE[I\{X \geq 1\}X(st)^{X-1}] - g'(s)g(t) \\ &\quad = tg'(ts) - g(t)g'(s) \\ &\quad = \lambda t g(ts) - \lambda g(t)g(s). \end{aligned} \tag{4.10}$$

$$E\{(X_1-\lambda)^2\}=\lambda. \tag{4.11}$$

$$\begin{aligned}
&E[(X-\lambda)\{I(X\geq 1)Xs^{X-1}-g'(s)\}]\\
&\quad=E[X^2s^{X-1}I\{X\geq 1\}]-g'(s)E[X]-\lambda E[I\{X\geq 1\}Xs^{X-1}]-\lambda g'(s)\\
&\quad=E[X^2s^{X-1}I\{X\geq 1\}]-g'(s)\lambda-\lambda g'(s)+\lambda g'(s).
\end{aligned}$$

Calculamos ahora el término $E[X^2s^{X-1}I\{X\geq 1\}]$. Sabemos que

$$g''(t)=\sum_{k=2}^{\infty}I\{X\geq 2\}k(k-1)s^{k-2}e^{-\lambda}\frac{\lambda^k}{k!}.$$

Por tanto,

$$\begin{aligned}
E[X^2s^{X-1}I\{X\geq 1\}] &= \sum_{k=1}^{\infty}k^2s^{k-1}e^{-\lambda}\frac{\lambda^k}{k!}\\
&= \sum_{k=1}^{\infty}k(k-1)s^{k-1}e^{-\lambda}\frac{\lambda^k}{k!}+\sum_{k=1}^{\infty}ks^{k-1}e^{-\lambda}\frac{\lambda^k}{k!}\\
&= \sum_{k=2}^{\infty}k(k-1)s^{k-1}e^{-\lambda}\frac{\lambda^k}{k!}+g'(s)\\
&= sg''(s)+g'(s),
\end{aligned}$$

de donde,

$$\begin{aligned}
E[(X-\lambda)\{I(X\geq 1)Xs^{X-1}-g'(s)\}] &= sg''(s)+g'(s)-\lambda g'(s)\\
&= \lambda^2sg(s)+\lambda g(s)-\lambda^2g(s)\\
&= (\lambda^2s-\lambda^2+\lambda)g(s).
\end{aligned} \tag{4.12}$$

$$\begin{aligned}
&E[\{I(X\geq 1)Xt^{X-1}-g'(t)\}\{I(X\geq 1)Xs^{X-1}-g'(s)\}]\\
&\quad=E[I\{X\geq 1\}X^2(ts)^{X-1}]-E[I\{X\geq 1\}Xt^{X-1}]g'(s)\\
&\qquad-E[I\{X\geq 1\}Xs^{X-1}]g'(t)+g'(t)g'(s)\\
&\quad=tsg''(ts)+g'(ts)-g'(t)g'(s)-g'(s)g'(t)+g'(t)g'(s)\\
&\quad=tsg''(ts)+g'(ts)-g'(t)g'(s)\\
&\quad=\lambda^2tsg(ts)+\lambda g(ts)-\lambda^2g(t)g(s).
\end{aligned} \tag{4.13}$$

Finalmente, sustituyendo (4.8)–(4.13) en (4.7) se obtiene

$$
\begin{aligned}
E\{Y(t)Y(s)\} &= \lambda^2\{g(ts) - g(t)g(s)\} \\
&\quad +\lambda^2(t-1)g(t)g(s) \\
&\quad -\lambda^2 tg(ts) + \lambda^2 g(t)g(s) \\
&\quad +\lambda^2(s-1)g(t)g(s) \\
&\quad +\lambda g(t)g(s) \\
&\quad -(\lambda^2 s - \lambda^2 + \lambda)g(t)g(s) \\
&\quad -\lambda^2 sg(ts) + \lambda^2 g(t)g(s) \\
&\quad -(\lambda^2 t - \lambda^2 + \lambda)g(t)g(s) \\
&\quad +\lambda^2 tsg(ts) + \lambda g(ts) - \lambda^2 g(t)g(s) \\
&= \{\lambda^2(1-t)(1-s) + \lambda\}g(ts) - \lambda g(t)g(s).
\end{aligned}
$$

Como

$$E\|Y_1\|_{L^2}^2 = \int_0^1 cov\{Y_1(t), Y_1(t)\}dt \;=\; \int_0^1 \{\lambda^2(1-t)^2 + \lambda\}g(t^2)dt - \int_0^1 \lambda g(t)^2 dt < \infty$$

Al ser las integrales en un compacto de funciones continuas, su resultado es finito. Se sigue que $Y_1, \ldots, Y_n$ son elementos aleatorios i.i.d. que toman valores en L^2. Por el teorema central del límite en espacios de Hilbert se sigue que

$$\sqrt{n}L_n \xrightarrow{\mathcal{L}} Z,$$

en L^2, donde Z es el proceso Gaussiano de L^2 definido en el enunciado.

Por el teorema de Slutsky, para probar el resultado es suficiente ver que

$$\|\sqrt{n}R_n\|_{L^2}^2 = \int_0^1 nR_n^2(t)dt = o_P(1). \tag{4.14}$$

Se tiene que

$$\sqrt{n}R_n(t) = (\bar{X} - \lambda)\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{t^{X_j} - g(t)\right\} := (\bar{X} - \lambda)W_n(t),$$

y por tanto,

$$\|\sqrt{n}R_n\|_{L^2}^2 = (\bar{X} - \lambda)^2\|W_n\|_{L^2}^2. \tag{4.15}$$

Por la ley fuerte de los grandes números se tiene que

$$\bar{X} - \lambda \xrightarrow{c.s.} 0,$$

y por el teorema de la aplicación continua se sigue que

$$(\bar{X} - \lambda)^2 \xrightarrow{c.s.} 0. \tag{4.16}$$

Nótese que

$$W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i'(t),$$

con $Y_1', \ldots, Y_n'$ iid de Y' definida como

$$Y'(t) = t^X - g(t).$$

Se tiene que

$$E\{Y'(t)\} = 0, \quad \forall t \in [0,1],$$

$$cov\{Y'(t), Y'(s)\} = E\{Y'(t)Y'(s)\} - E\{Y'(t)\}E\{Y'(s)\} = E\{Y'(t)Y'(s)\}.$$

Ahora bien,

$$E\{Y'(t)Y'(s)\} = E\{(ts)^X\} - g(t)g(s) = g(ts) - g(t)g(s).$$

Como

$$E\|Y'\|_{L^2}^2 = \int_0^1 cov\{Y'(t), Y'(t)\}dt = \int_0^1 e^{-\lambda[t^2-1]} - \int_0^1 e^{\lambda[2t-2]} < \infty.$$

Al ser las integrales en un compacto de funciones continuas, su resultado es finito. Se sigue que $Y_1', \ldots, Y_n'$ son elementos aleatorios i.i.d. que toman valores en L^2. Por el teorema central del límite en espacios de Hilbert se tiene que:

$$\sqrt{n}W_n \xrightarrow{\mathcal{L}} W,$$

en L^2, donde W es un proceso gaussiano de L^2 con media 0 y función de covarianzas

$$cov(W(t), W(s)) = g(ts) - g(t)g(s).$$

Aplicando el teorema de la aplicación continua, se sigue que

$$\|W_n\|_{L^2}^2 \xrightarrow{\mathcal{L}} \|W\|_{L^2}^2 \tag{4.17}$$

Aplicado el teorema de Slutsky se sigue que (4.14) se cumple. □

Corolario 1. *En las condiciones del Teorema 11 se tiene que*

$$\|Z_n\|_{L^2}^2 \xrightarrow{\mathcal{L}} \|Z\|_{L^2}^2.$$

Demostración El resultado es una consecuencia inmediata del Teorema 11 y del teorema de la aplicación continua. □

Como indican Baringhaus y Henze [2], la distribución de $\|Z_n\|_{L^2}^2$ coincide con la de

$$\sum_{i \geq 1} \theta_i \chi_{1i}^2,$$

donde $\chi^2_{11}, \chi^2_{12}, \ldots$ con variables independientes distribuidas según una ley χ^2_1, y $\{\theta_i\}_{i\geq 1}$ son los autovalores asociados del operador integral asociado a la función de covarianza de Z. Los $\{\theta_i\}_{i\geq 1}$ cumplen

$$\theta_i \geq 0, \qquad \forall i \geq 1,$$
$$\sum_{i\geq 1} \theta_i = \int_0^1 cov\{Z(t), Z(t)\}dt < \infty.$$

4.5. Consistencia

Sean $\alpha \in (0,1)$, $T_{n,obs}$ el valor observado del estadístico T_n en la muestra, y consideremos el test

$$\Psi = \begin{cases} 1, & \text{if} \quad T_{n,obs} \geq t_{n,1-\alpha}, \\ 0, & \text{otherwise}, \end{cases}$$

donde $t_{n,1-\alpha}$ es el percentil de orden $1-\alpha$ de la distribución nula de T_n, es decir,

$$t_{n,1-\alpha} = \text{ínf}\{x \,:\, P_0(T_n \leq x) \geq 1-\alpha),$$

o equivalentemente, el test que reachaza H_0 si

$$p = P_0(T_n \geq T_{n,obs}) \leq \alpha,$$

donde P_0 denota la distribución bajo H_0.

Veamos que Ψ es consistente frente a alternativas con momento de orden 1 finito, es decir, el test asintóticamente detecta este tipo de alternativas, en otras palabras, si $X \nsim Pois(\lambda), \forall \lambda > 0$, entonces el test rechaza H_0 para n "grande".

Teorema 12. *Sea X una variable de conteo con $\mu = E(X) < \infty$ tal que $X \nsim Pois(\lambda), \forall \lambda > 0$, entonces $P(\Psi = 1) \to 1$.*

Demostración El el Teorema 10 vimos que $\frac{1}{n}T_n$ tiene un límite finito c.s.. Por tanto para demostrar el resultado es sufiente ver que la distribución nula de $\frac{1}{n}T_n$ converge a 0, lo cual se sigue del Corolario 1. □

4.6. Extensiones del test

Treutler (1995) [21] propone la siguiente extensión del test T_n:

$$T_{n,a} = n \int_0^1 \{\bar{X}_n g_n(u) - g_n'(u)\}^2 u^a du \tag{4.18}$$

para valores de $a \geq 0$.

Capítulo 5

Los test de Puig y Weiß

5.1. Introducción

Existe una clase de funciones llamada la clase LC, la cual tiene grandes aplicaciones reales, particularmente en la biodosimetría (técnica que consiste en detectar dosis de radiación a las que alguien ha podido estar expuesto a partir del análisis de muestras biológicas). Además de la distribución Poisson, esta clase contiene a muchas otras distribuciones que son desviaciones con respecto a la Poisson, por ejemplo, distribuciones sobredispersas, esto es, aquellas que tienen una varianza mayor que la media ($\sigma^2 > \mu$) y cero-infladas, esto es, la probabilidad de 0 $p_0 = g(0) > \exp(-\mu)$.

Puig y Weiß (2020) [16] proponen, basándose en algunas caracterizaciones de la Poisson y en propiedades de las funciones de la clase LC que veremos a continuación, una serie de test de Poissonidad contra alternativas pertenecientes a la clase LC y, generalizando, test de Poissonidad contra alternativas generales.

5.2. Una caracterización de la ley Poisson

Las caracterizaciones que veremos están basadas en el operador thinning de la Binomial.

Definición 12. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable de conteo X. Sean $\xi_1, \xi_2, \ldots$ variables aleatorias Bernoulli independientes idénticamente distribuidas con probabilidad de éxito $p = \alpha \in (0,1)$, todas independientes de X. La variable de conteo

$$\alpha * X = \sum_{i=1}^{X} \xi_i$$

donde si $X = 0$, entonces tomamos $\alpha * X = 0$, es una Binomial α-thinning de X.

Teorema 13. *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable de conteo X, la función generatriz de probabilidad de la Binomial α-thinning de X se define como*

$$g_{\alpha * X}(s) = g_X(1-\alpha+\alpha s).$$

Demostración La función generatriz de probabilidad de una distribución es:

$$g_x(s) = \sum_{x>0} p_x(x)s^x$$

Aplicando esta definición para la distribución Bernoulli tenemos:

$$p_x(x) = \begin{cases} p & si \quad x = 1 \\ 1-p & si \quad x = 0 \\ 0 & si \quad x \notin \{0,1\} \end{cases}$$

Entonces:

$$g_{Ber}(s) = p_x(0)s^0 + p_x(1) + s^1 = (1-p) + ps$$

Por tanto, la función generatriz de probabilidad de ξ_i es

$$1 - \alpha + \alpha s$$

.

Luego, haciendo el cambio de variable $t = 1 - \alpha + \alpha s$, se tiene:

$$E[s^{\alpha * X}] = E[(1-\alpha+\alpha s)^X]$$

siendo $g_X(s) = E[s^X]$ la función generatriz de probabilidad de X.

Teorema 14. *Sean X_1, X_2 variables aleatorias de conteo independientes idénticamente distribuidas y sea $Y_\alpha = \alpha * X_1 + (1-\alpha) * X_2$, entonces X_i sigue una distribución Poisson de parámetro λ si y solo si alguna de las siguientes condiciones se cumple:*

1. *Y_α tiene la misma distribución que X_i $\forall \alpha \in (0,1)$.*

2. *X_i tiene momento de primer orden e Y_α tiene la misma distribución que X_i para algún $\alpha \in (0,1)$.*

Demostración Sea $g(s)$ la función generatriz de probabilidad de X_i, entonces tenemos que.

$$g_{\alpha * X_1}(s) = g(1-\alpha+\alpha s),$$
$$g_{(1-\alpha)*X_2} = g(1-(1-\alpha)+(1-\alpha+\alpha)s) = g(\alpha+(1-\alpha)s).$$

Por tanto, la función generatriz de probabilidad de Y_α es:

$$g_Y(s) = g(1-\alpha+\alpha s)g(\alpha+(1-\alpha)s). \tag{5.1}$$

- A continuación vamos a probar la suficiencia de la condición ***1***.

  Suponemos que las distribuciones de Y_α y de X_i son la misma $\forall \alpha \in (0,1)$. Esto es, $g_Y(s) = g(s)$ lo cual implica la siguiente identidad:

  $$g(s) = g(1-\alpha+\alpha s)g(\alpha+(1-\alpha)s). \tag{5.2}$$

  Haciendo el cambio de variables $s = t+1$, se tiene:

  $$g(t+1) = g(\alpha t+1)g((1-\alpha)t+1). \tag{5.3}$$

  Sea $\psi(z) = \log(g(z+1))$. La identidad anterior se transforma en:

  $$\psi(t) = \psi(\alpha t) + \psi((1-\alpha)t). \tag{5.4}$$

  Haciendo el cambio de variables $\alpha t = x$ y $(1-\alpha)t = y$, la identidad (5.4) se transforma en:

  $$\psi(x+y) = \psi(x) + \psi(y). \tag{5.5}$$

  Cualquier función generatriz de probabilidad $g(s)$ es una función analítica en el disco abierto y unitario continua por la izquierda en $s = 1$ con $g(1) = 1$. Esto implica que $\psi(z)$ es una función analítica en el disco abierto $|z+1| < 1$ y es continua por la izquierda en $z = 0$ con $\psi(0) = 0$.

  Sabemos que $\psi(t)$ es diferenciable en $-2 < t < 0$ y tomando derivadas respecto a x e y se tiene que $\psi'(x) = \psi'(y) \quad \forall x, y \in (-2, 0)$. Por tanto, la única solución en el dominio es la función lineal $\psi(t) = ct + b$, siendo c y b constantes. Generalizando para $|z+1| < 1$, tenemos que $\psi(0) = 0$, lo cual implica que $b = 0$. Luego $\psi(z) = cz$. Esto implica que $g(z+1) = \exp(cz)$ y $g(z) = \exp(c(z-1))$. Como $g(z)$ es una función generatriz de probabilidad, se sabe que es una función creciente que toma valores reales en $(0,1)$, por tanto $c = \lambda > 0$. Esto concluye la demostración de la suficiencia de la condición ***1***.

- A continuación vamos a demostrar la implicación contraria, es decir, que si $X_i \sim Pois(\lambda)$, entonces se cumple la condición ***1***.

  Si $X_i \sim Pois(\lambda)$, entonces, $g(s) = \exp(\lambda(s-1))$.

  Siendo $g_Y(s) = g(1-\alpha+\alpha s)g(\alpha+(1-\alpha)s)$, tenemos que

  $$g(1-\alpha+\alpha s) = g_\alpha(s)$$

  y

  $$g(\alpha+(1-\alpha)s) = g_{1-\alpha}(s).$$

Luego,

$$g_\alpha(s)g_{1-\alpha}(s) = g(s) = g_Y(s) = \exp(\lambda(s-1)),$$

$\forall \alpha \in (0,1)$.

Esto implica que se X_i tiene la misma distribución que X para todo $\alpha \in (0,1)$, es decir, que se cumple la condición ***1***.

- A continuación vamos a probar la suficiencia de la condición **2**.

  Sea un $\alpha \in (0,1)$ tal que las distribuciones de Y_α y de X_i sean la misma. Esto implica la existencia de la identidad (5.4) para un valor particular de α. Iterando la ecuación funcional (5.4) se tiene:

$$\psi(t) = \sum_{i=0}^{n} \binom{n}{i} \psi(\alpha^i (1-\alpha)^{n-i} t), \tag{5.6}$$

$\forall n \geq 1$.

Para $\alpha = 1/2$, tenemos:

$$\begin{aligned} \psi(t) = \sum_{i=0}^{n} \binom{n}{i} \psi(\frac{1}{2}^i \frac{1}{2}^{n-i} t) = \\ 2^n \psi(\frac{1}{2}^n t) = \\ 2^n \psi(t/2^n) \end{aligned} \tag{5.7}$$

$\psi(t)$ es diferenciable en $-2 < t < 0$ y su primera derivada es continua en 0. Tomando la primera derivada tenemos

$$\sum_{i=0}^{n} \binom{n}{i} \alpha^i (1-\alpha)^{n-i} \psi'(\alpha^i (1-\alpha)^{n-i} t). \tag{5.8}$$

Suponiendo $\alpha \leq (1-\alpha)$ (si $\alpha > (1-\alpha)$ se cambia α por $(1-\alpha)$) tenemos que $\alpha^i (1-\alpha)^{n-i} \geq \alpha^n, \quad \forall i$.

Esto es debido a que, si $\alpha \leq (1-\alpha)$, tenemos:

$$\alpha^n = \alpha^i \alpha^{n-i} \leq \alpha^i (1-\alpha)^{n-i} \ \forall i.$$

Esto implica la siguiente desigualdad:

$$\min_{s \in [-2\alpha^n, 0]} \psi'(s) \leq \psi'(\alpha^i (1-\alpha)^{n-i} t) \leq \max_{s \in [-2\alpha^n, 0]} \psi'(s).$$

Como

$$\sum_{i=0}^{n} \binom{n}{i} \alpha^i (1-\alpha)^{n-i} = 1,$$

(5.6) implica que

$$\min_{s\in[-2\alpha^n,0]} \psi'(s) \leq \psi'(t) \leq \max_{s\in[-2\alpha^n,0]} \psi'(s)$$

$\forall n \geq 1$.

Si n tiende a ∞, $\psi'(t) = \psi'(0) = \lambda \quad \forall t$. Vemos que $\lambda = \psi'(0) = g'(1)$ que es, precisamente, $E(X_i)$.

Por tanto $\psi(t) = \lambda t$ y, siguiendo el razonamiento usado para demostrar la condición ***1*** tenemos que $g(t+1) = exp(\lambda t)$ y $g(t) = exp(\lambda(t-1))$. Como $g(t)$ es una función generatriz de probabilidad, esto es, función creciente que toma valores reales en $(0,1)$, tenemos que $\lambda > 0$. Esto concluye la demostración de la suficiencia de **2**.

- A continuación vamos a demostrar la implicación contraria, es decir, que si $X_i \sim Pois(\lambda)$, entonces se cumple la condición **2**.

  Si $X_i \sim Pois(\lambda)$, entonces, $g(s) = \exp(\lambda(s-1))$.

  Siendo $g_Y(s) = g(1-\alpha+\alpha s)g(\alpha+(1-\alpha)s)$, tenemos que

  $$g(1-\alpha+\alpha s) = g_\alpha(s)$$

  y

  $$g(\alpha+(1-\alpha)s) = g_{1-\alpha}(s).$$

  Luego,

  $$g_\alpha(s)g_{1-\alpha}(s) = g(s) = g_Y(s) = \exp(\lambda(s-1)),$$

  $\forall \alpha \in (0,1)$.

  Esto implica que se X_i tiene la misma distribución que X para todo $\alpha \in (0,1)$, es decir, que se cumple la condición **2**.

□

La conclusión del Teorema 14 es que dada una distribución de conteo que satisfaga las condiciones ***1*** ó **2** con una función generatriz de probabilidad $g(s)$, la identidad

$$g(s) = g(1-\alpha+\alpha s)g(\alpha+(1-\alpha)s)$$

se satisface $\forall s$ si y solo si la distribución en cuestión sigue una ley Poisson.

Este teorema, además, nos conduce a algunas medidas de discrepancia respecto a la Poisson considerando un valor ajustado de α.

La primera medida basada la norma L^1:

$$\Delta_1(g,\alpha) = \int_0^1 |g(s) - g(1-\alpha+\alpha s)g(\alpha+(1-\alpha)s)|dt. \tag{5.9}$$

La segunda medida basada en la norma L^2:

$$\Delta_2(g,\alpha) = \int_0^1 (g(s) - g(1-\alpha+\alpha s)g(\alpha+(1-\alpha)s))^2 dt. \tag{5.10}$$

La tercera medida basada en la norma L^∞:

$$\Delta_\infty(g,\alpha) = \max_{t\in(0,1)} |g(s) - g(1-\alpha+\alpha s)g(\alpha+(1-\alpha)s)|. \tag{5.11}$$

5.3. Clase LC

La clase de funciones LC fue introducida por Puig y Kokonendji (2018) [15].

Definición 13. El conjunto de variables aleatorias de conteo tales que el logaritmo de su función generatriz de probabilidad en $[0,1]$ es convexo, se dice que forman parte de la clase LC.

El siguiente teorema demuestra que el valor absoluto puede ser eliminado de (5.9) y de (5.11) si la función para alternativas pertenecientes a la clase LC.

Teorema 15. *Sea $g(s)$ la función generatriz de probabilidad de una variable aleatoria de conteo perteneciente a la clase LC, entonces, $\forall s,\alpha \in [0,1]$*

$$g(s) \geq g(1-\alpha+\alpha s)g(\alpha+(1-\alpha)s).$$

Antes de demostrar este Teorema daremos una serie de definiciones y resultados previos útiles para su demostración.

Definición 14. Sea I un intervalo no vacío y no reducido a un punto y $f : I \to \mathbb{R}$ una función, se dice que f es convexa si verifica la siguiente condición:

Si $a,b \in I, a<b$, entonces, $f((1-t)a+tb) \leq (1-t)f(a)+tf(b)$ $,\forall t \in [0,1]$.

Definición 15. Se dice que una función $f(x)$ es superaditiva si

$$f(x+y) \geq f(x)+f(y),$$

para todo x e y en el dominio de f.

Lema 4. *Si f es convexa y creciente y $f(0)=0$, entonces f es superaditiva.*

Demostración Para demostrarlo, dibujamos la línea secante que une el punto $(0,0)$ con el punto $(x+y, f(x+y))$. La línea resultante es la representación gráfica de una función, $\mathbb{L}$ lineal y aditiva, esto es, $\mathbb{L}(x+y) = \mathbb{L}(x) + \mathbb{L}(y)$.

Recordar que una función es convexa si el segmento que une dos puntos cualesquiera de su gráfica queda por arriba de la curva de la función.

Por tanto, por convexidad tenemos que $f(x) \leq \mathbb{L}(x)$ y $f(y) \leq \mathbb{L}(y)$. Así, queda demostrada que la función f es superaditiva. □

Demostración del Teorema 12 Si $\log(g(s))$ es una función convexa, entonces $\psi(t) = \log(g(s+1))$ es también una función convexa tal que $\psi(0) = 0$. Una función convexa creciente que se anula en 0 es superaditiva, esto es, $\psi(x+y) \geq \psi(x) + \psi(y)$.

Tomando $x = \alpha s$ e $y = (1-\alpha)s$, concluimos la demostración. □

Como α es arbitrario, nos interesa el valor de α que maximice las discrepancias (5.9), (5.10) y (5.11).

Proposición 3. *Sea $\phi(\alpha) = g(1+\alpha(s-1))g(s-\alpha(s-1))$, con $s \neq 1$ siendo $g(s)$ la función generatriz de probabilidad de una distribución de conteo perteneciente a la clase LC, entonces $\phi(\alpha)$ se maximiza para $\alpha = 1/2$.*

Demostración Como $\phi(\alpha)$ es diferenciable, tenemos:

$$\phi'(\alpha) = (1-s)[g'(1+\alpha(s-1))g(s-\alpha(s-1)) - g'(s-\alpha(s-1))g(1+\alpha(s-1))].$$

Para $\alpha = 1/2$:

$$\phi'(1/2) = (1-s)[g'((1+s)/2)g((1+s)/2) - g'((1+s)/2)g((1+s)/2)] = 0 \;\; \forall s.$$

Con el fin de hallar otros puntos críticos suponemos que $\phi'(\alpha) = 0$ para $s \neq 1$. Esto implica:

$$\frac{g'}{g}(1+\alpha(s-1)) = \frac{g'}{g}(s-\alpha(s-1)).$$

$\log(g(s))$ es una función convexa, por lo que su primera derivada es una función creciente. La igualdad,

$$1+\alpha(s-1) = s-\alpha(s-1)$$

solo se cumple para $\alpha = 1/2$, lo cual demuestra que $\alpha = 1/2$ es el único punto crítico.

Además, tenemos que:

$$\begin{aligned} \phi''(\alpha) &= (1-s)^2[2g'(1+\alpha(s-1))g'(s-\alpha(s-1)) \\ - \;& g''(1+\alpha(s-1))g(s-\alpha(s-1)) \\ - \;& g''(s-\alpha(s-1))g(1+\alpha(s-1))] \end{aligned} \tag{5.12}$$

y

$$\phi''(1/2) = 2(1-s)^2((g'(1/2+t/2))^2 - g''(1/2+t/2)g(1/2+t/2)). \tag{5.13}$$

Como $\log(g(s))$ es una función convexa, su segunda derivada es positiva para $s \in (0,1)$. Por tanto:

$$(\log(g))''(s) = \frac{g''(s)g(s) - g'(s)^2}{g(s)^2} \geq 0.$$

Vemos que el segundo factor de (5.12) es el numerador con el signo invertido, Entonces, $g''(1/2) \leq 0$, lo que concluye la demostración. □

5.4. Estadísticos propuestos

Sea $X_1, \ldots, X_n$ observaciones independientes e idénticamente distribuidas provenientes de una variable de conteo. La muestra puede ser resumida mediante las frecuencias: $f_0, f_1, \ldots, f_m$, siendo m la mayor observación: $m = \max\{X_1, \ldots, X_n\}$.

Por la ley de los grandes números sabemos que $g_n(s)$ es un estimador consistente de $g(s)$.

Basados en (5.9) y (5.11) evaluando en $\alpha = 1/2$, siendo las hipótesis alternativas distribuciones de la clase LC, Puig y Weiß (2020) [16] propone rechazar la hipótesis nula para valores grandes de estos dos test:

$$\hat{\Delta}_1 = \int_0^1 [g_n(s) - [g_n(\frac{s+1}{2})]^2] ds. \tag{5.14}$$

$$\hat{\Delta}_\infty = \max_{s\in[0,1]} (g_n(s) - [g_n(\frac{s+1}{2})]^2). \tag{5.15}$$

Para alternativas más generales, fuera de la clase LC y basados en (5.9), (5.10), (5.11), evaluando en $\alpha = 1/2$, Puig y Weiß (2020) [16] propone rechazar la hipótesis nula para valores grandes de los siguientes estadísticos:

$$\hat{\Delta}_1^\star = \int_0^1 |g_n(s) - [g_n(\frac{s+1}{2})]^2| ds. \tag{5.16}$$

$$\hat{\Delta}_2 = \int_0^1 (g_n(s) - [g_n(\frac{s+1}{2})]^2)^2 ds. \tag{5.17}$$

$$\hat{\Delta}_\infty^\star = \max_{s\in[0,1]} (|g_n(s) - [g_n(\frac{s+1}{2})]^2|). \tag{5.18}$$

5.4.1. Extensiones de los estadísticos

Puig y Weiß, de forma similar al esquema de ponderación propuesto por Gürtel y Henze (2000) [7] que hemos visto para desarrollar los test $T_{n,a}$ a partir de T_n y $R_{n,a}$ a partir de R_n, proponen una ponderación de los test.

Este esquema de ponderación propone utilizar para los estadísticos 5.14, 5.16 y 5.17 añadiendo el factor s^a a la integral. Se propone, por tanto, rechazar la hipótesis nula para los siguientes estadísticos:

$$\hat{\Delta}_{1,a} = \int_0^1 g_n(s) - [g_n(s)(\frac{s+1}{2})]^2)s^a ds,$$

$$\hat{\Delta}^\star_{1,a} = \int_0^1 |g_n(s) - [g_n(s)(\frac{s+1}{2})]^2|s^a ds,$$

$$\hat{\Delta}_{2,a} = \int_0^1 (g_n(s) - [g_n(s)(\frac{s+1}{2})]^2)^2 dss^a$$

También se hicieron pruebas con la elaboración de otros esquemas de ponderación para los test como ponderar con $(1-s)^a$ en vez de con s^a, pero los resultados de la potencia de los test fueron mucho menos óptimos.

Capítulo 6

El test de Székely y Rizzo

6.1. Introducción

Basándose en una caracterización por distancias medias de las funciones particularizando para la distribución Poisson, Székely y Rizzo (2004) [20] proponen un test para determinar la Poissonidad de las distribuciones basado en la esperanza de las distancias.

6.2. Caracterización general

Consideremos $E_X[k - X]$ la distancia media de un valor $k \in \mathbb{N}_0$, a una variable discreta X. Esto es, para cada k, $E_X|k - X| = \sum_{j=0}^{\infty} |k - j| f_X(j)$, siendo f_X la función de probabilidad de X.

Teorema 16. *Sean X e Y dos variables aleatorias discretas que toman valores enteros no negativos con $E[X] < \infty$ y $E[Y] < \infty$. Entonces, X e Y están idénticamente distribuidas si y solo si*

$$E_X|k - X| = E_Y|k - Y|, \tag{6.1}$$

para todo k entero no negativo.

Demostración Suponemos que X e Y son variables de valores enteros no negativos con esperanzas finitas. Claramente, (6.1) se cumple para todo valor real k si X e Y están idénticamente distribuidas. Para probar que si (6.1) se cumple, entonces, X e Y están idénticamente distribuidas es suficiente demostrar que el conjunto de distancias $\{E_X|k - X| : k = 0, 1, \ldots\}$ determina de forma única la función de probabilidad de X.

Sea $F_X(k) = \sum_{j=0}^{k} f_X(j)$ la función de distribución acumulada de X, definimos $m_X = E_X|k - X|$ y $\mu = E[X]$. Por tanto $\mu = m_0$, para todo entero positivo k,

$$\begin{aligned} m_k &= \sum_{j=0}^{\infty} |k-j| f_X(j) = 2\sum_{j=0}^{k-1}(k-j)f_X(j) + \mu - k, \\ &= 2kF_X(k-1) - 2\sum_{j=0}^{k-1} j f_X(j) - (k-\mu) \end{aligned}$$

Sea $G_X(k) = \sum_{j=0}^{k} j f_X(j)$, entonces

$$\begin{aligned} m_k &= 2kF_X(k-1) - 2G_X(k-1) - (k-\mu) \\ &= 2(kF_X(k-2) - G_X(k-2)) + 2f_X(k-1) - (k-\mu). \end{aligned}$$

Por tanto

$$f_X(0) = (m_1 + 1 - m_0)/2,$$

y para $k = 1, 2, \ldots,$

$$f_X(k-1) = (m_k - 2(kF_X(k-2) - G_X(k-2)) + k - m_0)/2.$$

Simplificando, obtenemos la fórmula recursiva

$$f_X(k-1) = (m_k - 2\sum_{j=0}^{k-2}(k-j)f_X(j) + k - m_0)/2, \tag{6.2}$$

la cual determina de forma única la distribución de X, por tanto, X e Y están idénticamente distribuidos si $E_X[k-X] = E_Y[k-Y]$ para todo entero $k \geq 0$. $\square$

6.3. Caracterización para la distribución Poisson

Teorema 17. *Sea X una variable aleatoria discreta que toma valores enteros no negativos con función de probabilidad f_X, y función de distribución acumulada F_X y $E[X] < \infty$. Entonces, X tiene una distribución Poisson de media λ si y solo si*

$$E_X|k-X| = 2(k-\lambda)F_X(k-1) + 2\lambda f_X(k-1) - (k-\lambda), \tag{6.3}$$

se cumple para todo entero no negativo k.

Demostración Aplicando el Teorema 16 se demuestra que si X sigue una distribución Poisson, el Teorema 17 es cierto. Para demostrar que la otra implicación es cierta, tenemos que, si X sigue una distribución Poisson de parámetro λ, entonces

$$f_X(k) = e^{-\lambda}\lambda^k/k!$$

y

$$E_X|k-X| = \lambda - k + 2e^{-\lambda}\sum_{j=0}^{k-1}(k-j)\frac{\lambda^j}{j!}.$$

Aplicando las identidades $\lambda f_X(k-1) = kf_X(k-1)$ y $\sum_{j=0}^{k} jf_X(j) = \lambda F_X(k-1)$ se obtiene:

$$\begin{aligned} E_X|k-X| &= \lambda - k + 2kF_X(k-1) - 2\lambda F_X(k-2), \\ &= 2(k-\lambda)F_X(k-1) + 2\lambda f_X(k-1) - (k-\lambda). \end{aligned}$$

□

Bajo una distribución Poisson de parámetro λ, $m_0 = E[X] = \lambda$ y $m_1 = 2f_X(0) - (1-\lambda)$. Por tanto,

$$F_X(0) = f_X(0) = (m_1 + 1 - \lambda)/2.$$

Resolviendo para $f_X(k)$ en (4.8), se obtiene la siguiente fórmula recursiva:

$$f_X(k) = \frac{m_{k+1} - (k+1-\lambda)(2F_X(k-1)-1)}{2(k+1)}, \forall k \geq 1. \tag{6.4}$$

6.4. Estadísticos propuestos

A partir del Teorema 16 se pueden aplicar test de homogeneidad entre muestras. A partir del Teorema 17 se pueden aplicar test de Poissonidad.

Sea $X_1, \ldots, X_n$ una variable aleatoria con función de distribución F_X, F_X puede ser estimada aplicando la ecuación recursiva (6.2) reemplazando m_k por

$$\hat{m}_k = \frac{1}{n}\sum_{j=1}^{n}|k - X_j|, k \in \mathbb{N}_0. \tag{6.5}$$

Para el caso de comprobar la Poissonidad de X, definimos $f(;\lambda)$ como la función de probabilidad y $F(;\lambda)$ como la función de distribución acumulada de una Poisson de parámetro λ. m_k se estima a partir de (6.5). Entonces, $\hat{m}_0 = \hat{\lambda}$, donde $\lambda = \frac{1}{n}\sum_{i=1}^{n} X_i$ y $\hat{F_X}(0) = \hat{f_X}(0) = (\hat{m}_1 + 1 - \hat{\lambda})/2$.

Sustituyendo el estimador $\{\hat{m}_k\}$ en (6.4), $\hat{f_X}$ y $\hat{F_X}$ son determinados por la fórmula recursiva

$$\hat{f_X}(k) = \frac{\hat{m}_{k+1} - (k+1-\hat{\lambda})(2\hat{F}_X(k-1)-1)}{2(k+1)}.$$

Székely y Rizzo (2004) proponen rechazar la hipótesis nula para valores grandes del siguiente estadístico, basado en la proximidad entre la función de distribución estimada de X, caso de que siga una distribución Poisson, y la función de distribución acumulada de la distribución Poisson, usando la distancia de Cramér Von Misses:

$$M_n = n \sum_{0}^{\infty} (\hat{F}_X(j) - F(j;\hat{\lambda}))^2 f(j;\hat{\lambda}).$$

El cálculo de los puntos críticos de este test se lleva a cabo mediante un bootstrap paramétrico que veremos en el capítulo siguiente.

Capítulo 7

Simulaciones siguiendo el método bootstrap paramétrico

7.1. Método de aproximación bootstrap paramétrico

En los casos que nos ocupan, para obtener una aproximación a la distribución nula no podemos usar la clásica distribución asintótica, pues no aporta una aproximación útil debido a que depende de un parámetro desconocido, λ. Es por esto que usaremos el método del bootstrap paramétrico, descrito en Gürtler y Henze [7].

Sea $W_n = W_n(X_1, \ldots, X_n)$ un estadístico para el contraste (1.1). Supongamos que la hipótesis nula H_0 es rechazada si

$$W_n(X_1, \ldots, X_n) > c = c(n, \alpha, \lambda).$$

El valor de $c = c(n, \alpha, \lambda)$ depende del tamaño de la muestra n, del nivel de significación α y también del "verdadero" y desconocido valor del parámetro λ. El bootstrap paramétrico aproxima la distribución nula de W_n, $P(W_n \leq w; \lambda)$, donde incluimos λ como argumento de la función de probabilidad para enfatizar que depende de este parámetro, mediante la distribución condicional, dada la muestra, del estadístico con el parámetro igual a $\hat{\lambda} = \bar{X}_n$, esto es, estima

$$H(w) = P(W_n \leq w; \lambda)$$

mediante

$$H^*(w) = P(W_n^* \leq w; \hat{\lambda} \mid X_1, \ldots, X_n) := P^*(W_n^* \leq w; \hat{\lambda}),$$

donde $W_n^* = W_n(X_1^*, \ldots, X_n^*)$ es el estadístico W_n evaluado en $X_1^*, \ldots, X_n^*$ que es una muestra procedente de una ley $Pois(\hat{\lambda})$. Decimos que la aproximación es en la distribución condicional, dada la muestra, porque en la aproximación $\hat{\lambda}$ es considerado como un parámetro fijo, esto es, en la aproximación bootstrap la variabilidad de W_n^* es debida a que las muestras proceden de una ley $Pois(\hat{\lambda})$, ignorando la variabilidad de $\hat{\lambda}$.

Intuitivamente, como $\hat{\lambda}$ es un buen estimador de λ, esto es, $\hat{\lambda}$ "está cerca"de λ, cabe esperar que las leyes $Pois(\hat{\lambda})$ y $Pois(\lambda)$ sean "similares", y por tanto que $P(W_n \leq w; \lambda)$ y $P_*(W_n^* \leq w; \hat{\lambda})$ también estén "próximas".

Nótese que en la distribución bootstrap todos los parámetros son conocidos, y por tanto, en teoría, podría ser calculada exactamente, bien analíticamente o bien numéricamente, generando todas las posibles muestras. Desde un punto de vista práctico, ambas son inviables, por lo que en la práctica, la distribución bootstrap es numéricamente aproximada mediante simulación como sigue.

(a) Para $b = 1, \ldots, B$ repetir

(a.1) Generar $X_1^{*b}, \ldots, X_n^{*b}$, independientes e idénticamente distribuidas según una ley $Pois(\hat{\lambda})$ (las muestras bootstrap).

(a.2) Calcular $W_n^{*b} = W_n(X_1^{*b}, \ldots, X_n^{*b})$ (las replicaciones bootstrap del estadístico del contraste).

(b) Aproximar $P^*(W_n^* \leq w; \hat{\lambda})$ mediante $P_B^*(W_n^* \leq w; \hat{\lambda})$ que es la función de distribución empírica asociada a las replicaciones bootstrap del estadístico, $W_n^{*1}, \ldots, W_n^{*B}$, es decir,

$$H_B^*(w) = \frac{1}{B} \sum_{b=1}^{B} I(W_n^{*b} \leq w).$$

En particular, el punto crítico $c = c(n, \alpha, \lambda)$ es estimado mediante

$$c_B^*(n, \alpha, \hat{\lambda}) = H_B^{*\,-1}(w)(\alpha) = \text{ínf}\{x \,:\, H_B^*(w) \geq \alpha\} = W_n^{*([B\alpha])},$$

donde $W_n^{*(1)} \leq \ldots \leq W_n^{*(B)}$ son las las replicaciones bootstrap del estadístico ordenadas en orden creciente y $[\cdot]$ es la función parte entera. Equivalentemente, si denotamos mediante $W_{n,obs} = W_n(X_1, \ldots, X_n)$ al valor observado del estadístico en la muestra original, el p-valor es estimado mediante

$$p_B^* = \frac{1}{B} \sum_{b=1}^{B} I(W_n^{*b} \geq W_{n,obs}).$$

Valores usuales para B son $B = 500, 1000, 2000$.

A continuación veremos la forma en que hemos llevado a cabo la programación de este método para determinar el p-valor de los diferentes estadísticos asociados a una muestra determinada.

Lo veremos para un estadístico general al que llamaremos W_n, siendo ***"n.boot"*** el valor del parámetro B asociado a una muestra determinada, a la que llamaremos ***"sample"***.

Algoritmo 1: Programación en R de la determinación del pvalor para un estadístico general con $B = n.boot$ asociado a una muestra, ***"sample"***

```
pvalores = 0;
x = sample;
lambda.hat = mean(x);
T.obs = Wn(x);
para i ∈ (1 : n.boot) hacer
    x.boot = rpois(length(x), lambda.hat);
    T.boot = Wn(x.boot);
    pvalores = (pvalores + as.numeric(T.boot > T.obs));
fin
pvalores = pvalores / n.boot
```

7.2. Simulaciones

En esta sección vamos a comprobar numéricamente la bondad de aproximación bootstrap a la distribución nula de los estadísticos considerados para distintos valores del parámetro λ de la ley Poisson. También vamos a comparar la potencia de algunos de los test definidos previamente frente a distribuciones alternativas.

7.2.1. Bondad de la aproximación bootstrap a la distribución nula de los estadísticos

En este apartado estudiaremos numéricamente si el nivel del test con p-valores bootstrap es cercano al valor fijado, α.

A continuación detallaremos las distribuciones elegidas para determinar la bondad de aproximación bootstrap:

- $Pois(1)$: Distribución Poisson con media, $\lambda = 1$.
- $Pois(5)$: Distribución Poisson con media, $\lambda = 5$.
- $Pois(10)$: Distribución Poisson con media, $\lambda = 10$

En las siguientes tablas veremos el porcentaje de veces que la hipótesis nula es rechazada para varios de los estadísticos mencionados en la memoria en mil muestras de las distribuciones Poisson anteriormente mencionadas con $n = 20, 30, 40$ y $B = 1000$ para valores críticos de $\alpha = 0.05$ y $\alpha = 0.10$. En rojo están marcadas las observaciones en que la diferencia entre el porcentaje de veces que la hipótesis nula es rechazada y el valor crítico de α es mayor o igual a 0.15.

Distrib. $n=20$	α	T_n	$T_{n,5}$	$\hat{\Delta}_1$	$\hat{\Delta}_2^\star$	$\hat{\Delta}_\infty$	R_n	$R_{n,3}$	$\hat{\Delta}_{1,5}$	$\hat{\Delta}_{2,5}^\star$	u	K_n	C_n	C_n*	L_n
$Pois(1)$	0.05	0.063	0.051	0.052	0.068	0.070	0.069	0.058	0.056	0.053	0.049	0.039	0.059	0.056	0.043
	0.1	0.105	0.112	0.099	0.105	0.109	0.117	0.114	0.114	0.110	0.108	0.082	0.114	0.119	0.098
$Pois(5)$	0.05	0.055	0.055	0.051	0.052	0.052	0.044	0.049	0.053	0.058	0.052	0.032	0.044	0.042	0.054
	0.1	0.111	0.103	0.100	0.098	0.098	0.102	0.098	0.111	0.110	0.090	0.088	0.095	0.099	0.099
$Pois(10)$	0.05	0.049	0.052	0.048	0.046	0.050	0.046	0.048	0.047	0.044	0.054	0.044	0.045	0.051	0.045
	0.1	0.095	0.098	0.094	0.086	0.085	0.084	0.091	0.093	0.095	0.103	0.077	0.097	0.094	0.102

Cuadro 7.1: Proporción de veces que la hipótesis nula se rechaza bajo H_0.

Distrib. $n=30$	α	T_n	$T_{n,5}$	$\hat{\Delta}_1$	$\hat{\Delta}_2^\star$	$\hat{\Delta}_\infty$	R_n	$R_{n,3}$	$\hat{\Delta}_{1,5}$	$\hat{\Delta}_{2,5}^\star$	u	K_n	C_n	C_n*	L_n
$Pois(1)$	0.05	0.037	0.034	0.046	0.037	0.040	0.040	0.036	0.045	0.049	0.050	0.051	0.061	0.054	0.049
	0.1	0.082	0.089	0.108	0.084	0.086	0.089	0.093	0.092	0.095	0.115	0.119	0.105	0.106	0.100
$Pois(5)$	0.05	0.050	0.060	0.052	0.050	0.044	0.047	0.058	0.062	0.061	0.057	0.054	0.048	0.049	0.049
	0.1	0.105	0.100	0.097	0.093	0.106	0.097	0.100	0.103	0.106	0.104	0.104	0.089	0.101	0.084
$Pois(10)$	0.05	0.051	0.053	0.052	0.044	0.048	0.046	0.052	0.055	0.054	0.056	0.056	0.046	0.056	0.055
	0.1	0.090	0.092	0.107	0.091	0.091	0.087	0.089	0.105	0.117	0.097	0.094	0.094	0.104	0.097

Cuadro 7.2: Proporción de veces que la hipótesis nula se rechaza bajo H_0.

Distrib. $n = 40$	α	T_n	$T_{n,5}$	$\hat{\Delta}_1$	$\hat{\Delta}_2^\star$	$\hat{\Delta}_\infty$	R_n	$R_{n,3}$	$\hat{\Delta}_{1,5}$	$\hat{\Delta}_{2,5}^\star$	u	K_n	C_n	C_n*	L_n
$Pois(1)$	0.05	0.041	0.047	0.061	0.045	0.047	0.046	0.048	0.059	0.044	0.043	0.060	0.054	0.064	0.038
	0.1	0.099	0.095	0.110	0.100	0.102	0.103	0.103	0.117	0.099	0.094	0.113	0.096	0.098	0.087
$Pois(5)$	0.05	0.047	0.049	0.051	0.045	0.036	0.048	0.048	0.057	0.044	0.052	0.061	0.054	0.056	0.056
	0.1	0.098	0.108	0.114	0.105	0.094	0.095	0.105	0.103	0.085	0.098	0.096	0.102	0.102	0.103
$Pois(10)$	0.05	0.054	0.058	0.048	0.049	0.051	0.047	0.055	0.044	0.041	0.098	0.096	0.102	0.102	0.103
	0.1	0.107	0.117	0.101	0.096	0.106	0.103	0.110	0.089	0.089	0.098	0.096	0.102	0.102	0.103

Cuadro 7.3: Proporción de veces que la hipótesis nula se rechaza bajo H_0.

En la Tabla 7.1 observamos que el nivel real de los test se ajusta bastante bien al valor teórico α. No obstante, algunos estadísticos como $\hat{\Delta}_2^{\star}$, $\hat{\Delta}_\infty$ o R_n tienen desviaciones más o menos considerables con respecto al nivel real del test, sobre todo en la distribución Poisson de parámetro $\lambda = 1$. Esto puede ser debido a que el valor de n no es excesivamente grande y a que al ser $\lambda = 1$ pueden aparecer muchos 0s.

Para evitar el problema del tamaño de la muestra, vamos a probar esta misma simulación para valores de $n = 30$. En este caso, como podemos ver en la Tabla 7.2, el número de casillas en los que la diferencia entre el nivel real del test y el porcentaje de observaciones en los que la hipótesis nula se rechaza difieren en más de 0.15, baja levemente respecto a $n = 20$, pero la bajada apenas se observa con claridad y no es clara. La distribución $Pois(1)$ sigue siendo la que más problemas presenta.

En la Tabla 7.3 observamos que, para $n = 40$, ya el número de casillas en los que la diferencia entre el nivel real del test y el porcentaje de observaciones en los que la hipótesis nula se rechaza difieren en más de 0.15 baja de forma bastante considerable respecto a $n = 30$ y $n = 20$. Además, en la distribución $Pois(1)$ ya no existen tantas desviaciones como para $n = 20$ o $n = 30$.

Esto nos lleva a confirmar que la igualdad entre el nivel real de los test el porcentaje de observaciones en los que la hipótesis nula se rechaza se va estabilizando a medida que va aumentando el tamaño de la muestra.

7.2.2. Potencia frente a alternativas

En este apartado compararemos numéricamente la potencia de los test considerados frente a distintas hipótesis alternativas.

A continuación, detallaremos las distribuciones elegidas para determinar la potencia:

- $B(10, 0.5)$: Distribución binomial con número de ensayos, $n = 10$ y probabilidad de éxito, $p = 0.5$.
- $B(10, 0.1)$: Distribución binomial con número de ensayos, $n = 10$ y probabilidad de éxito, $p = 0.1$.
- $NB(5, 0.5)$: Distribución binomial negativa con número deseado de resultados favorables, $k = 5$ y con una probabilidad de éxito, $p = 0.5$.
- $NB(10, 2/3)$: Distribución binomial negativa con número deseado de resultados favorables, $k = 5$ y con una probabilidad de éxito, $p = 0.5$.
- $Pois(1) : Pois(5)$: Mezcla 50-50 de dos distribuciones Poisson, una con media, $\lambda_1 = 1$ y otra con media, $\lambda_2 = 5$.

- $Pois(3) : Pois(5)$: Mezcla 50-50 de dos distribuciones Poisson, una con media, $\lambda_1 = 3$ y otra con media, $\lambda_2 = 5$.
- $ZIP(0.3, 3)$: Distribución Poisson cero-inflada. En este caso, una mezcla $0.3 : 0.7$ de dos componentes, una constante de ceros y una Poisson de media, $\lambda = 3$ respectivamente.
- $ZIP(0.2, 5)$: Distribución Poisson cero-inflada. En este caso, una mezcla $0.2 : 0.8$ de dos componentes, una constante de ceros y una Poisson de media, $\lambda = 5$ respectivamente.

En las Tablas 7.4 y 7.5 veremos el porcentaje de veces que la hipótesis nula es rechazada para varios de los estadísticos mencionados en la memoria en mil muestras de las distribuciones alternativas anteriormente mencionadas con $n = 20, 30, 40$ y $B = 1000$ y puntos críticos de $\alpha = 0.05$ y $\alpha = 0.1$.

Los valores en negrita indican los dos estadísticos para los que la hipótesis nula es rechazada en un mayor número de veces para cada distribución alternativa.

Podemos extraer las siguientes conclusiones de las Tablas 7.4 y 7.5:

1. Exceptuando para las mezclas de distribuciones Poisson y para las distribuciones cero-infladas de la Poisson, $R_{n,a}$ y $T_{n,a}$ mejora la potencia de R_n y T_n.
2. Para las distribuciones pertenecientes a la clase LC los test propuestos por Puig y Weißtienen una potencia muy elevada. Esto confirma que, como vimos en el Capítulo 5, estos test están propuestos de forma específica contra alternativas de la clase LC.
3. En contraparte, para las distribuciones binomiales, los test propuestos por Puig y Weiß tienen una potencia muy baja.
4. El test u se basa únicamente en los dos primeros momentos del conjunto de datos, por lo que depende exclusivamente de estos si su potencia es mayor o menor.
5. Sorprende que el nivel real de algunos de los test se ajustan más o menos bien a la distribución $B(10, 0.1)$. Esto es debido al Teorema 18.
6. La potencia de los test aumenta generalmente de forma suave a medida que aumenta el valor de n.
7. El test C_n* no mejora especialmente la potencia del test C_n, salvo cuando la hipótesis alternativa es una distribución binomial.
8. Dentro de los test basados en la función de distribución empírica, K_n y L_n son los test que mejores resultados dan en cuanto a potencia.

9. En términos generales, los test basados en la función generatriz de probabilidad dan resultados más o menos semejantes a los de los test basados en la función de distribución empírica.

Teorema 18. *Sea $X_1, \ldots, X_n$ una variable aleatoria que sigue una distribución binomial*

$$X \sim B(n, p),$$

para valores de n grande y p pequeña la distribución X se aproxima a una distribución $Pois(\lambda)$, siendo $\lambda = n * p$.

Demostración: Suponemos X sigue una distribución $B(n, p)$ y sea n grande, p pequeña y $\lambda = n * p$. Entonces:

$$\begin{aligned} P(X=i) = \frac{n!}{(n-i)!i!} p^i (1-p)^{n-i} = \\ \frac{n!}{(n-i)!i!} (\frac{\lambda}{n})^i (1-\frac{\lambda}{n})^{n-i} = \\ \frac{n(n-1)\ldots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}. \end{aligned} \tag{7.1}$$

Para n grande y p pequeña, con $\lambda = np$ constante tenemos que:

$$(1-\frac{\lambda}{n})^n \approx e^{-\lambda},$$

$$\frac{n(n-1)\ldots(n-i+1)}{n^i} \approx 1,$$

$$(1-\frac{\lambda}{n})^i \approx 1,$$

Entonces, para n grande y p pequeña tenemos que

$$P(X=i) \approx e^{-\lambda} \frac{\lambda^i}{i!},$$

y por tanto, X sigue aproximadamente una distribución $Pois(\lambda)$.

Distrib ($n = 20$)	α	T_n	$T_{n,5}$	$\hat{\Delta}_1$	$\hat{\Delta}_2^\star$	$\hat{\Delta}_\infty$	R_n	$R_{n,3}$	$\hat{\Delta}_{1,5}$	$\hat{\Delta}_{2,5}^\star$	u	K_n	C_n	C_n*	L_n
$B(10, 0.5)$	0.05	0.015	**0.351**	0.001	0.005	0.013	0.011	0.300	0.000	0.306	0.000	0.139	0.286	0.294	**0.461**
	0.1	0.226	**0.570**	0.011	0.041	0.061	0.216	0.534	0.000	0.521	0.000	0.272	0.402	0.406	**0.565**
$B(10, 0.1)$	0.05	0.056	0.050	0.034	0.066	0.062	0.064	0.056	0.002	0.005	0.022	0.033	**0.072**	0.064	**0.068**
	0.1	0.114	0.117	0.066	0.120	0.116	0.126	0.124	0.005	0.094	0.065	0.077	0.124	**0.137**	**0.140**
$NB(5, 0.5)$	0.05	0.541	0.622	0.407	0.398	0.288	0.527	0.616	**0.646**	0.581	**0.695**	0.394	0.354	0.242	0.303
	0.1	0.658	0.691	0.550	0.527	0.474	0.634	0.681	**0.758**	0.646	**0.779**	0.520	0.464	0.339	0.432
$NB(10, 2/3)$	0.05	0.235	0.277	0.189	0.172	0.112	0.231	0.266	**0.354**	0.287	**0.364**	0.173	0.156	0.132	0.126
	0.1	0.334	0.368	0.293	0.263	0.230	0.314	0.345	**0.467**	0.355	**0.500**	0.286	0.239	0.195	0.219
$Pois(1) : Pois(5)$	0.05	0.830	**0.908**	0.791	0.714	0.570	0.846	0.900	**0.919**	0.877	**0.919**	0.824	0.746	0.568	0.591
	0.1	0.886	0.933	0.864	0.781	0.671	0.892	0.933	**0.960**	0.909	**0.958**	0.895	0.824	0.681	0.725
$Pois(3) : Pois(5)$	0.05	0.132	0.148	0.112	0.098	0.075	0.121	0.146	**0.171**	0.129	**0.213**	0.100	0.084	0.082	0.048
	0.1	0.207	0.211	0.196	0.153	0.122	0.190	0.197	**0.271**	0.183	**0.317**	0.166	0.137	0.135	0.110
$ZIP(0.3, 3)$	0.05	**0.894**	0.804	**0.938**	0.881	0.845	0.875	0.818	0.874	0.843	0.726	0.797	0.620	0.514	0.381
	0.1	**0.933**	0.860	**0.968**	0.932	0.907	0.922	0.884	0.932	0.891	0.832	0.869	0.723	0.629	0.638
$ZIP(0.2, 5)$	0.05	0.951	0.884	0.952	**0.954**	**0.958**	0.946	0.906	0.929	0.934	0.760	0.647	0.499	0.454	0.502
	0.1	0.967	0.909	**0.972**	**0.972**	**0.973**	0.960	0.928	0.957	0.947	0.836	0.791	0.611	0.552	0.632

Cuadro 7.4: Proporción de veces que la hipótesis nula se rechaza bajo H_0.

Distrib. ($n = 30$)	α	T_n	$T_{n,5}$	$\hat{\Delta}_1$	$\hat{\Delta}_2^\star$	$\hat{\Delta}_\infty$	R_n	$R_{n,3}$	$\hat{\Delta}_{1,5}$	$\hat{\Delta}_{2,5}^\star$	u	K_n	C_n	C_n*	L_n
$B(10, 0.5)$	0.05	0.177	**0.569**	0.001	0.001	0.009	0.159	**0.514**	0.000	0.510	0.000	0.258	0.445	0.489	0.695
	0.1	0.501	**0.746**	0.003	0.068	0.041	0.507	0.714	0.000	0.707	0.000	0.424	0.590	0.626	**0.793**
$B(10, 0.1)$	0.05	0.067	0.057	0.026	0.072	**0.076**	0.069	0.060	0.018	0.056	0.018	0.048	0.058	0.063	**0.075**
	0.1	0.135	0.118	0.054	**0.137**	**0.149**	0.133	0.122	0.039	0.110	0.044	0.094	0.110	0.118	0.124
$NB(5, 0.5)$	0.05	0.690	**0.763**	0.503	0.497	0.417	0.571	0.750	0.762	0.691	**0.813**	0.580	0.472	0.349	0.435
	0.1	0.769	0.814	0.635	0.614	0.548	0.755	0.798	**0.854**	0.758	**0.893**	0.700	0.593	0.464	0.571
$NB(10, 2/3)$	0.05	0.327	0.379	0.265	0.253	0.178	0.313	0.369	**0.417**	0.338	**0.479**	0.243	0.195	0.155	0.173
	0.1	0.434	0.455	0.363	0.340	0.292	0.406	0.441	**0.549**	0.409	**0.592**	0.370	0.282	0.236	0.263
$Pois(1) : Pois(5)$	0.05	0.842	0.898	0.795	0.737	0.585	0.852	0.895	**0.991**	**0.979**	0.896	0.919	0.868	0.772	0.811
	0.1	0.890	0.927	0.871	0.797	0.699	0.892	0.921	**0.994**	**0.988**	0.948	0.946	0.930	0.855	0.894
$Pois(3) : Pois(5)$	0.05	0.131	0.134	0.116	0.099	0.074	0.122	0.128	**0.209**	0.142	**0.185**	0.136	0.101	0.110	0.090
	0.1	0.200	0.201	0.220	0.168	0.125	0.182	0.200	**0.325**	0.210	**0.279**	0.224	0.191	0.183	0.179
$ZIP(0.3, 3)$	0.05	**0.968**	0.933	**0.981**	**0.968**	0.947	0.966	0.941	0.954	0.943	0.855	0.826	0.768	0.723	0.586
	0.1	**0.981**	0.961	**0.992**	**0.981**	0.977	0.979	0.968	0.977	0.965	0.941	0.878	0.858	0.812	0.715
$ZIP(0.2, 5)$	0.05	0.953	0.868	0.954	0.953	0.952	0.950	0.893	**0.974**	**0.974**	0.751	0.641	0.674	0.650	0.710
	0.1	0.968	0.899	0.968	0.968	0.972	0.963	0.921	**0.987**	**0.983**	0.837	0.802	0.764	0.735	0.815

Cuadro 7.5: Proporción de veces que la hipótesis nula se rechaza bajo H_0.

Capítulo 8

Estudios computacionales

A la hora de programar, en nuestro caso, en el lenguaje R, un estadístico, que no es más que una función, para posteriormente programar un test basado en ese estadístico, en ocasiones existe más de una alternativa. Para elegir una alternativa con respecto a otra nos basaremos exclusivamente en el tiempo computacional que tarda en llevar a cabo la ejecución del test. El objetivo de este capítulo no es otro que determinar para varios estadísticos cuál es la mejor alternativa a utilizar a la hora de realizar su programación en R.

En ocasiones, ocurre que para unos tamaños de muestra es preferible una alternativa mientras que para otros tamaños de muestra, es preferible otra.

8.1. Estudio computacional para el test de Baringhaus y Henze

Para el cálculo del test T_n a la hora de realizar su programación computacional hay, al menos, dos alternativas. Estas son (4.2) y (8.1), que veremos a continuación. Para programarlas en el programa R, se llevan a cabo los siguientes métodos.

- Para (4.2) utilizando la función ***integrate***, integrada en el programa R, la cual desarrolla integración numérica de funciones reales de una variable.

- Para (8.1) se programa de forma sencilla mediante funciones de R.

$$T_n = \frac{1}{n}\sum_{i,j=1}^{n}[\frac{\bar{X}_n^2}{X_i + X_j - 1} + \frac{X_i X_j}{X_i + X_j - 1}] - \bar{X}_n[n - \frac{1}{n}(\sum_{i=1}^{n} 1\{X_i = 0\})^2] \tag{8.1}$$

Por otro lado, para el cálculo del test $T_{n,a}$ a la hora de realizar la programación computacional hay también, al menos, dos alternativas. Una de ellas es (4.18), programable en R mediante la función ***integrate***. La otra, mediante cálculo numérico, la veremos a continuación:

$$T_{n,a} = \frac{1}{n}\sum_{i,j=1}^{n}\left[\frac{\hat{X}_n^2}{X_i + X_j + a + 1} - \frac{\bar{X}_n(X_i + X_j)}{X_i + X_j + a} + \frac{X_i X_j}{X_i + X_j + a - 1}\right] \tag{8.2}$$

A continuación veremos en el programa R la programación del cálculo de los estadísticos T_n y $T_{n,a}$ mediante integración (4.2) y (4.18) y mediante cálculo numérico (8.1) y (8.2). También determinaremos que mediante ambos métodos el estadístico ha sido bien calculado comprobando que para una misma muestra devuelve el mismo valor.

Cálculo numérico e integración numérica para Test de Baringhaus

Cálculo numérico, Tn

```
fTn=function(z) {auxsuma=outer(z,z,"+")
  auxmult=outer(z,z,"*")
  tt=table(z)

  dd=as.numeric(names(tt))+1
  ni=rep(0,(max(z)+1))
  ni[dd]=as.numeric(tt)



  mauxsum=auxsuma[auxmult>0]
  mauxmult=auxmult[auxmult>0]



  TN=(1/length(z))*(sum(mean(z)^2/(auxsuma+1))+sum(mauxmult/(mauxsum-1))) -
  mean(z)*(length(z)-(1/length(z))*(ni[1])^2)
  return(TN)
  }
```

Integración numérica Tn

```
fTn1=function(z) {
  f2noder= function(t) {mean(z) * mean(sum(t^z))}
  f2der= function(t) {mean(sum(t^z*z/t))}
  f3_a=function(t,b) {((f2noder(t) - f2der(t)) ^2)}
  b_a= integrate(Vectorize(f3_a, vectorize.args = 't'), lower=0, upper=1)
  b_aSol=as.numeric(b_a[1])
  return(b_aSol/length(z))
}
```

Prueba Tn

Comprobamos que los resultados mediante ambos métodos son los mismos.

```
X = rpois(20,4)
(fTn(X))
```

```
## [1] 0.07222445
```

```
(fTn1(X))
```

```
## [1] 0.07222445
```

Integración numérica Tn,a

```
fTn_a=function(z,a) {
    f2noder= function(t) {mean(z) * mean(sum(t^z))}
    f2der= function(t) {mean(sum(t^z*z/t))}
    f3_a=function(t,b) {((f2noder(t) - f2der(t)) ^2) * t^b}
    b_a= integrate(Vectorize(f3_a, vectorize.args = 't'), lower=0, upper=1, b = a)
    b_aSol=as.numeric(b_a[1])
    return(b_aSol/length(z))
  }
```

Cálculo numérico Tn,a

```
fTn_a1 = function(x,a){
  aux = outer(x,x,"+")
  aux1 = outer(x,x,"*")
  num1 = mean(x)^2
  den1 = aux + a + 1
  coc1 = sum(num1/den1)
  num2 = mean(x)*(aux)
  den2 = aux + a
  coc2 = sum(num2/den2)
  num3 = aux1
  den3 = aux + a - 1
  coc3 = sum(num3/den3)
  (coc1 - coc2 + coc3)/length(x)
}
```

Prueba Tn,a

Comprobamos que los resultados mediante ambos métodos son los mismos.

```
X = rpois(20,4)
(fTn_a(X,3))
```

```
## [1] 0.007450308
```

```
(fTn_a1(X,3))
```

```
## [1] 0.007450308
```

8.1.1. Cálculo de eficiencia

En la imagen 8.1 veremos el tiempo computacional respecto al valor de n que se emplea para la ejecución de un test a partir del estadístico T_n para valores de $B = 1000$, para $10 \leq n \leq 200$, y para $\lambda = 4$.

La línea verde representa el cálculo numérico y la roja la integración numérica.

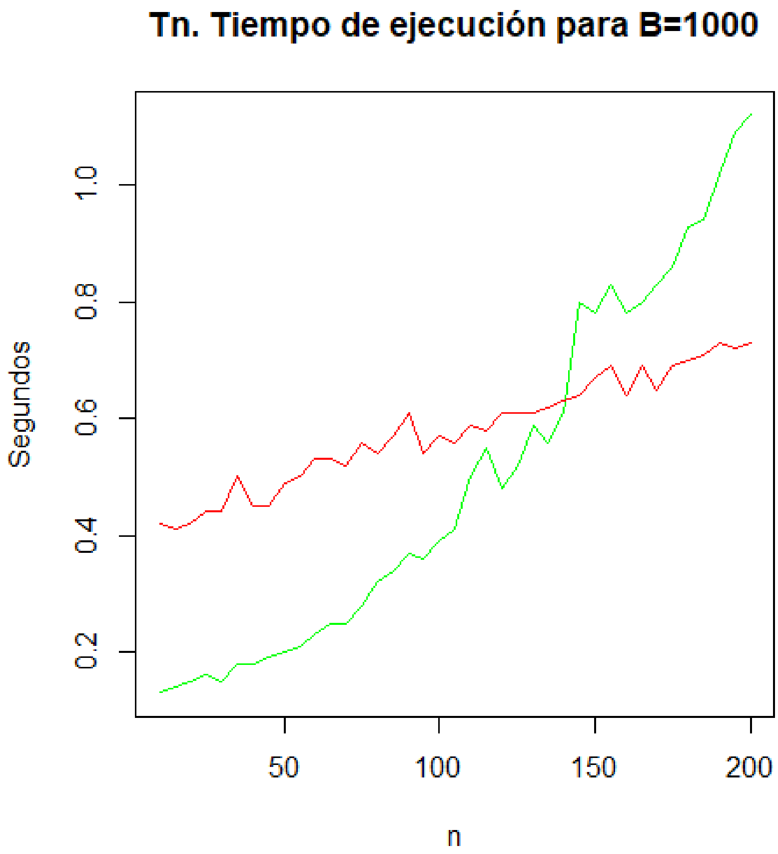


Figura 8.1

Conclusión: Podemos concluir viendo la gráfica que para valores mayores a $n \sim 140$ la integración numérica es la alternativa más eficiente en términos de tiempo computacional , mientras que para valores menores a $n \sim 140$ lo es el cálculo numérico.

Lógicamente, el valor de B no afecta a la evolución del tiempo computacional. Esto es aplicable para el estudio de todos los test.

En la siguiente imagen, 8.2, veremos el tiempo computacional respecto al valor de n que se emplea para la ejecución de un test a partir del estadístico $T_{n,3}$ para valores de $B = 1000$, para $10 \leq n \leq 200$, y para $\lambda = 4$.

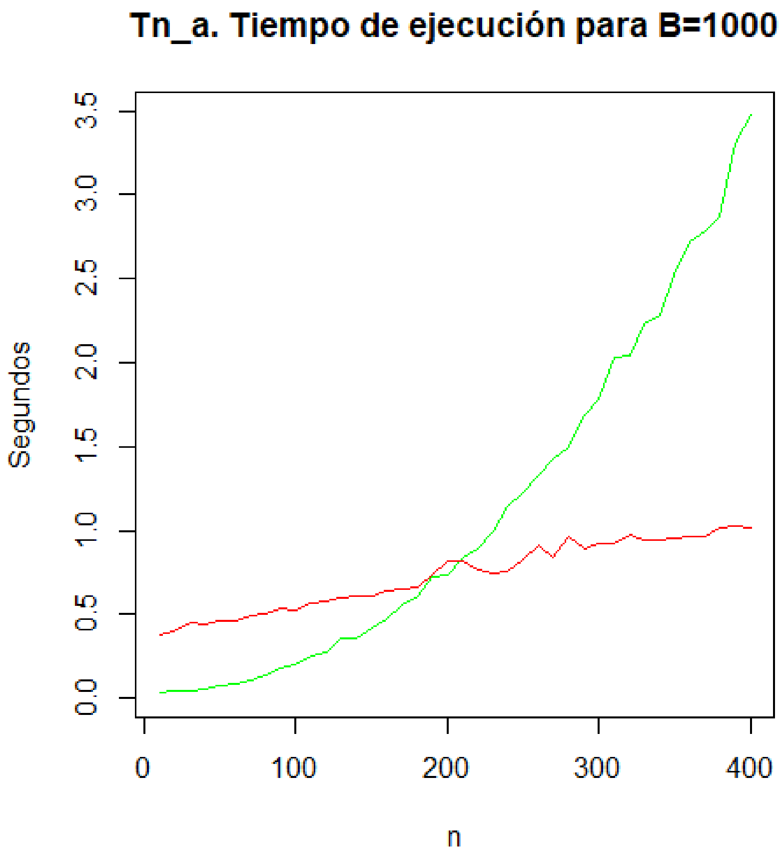


Figura 8.2

Conclusión: En este caso concluimos que para valores mayores a n 200 la integración numérica es la alternativa más eficiente en términos de tiempo computacional, mientras que para valores menores a $n \sim 200$ lo es el cálculo numérico.

8.2. Estudio computacional para los test de Puig y Weiß

Para el calculo de los test (5.14) y (5.17), a la hora de programarlos computacionalmente en el programa R existen, al menos, las siguientes dos alternativas:

- Utilizando la función ***integrate***, integrada en el programa R, la cual desarrolla integración numérica de funciones reales de una variable.

- Mediante cálculos podemos solucionar la integral en cuestión y programar esta solución en el programa R. A este método lo llamaremos cálculo numérico.

Este es el desarrollo del cálculo numérico de $\hat{\Delta}_1$ y $\hat{\Delta}_2^{\star}$:

$$\begin{aligned}\hat{\Delta}_1 &= \int_0^1 [\frac{1}{n}\sum_{i=1}^{n} t^{X_i} - \frac{1}{n^2}\sum_{i,j=1}^{n}\left(\frac{t+1}{2}\right)^{X_i+X_j}]dt \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 t^{X_i}dt - \frac{1}{n^2}\int_0^1 (\frac{1}{n^2})^{X_i+X_j}dt \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i+1} - \frac{1}{n^2}\sum_{i,j=1}^{n}\frac{2-2^{-X_i-X_j}}{X_i+X_j+1}.\end{aligned} \tag{8.3}$$

$$\begin{aligned}&\hat{\Delta}_2^{\star} = \int_0^1 (\hat{g}(t)^2 - 2\hat{g}(t)[\hat{g}(\frac{t+1}{2})]^2 + [\hat{g}(\frac{t+1}{2})]^4)dt = \\ &\frac{1}{n^2}\sum_{i,j=1}^{n}\int_0^1 t^{X_i+X_j}dt - \frac{2}{n^3}\sum_{i,j,k=1}^{n}\int_0^1 t^{X_i}(\frac{t+1}{2})^{X_j+X_k}dt + \\ &\frac{1}{n^4}\sum_{i,j,k,l=1}^{n}\int_0^1 (\frac{t+1}{2})^{X_i+X_j+X_k+X_l}dt = \\ &\frac{1}{n^2}\sum_{i,j=1}^{n}\frac{1}{X_i+X_j+1} + \frac{1}{n^4}\sum_{i,j,k,l=1}^{n}\frac{2-2^{-X_i-X_j-X_k-X_l}}{X_i+X_j+X_k+X_l+1} - \\ &\frac{2}{n^3}\sum_{i,j,k=1}^{n}2^{-X_j-X_k}\sum_{r=0}^{X_j+X_k}\binom{X_j+X_k}{r}\frac{1}{X_i+r+1}.\end{aligned} \tag{8.4}$$

A continuación veremos en el programa R la programación del cálculo de los estadísticos $\hat{\Delta}_1$ y $\hat{\Delta}_2^{\star}$ mediante integración (5.17) y (5.14) y mediante cálculo numérico (8.3) y (8.4). También determinaremos que mediante ambos métodos el estadístico ha sido bien calculado comprobando que para una misma muestra devuelve el mismo valor.

Cálculo numérico e integración numérica para Test Delta

Cálculo numérico, Delta1

```
fdelta1=function(z) {
    aux=outer(z,z,"+")
    num= 2-2^(-aux)
    den=aux+1
    coc=num/den
    mean(coc)
    Delta1Clas=mean(1/(z+1))-mean(coc)
    Delta1Clas
  }
```

Integración numérica Delta1

```
fdelta1b=function(z){
  fdelta1b1 =function(y,t) {
    (1/length(y)) * sum(t^y) -
          ((1/length(y)) * sum(((t+1)/2)^y))^2
  }
  i =integrate(f = Vectorize(fdelta1b1,vectorize.args = 't'), lower = 0, upper = 1,
              y = z)
  i$value
}
```

Prueba Delta1

Comprobamos que los resultados mediante ambos métodos son los mismos.

```
X = rpois(20,4)
(fdelta1(X))
```

```
## [1] -0.01305832
```

```
(fdelta1b(X))
```

```
## [1] -0.01305832
```

Integración numérica Delta2

```
fdelta2=function(y) {
    aux=outer(y,y,"+")
    aux2=outer(aux,aux,"+")
    num1=2-2^-aux2
    den1=aux2+1
    coc1=num1/den1
    length(coc1)
    sum2=mean(coc1)
```

```r
    aux=as.vector(outer(y,y,"+"))
    num2=1
    den2=aux+1
    coc2=num2/den2
    coc2
    sum1=mean(coc2)
    ff1=function(u,x){
      z=0:u;
      yy=choose(u,z);
      w=1/(z+x+1);
      sal=sum(yy*w)*2^(-u); return(sal)}
    ff2=function(x,u){
      sal=mean(unlist(lapply(u,ff1,x))); return(sal)}

    suma3=mean(unlist(lapply(y,ff2,aux)))
    sumatotal=sum1+sum2-2*suma3
    return(sumatotal)
  }
```

Cálculo numérico Delta2

```r
fdelta2b=function(z){
  fdelta2b1 =function(y,t) {
    ((1/length(y)) * sum(t^y) -
       ((1/length(y)) * sum(((t+1)/2)^y))^2)^2
  }
  i =integrate(f = Vectorize(fdelta2b1,vectorize.args = 't'), lower = 0, upper = 1,
               y = z)
  i$value
}
```

Prueba Delta2

Comprobamos que los resultados mediante ambos métodos son los mismos.

```r
X = rpois(20,4)
(fdelta2(X))
```

```
## [1] 0.0001178003
```

```r
(fdelta2b(X))
```

```
## [1] 0.0001178003
```

8.2.1. Cálculo de eficiencia

En la imagen 8.3 veremos el tiempo computacional respecto al valor de n que se emplea para la ejecución de un test a partir del estadístico $\hat{\Delta}_1$ para valores de $B = 1000$, para $10 \leq n \leq 110$, y para $\lambda = 4$.

La línea verde representa el cálculo numérico y la roja la integración numérica.

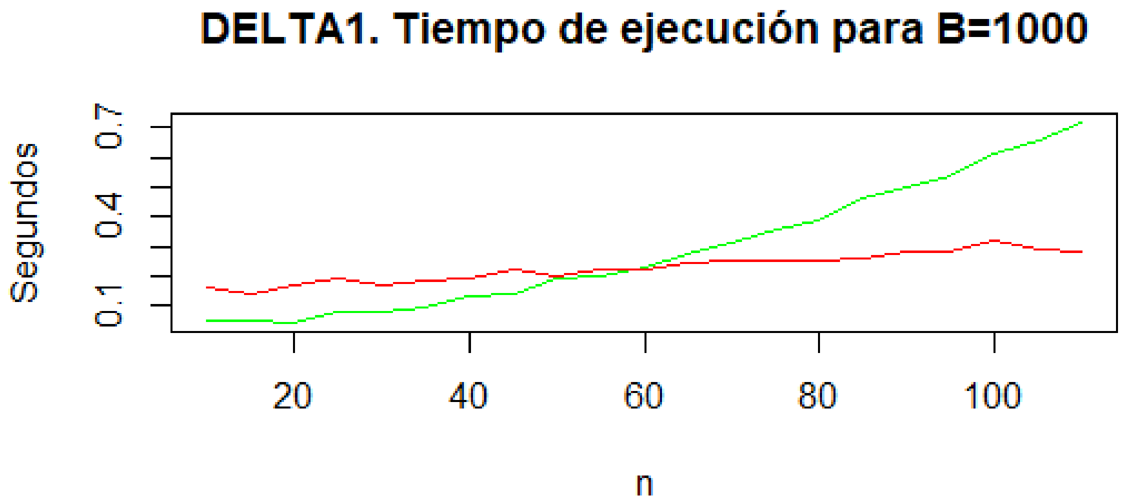


Figura 8.3

Conclusión: Podemos concluir que para valores mayores a $n \sim 60$ la integración numérica es más eficiente en términos de tiempo computacional que el cálculo numérico y para valores menores a $n \sim 60$, lo es el cálculo numérico.

En la siguiente imagen, 8.4, veremos el tiempo computacional respecto al valor de n que se emplea para la ejecución de un test a partir del estadístico $\hat{\Delta}_2^{\star}$ para valores de $B = 1000$, para $5 \leq n \leq 30$, y para $\lambda = 4$.

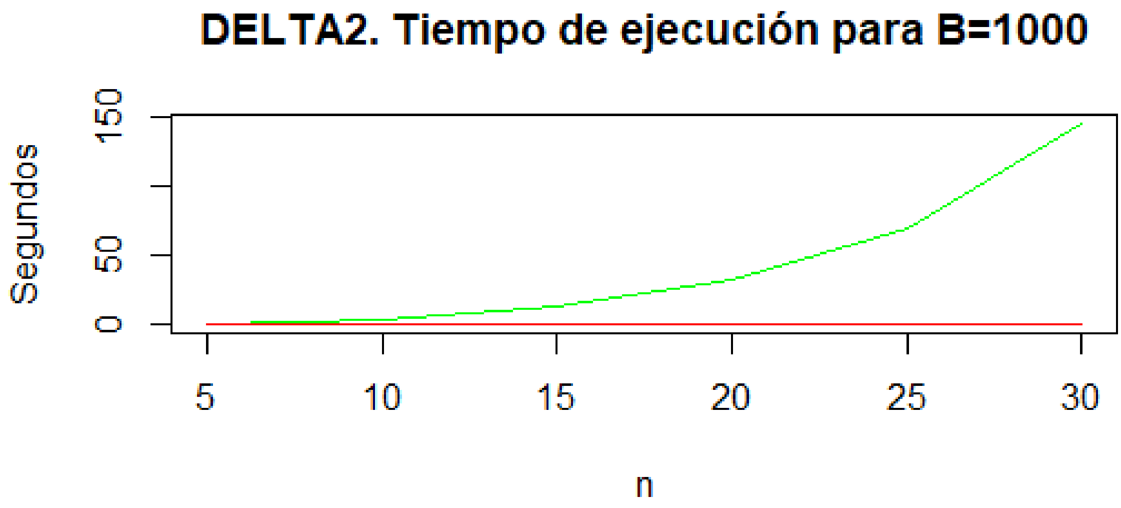


Figura 8.4

Conclusión: En este caso, para todos los valores de n vemos que es claramente más eficiente la integración numérica. Esto tiene una explicación bastante obvia, y es la necesidad de realizar una suma cuádruple y una suma triple a la hora de realizar el cálculo numérico, lo cual computacionalmente tiene un coste muy elevado. Aunque no sea aprecia bien, los valores de la línea roja (integración numérica) se encuentran entre 0.15 y 0.20, mientras que los valores de la línea verde (cálculo numérico) crecen exponencialmente desde un valor de 0.58 segundos para $n = 5$ hasta un valor de 145.74 segundos para $n = 30$.

8.3. Estudio computacional para el test de Rueda y otros

Para el cálculo de los test R_n y $R_{n,a}$ a la hora de realizar su programación computacional en el programa R hay, al menos, dos alternativas para cada uno. Estas son:

- Utilizando la función ***integrate***, integrada en el programa R, desarrollando la integración numérica de (3.9) y (3.10) respectivamente.

- Mediante cálculos solucionando las integrales en cuestión y programarla en el programa R. A este método lo llamaremos cálculo numérico.

Este es el desarrollo del cálculo numérico de $R_{n,a}$. Para R_n, es el mismo desarrollo pero para un valor de $a = 0$.

$$R_{n,a} = \frac{1}{n}\sum_{i,j=1}^{n}\frac{1}{X_i+X_j+a+1} - 2\sum_{i=1}^{n}\int_0^1 e^{\bar{X}_n(t-1)}t^{X_i+a}dt + n\int_0^1 e^{2\bar{X}_n(t-1)}t^a dt =$$
$$\frac{1}{n}\sum_{i,j=1}^{n}\frac{1}{X_i+X_j+a+1} - 2\sum_{i=1}^{n}\frac{e^{-\bar{X}_n}}{(-\bar{X}_n)^{X_i+a+1}}\gamma(X_i+a+1,-\bar{X}_n) + \frac{ne^{-2\bar{X}_n}}{(-2\bar{X}_n)^{a+1}}\gamma(a+1,-2\bar{X}_n) \tag{8.5}$$

donde $\gamma(c,x) = \int_0^x e^{-t}t^{c-1}dt$, es decir, la función gamma incompleta.

A continuación veremos en el programa R la programación del cálculo de los estadísticos R_n y $R_{n,a}$ mediante integración (3.9) y (3.10) y mediante cálculo numérico (8.5) para $R_{n,a}$, y con $a = 0$ para R_n. También determinaremos que mediante ambos métodos el estadístico ha sido bien calculado comprobando que para una misma muestra devuelve el mismo valor.

Cálculo numérico e integración numérica para Test de Rueda et al

```r
#install.packages("pracma")
library(pracma)
```

```
## Warning: package 'pracma' was built under R version 4.0.4
```

Cálculo numérico, Rn

```r
fRn=function(z){
    aux = outer(z,z, "+") + 1
    t = sum(1/aux)
    coc1 = t/length(z)
    r = 0
    for (i in 1:length(z)) {
      r = r + ((exp(-mean(z))/((-mean(z))^(z[i] + 1))) *
                 as.numeric(gammainc(a = (z[i] + 1), x = (-mean(z)))[1]))
    }
    coc2 = 2*r
    num3 = length(z) * exp(-2*mean(z))
    den3 = (-2 * mean(z))
    coc3 = (num3/den3)* as.numeric(gammainc(a = 1, x = (-2*mean(z)))[1])
    return(coc1 - coc2 + coc3)
  }
```

Integración numérica Rn

```r
fRn1 = function(z){
    Gn = function(t,y){
      (sqrt(length(y))*((1/length(y)) * sum(t^y) -
                          exp(mean(y)*(t-1))))^2
    }
    i = integrate(f = Vectorize(Gn,vectorize.args = 't'), lower = 0, upper = 1,
                  y = z)
    i$value
  }
```

Prueba Rn

Comprobamos que los resultados mediante ambos métodos son los mismos.

```r
X = rpois(20,4)
(fRn(X))
```

```
## [1] 0.00104409
```

```r
(fRn1(X))
```

```
## [1] 0.00104409
```

Integración numérica Rn,a

```
fRn_a = function(z,c){
    Gna = function(t,y,b){
      (sqrt(length(y))*((1/length(y)) * sum(t^y) -
                          exp(mean(y)*(t-1))))^2 * t^b
    }
    i = integrate(f = Vectorize(Gna,vectorize.args = 't'), lower = 0, upper = 1,
                 y = z, b = c)
    i$value
  }
```

Cálculo numérico Rn,a

```
fRn_a1=function(z,b){
    aux = outer(z,z, "+") + b + 1
    t = sum(1/aux)
    coc1 = t/length(z)
    r = 0
    for (i in 1:length(z)) {
      r = r + ((exp(-mean(z))/((-mean(z))^(z[i] + b + 1))) *
                as.numeric(gammainc(a = (z[i] + b + 1), x = (-mean(z)))[1]))
    }
    coc2 = 2*r
    num3 = length(z) * exp(-2*mean(z))
    den3 = (-2 * mean(z))^(b + 1)
    coc3 = (num3/den3)* as.numeric(gammainc(a = (b + 1), x = (-2*mean(z)))[1])
    coc1 - coc2 + coc3
  }
```

Prueba Rn,a

Comprobamos que los resultados mediante ambos métodos son los mismos.

```
X = rpois(20,4)
(fRn_a(X,3))
```

```
## [1] 0.0009752881
```

```
(fRn_a1(X,3))
```

```
## [1] 0.0009752881
```

8.3.1. Cálculo de eficiencia

En la imagen 8.5 veremos el tiempo computacional respecto al valor de n que se emplea para la ejecución de un test a partir del estadístico R_n para valores de $B = 1000$, para $10 \leq n \leq 200$ y para $\lambda = 4$. La línea verde representa el cálculo numérico y la roja la integración.

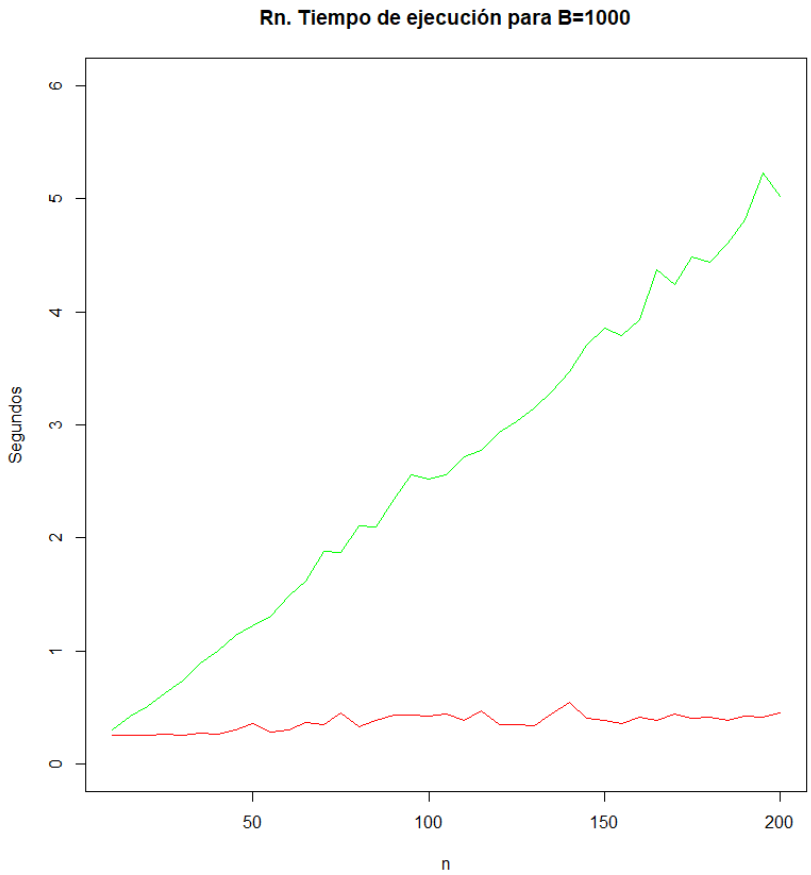


Figura 8.5

Conclusión: Podemos concluir que para todo valor de $n > 10$ es mayor el tiempo empleado para obtener el valor del estadístico mediante el cálculo numérico que mediante la integración. Mientras que el tiempo que se emplea con el cálculo numérico crece de forma muy pronunciada a medida que crece el valor de n, el tiempo que se emplea con la integración no varía independientemente de cuál sea el valor de n.

En la imagen 8.6, vamos a ver el gráfico anterior pero para valores de n entre 3 y 20 para así analizar este tramo concreto con profundidad.

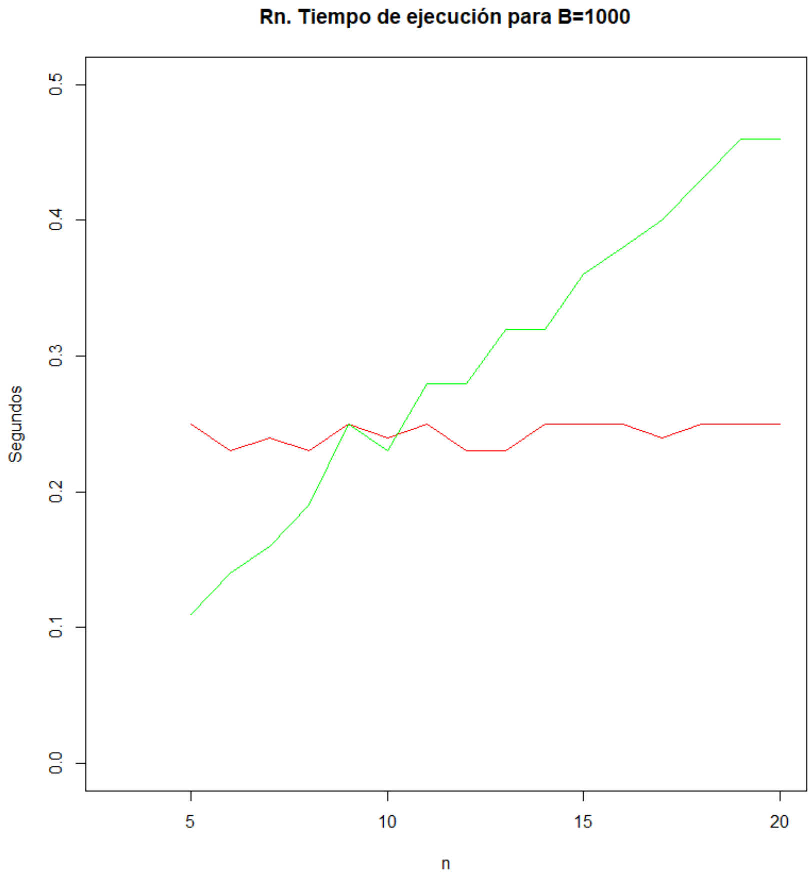


Figura 8.6

A partir de esta imagen podemos determinar que el tiempo empleado en calcular el estadístico mediante cálculo numérico es inferior al empleado para calcularlo mediante integración para valores inferiores a $n \sim 10$.

En la imagen 8.7 veremos el tiempo computacional respecto al valor de n que se emplea para la ejecución de un test a partir del estadístico $R_{n,3}$ para valores de $B = 1000$, para $10 \leq n \leq 100$ y para $\lambda = 4$. La línea verde representa el cálculo numérico y la roja la integración.

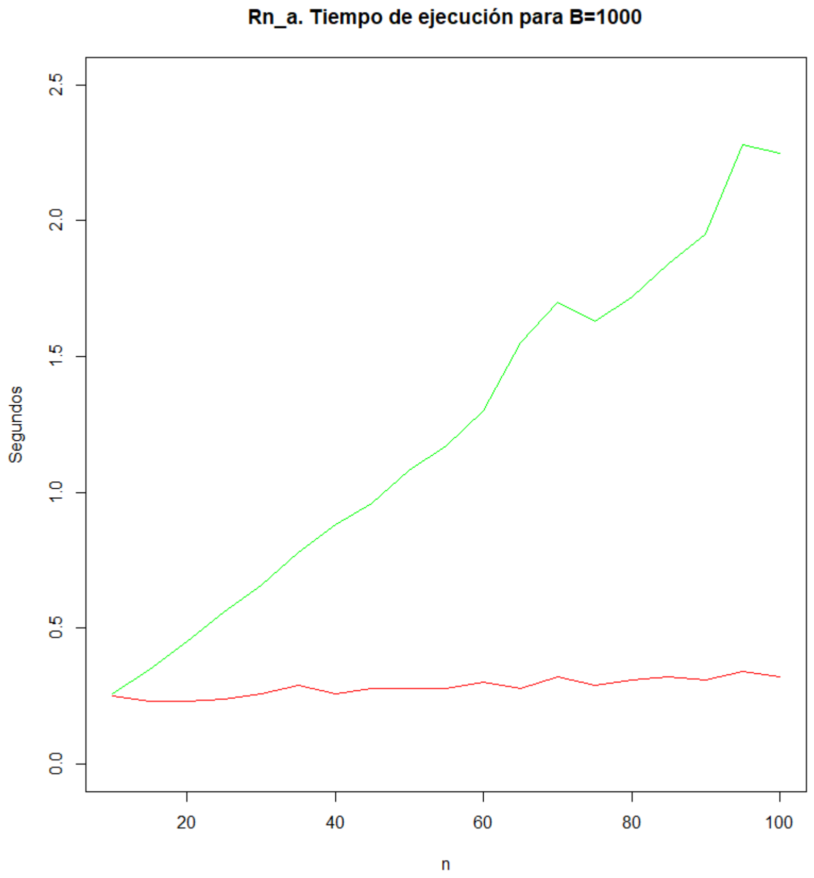


Figura 8.7

Conclusión: De forma semejante a como ocurría con R_n, el tiempo empleado es mucho mayor, al menos, a partir de valores de $n > 10$ aplicando cálculo numérico que aplicando integración. Además, también vemos que el crecimiento es bastante pronunciando cuando al cálculo numérico se refiere a medida que se incrementa el valor de n, mientras que en la integración el incremento del valor de n no influye.

En la imagen 8.8, veremos el gráfico anterior pero para valores de n entre 3 y 20 para analizar ese tramo concreto con profundidad.

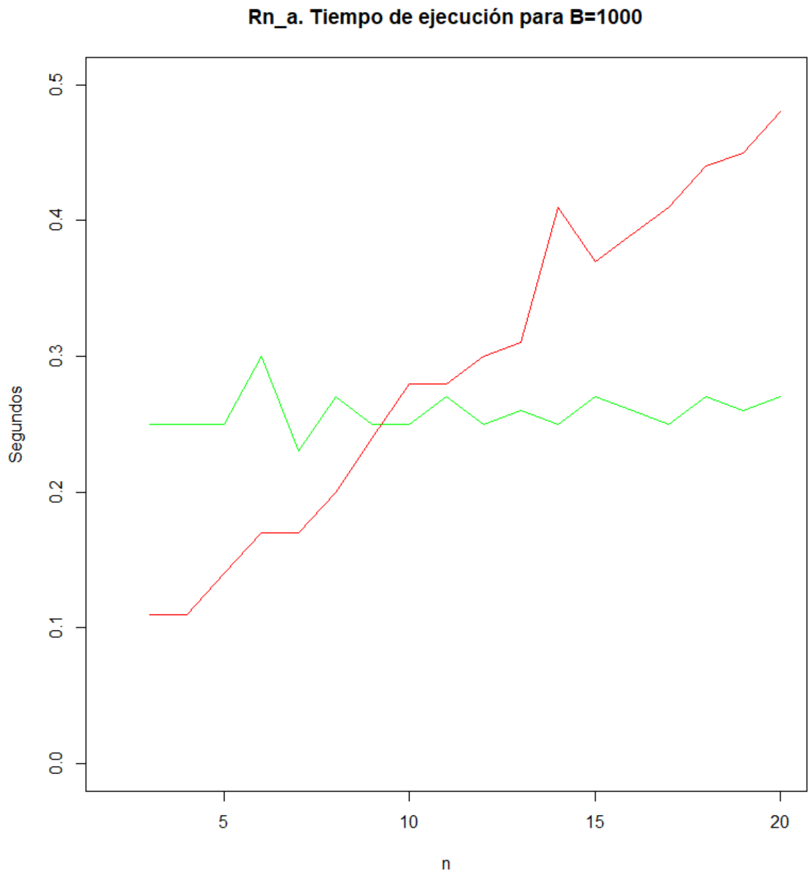


Figura 8.8

Nuevamente, de forma semejante a lo ocurrido con R_n, podemos determinar a partir de este gráfico que el tiempo empleado utilizando el cálculo numérico para calcular el estadístico es menor que el empleado utilizando la integración para valores de $n \sim 10$.

Capítulo 9

Aplicación de los test a datos reales

9.1. Introducción

En este capítulo vamos a realizar una aplicación de lo visto en los capítulos anteriores a un conjunto de datos reales. En este caso utilizaremos los conjunto de datos referidos al número de goles en cada partido de los mundiales de fútbol desde 1990 hasta la actualidad (el último, en 2018). Cada uno de estos trece conjuntos de datos de mundiales (uno cada cuatro años), supondremos que siguen una distribución Poisson con un parámetro λ determinado por la media de los conjuntos de datos.

Aplicaremos varios de los test vistos en los capítulos anteriores para determinar la Poissonidad de estos datos y también compararemos gráficamente tanto las frecuencias relativas como las frecuencias relativas acumuladas de estos conjuntos de datos con los de una distribución Poisson con un parámetro λ determinado por la media de los conjuntos de datos.

9.2. Relación de los goles en el fútbol con la distribución Poisson

Mahrer (1982) [12] proponen un modelo de predicción de resultados de fútbol basados asumiendo las siguientes dos suposiciones:

- El número de goles anotados por un equipo en un partido de fútbol se distribuye mediante una distribución Poisson.
- El número de goles anotados por un equipo es independiente al número de goles anotados por el equipo rival.

Para un encuentro entre un equipo local (indexado por i) y un equipo visitante (indexado por j), proponen modelar los goles del equipo local por la siguiente distribución:

$$X_{i,j} \sim Pois(\alpha_i \beta_j)\tau,$$

siendo α_i un parámetro que denota un ratio de ataque del equipo local, β_j un parámetro que denota un ratio de defensa para el equipo visitante y τ un parámetro que denota la ventaja de jugar en casa para el equipo local. Por otro lado, para modelar los goles del equipo visitante proponen la siguiente distribución:

$$Y_{i,j} \sim Pois(\alpha_j \beta_i), \tag{9.1}$$

donde cada parámetro representa lo mismo que en la distribución anterior pero para el equipo opuesto.

Para los marcadores bajos (0-0, 1-1, 1-0, 0-1) el modelo (9.1) presentaba algunos fallos, debido a que el supuesto de independencia no se cumplía. Por ello, Dixon y Coles (1997) [6] proponen la siguiente modificación al modelo:

$$P(X_{i,j} = x, Y_{i,j} = y) = \kappa_{\lambda,\mu}(x,y)\frac{\lambda^x e^{-\lambda}}{x!} + \frac{\mu^y e^{-\mu}}{y!}, \tag{9.2}$$

donde

$$\lambda = \alpha_i \beta_j \tau,$$
$$\mu = \alpha_j \beta_i$$

y

$$\kappa_{\lambda,\mu}(x,y) = \begin{cases} 1 - \lambda\mu\rho & \text{si } x = y = 0 \\ 1 + \lambda\rho & \text{si } x = 0, y = 1 \\ 1 + \mu\rho & \text{si } x = 1, y = 0 \\ 1 - \rho & \text{si } x = y = 1 \\ 1 & \text{otro caso} \end{cases}$$

siendo $\text{máx}(-1/\lambda, -1/\mu) \leq \rho \leq \text{mín}(1/\lambda * \mu, 1)$.

ρ representa un parámetro de dependencia, donde $\rho = 0$ corresponde a la independencia total, la cual se ve perturbada para valores de $x \leq 1$ y $y \leq 1$.

Wang (2010) [22] trata de predecir los goles y resultados finales que se darían en el Mundial de Sudáfrica de 2010, modelando los goles de los diferentes equipos mediante de una distribución Poisson. Esta distribución Poisson, para un partido entre dos equipos indexados por i y j, tiene un parámetro

$$\lambda_{i,j} = [\frac{R_i}{R_j}]^\alpha,$$

donde R_k se corresponde con con el rating que asociaba la FIFA en el mes previo al mundial al equipo k y α es un parámetro estimado a partir de los enfrentamientos históricos entre ambos equipos.

9.3. Análisis gráfico y testeo

Una vez hemos visto algo de historia sobre el uso de la distribución Poisson para modelar los goles en el fútbol, vamos a corroborar tanto de forma gráfica como de forma numérica (aplicando los test) el principal supuesto en que se basan estos artículos (y que no se demuestra en dichos) artículos, esto es, que el número de goles anotados por un equipo en un partido de fútbol se distribuye mediante una distribución Poisson.

Teorema 19. *Sean ξ_1 y ξ_2 dos variables aleatorias independientes que siguen una distribución Poisson de parámetro λ_1 y λ_2 respectivamente, entonces, la variable aleatoria $\xi = \xi_1 + \xi_2$ sigue una distribución Poisson de parámetro $\lambda = \lambda_1 + \lambda_2$.*

Por el Teorema 19, si se cumple el supuesto de que el número de goles anotados por un equipo en un partido de fútbol se distribuye mediante una distribución Poisson, también ha de cumplirse el supuesto de que el número de goles anotados en un partido de fútbol se distribuye mediante una distribución Poisson.

9.3.1. Datos

Los datos que hemos elegido para comprobar el supuesto han sido los goles anotados en los partidos de la Copa del Mundo de 2018, en los de la Copa del Mundo de 2014, en los de la Copa del Mundo de 2010 y en la de los goles anotados en todas las copas del mundo desde la Copa del Mundo de 1990 a la Copa del Mundo de 2018. Las frecuencias absolutas son las siguientes:

Cuadro 9.1: Frecuencias de goles por partido en el Mundial 2018

Núm. goles	0	1	2	3	4	5	6	7
Frecuencia	1	15	17	19	5	2	2	3

Cuadro 9.2: Frecuencias de goles por partido en el Mundial 2014

Núm. goles	0	1	2	3	4	5	6	7	8
Frecuencia	7	12	8	20	9	4	2	1	1

Cuadro 9.3: Frecuencias de goles por partido en el Mundial 2010

Núm. goles	0	1	2	3	4	5	6	7
Frecuencia	7	17	13	14	7	5	0	1

Cuadro 9.4: Frecuencias de goles por partido en los Mundiales entre 1990 y 2018

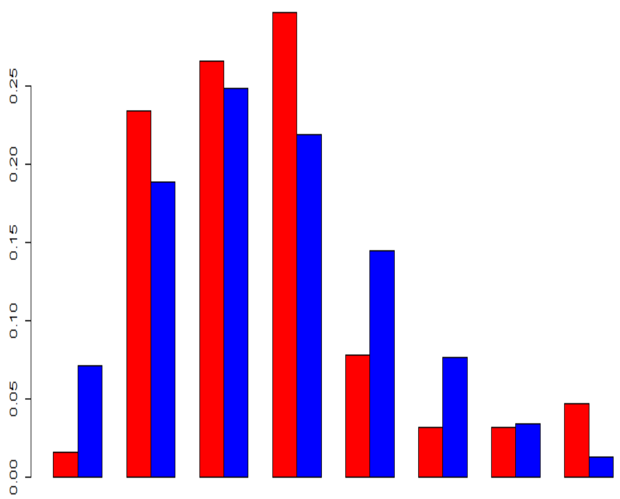
Núm. goles	0	1	2	3	4	5	6	7	8
Frecuencia	25	72	76	76	39	17	7	6	2

9.3.2. Mundial 2018

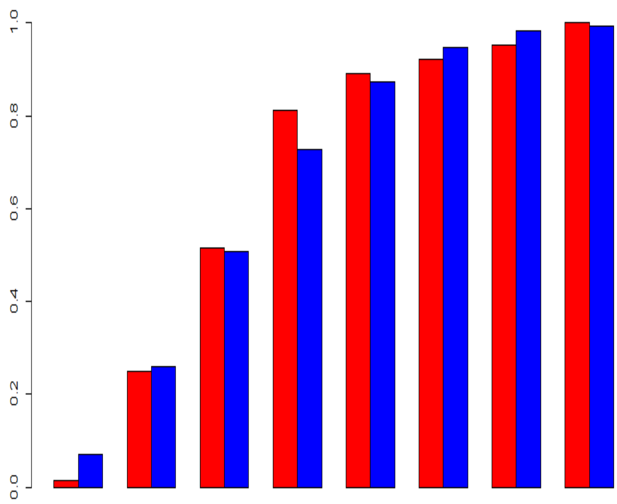
La frecuencia de distribución de los goles por partido en este mundial la podemos ver en 9.1.

Cuadro 9.5: p-valores de algunos de los test estudiados para los datos del Mundial 2018, basados en 10000 muestras bootstrap

T_n	$T_{n,1}$	$T_{n,3}$	$T_{n,5}$	R_n	$R_{n,1}$	$R_{n,3}$	$R_{n,5}$	u
0.0606	0.0981	0.1721	0.2379	0.0836	0.1206	0.1934	0.2614	0.6107
$\hat{\Delta}_1$	$\hat{\Delta}_{1,1}$	$\hat{\Delta}_{1,3}$	$\hat{\Delta}_{1,5}$	$\hat{\Delta}_2^\star$	$\hat{\Delta}_{2,1}^\star$	$\hat{\Delta}_{2,3}^\star$	$\hat{\Delta}_{2,5}^\star$	$\hat{\Delta}_\infty$
0.9758	0.9593	0.9163	0.8683	0.0379	0.0532	0.0966	0.1462	0.0417
K_n	C_n	C_n*	L_n					
0.0313	0.0394	0.665	0.0772					



(a) Frecuencias empíricas relativas de una $Pois(2.64)$ (azul) y de los datos del Mundial 2018 (rojo)



(b) Frecuencias empíricas relativas acumuladas de una $Pois(2.64)$ (azul) y de los datos del Mundial 2018 (rojo)

Figura 9.1

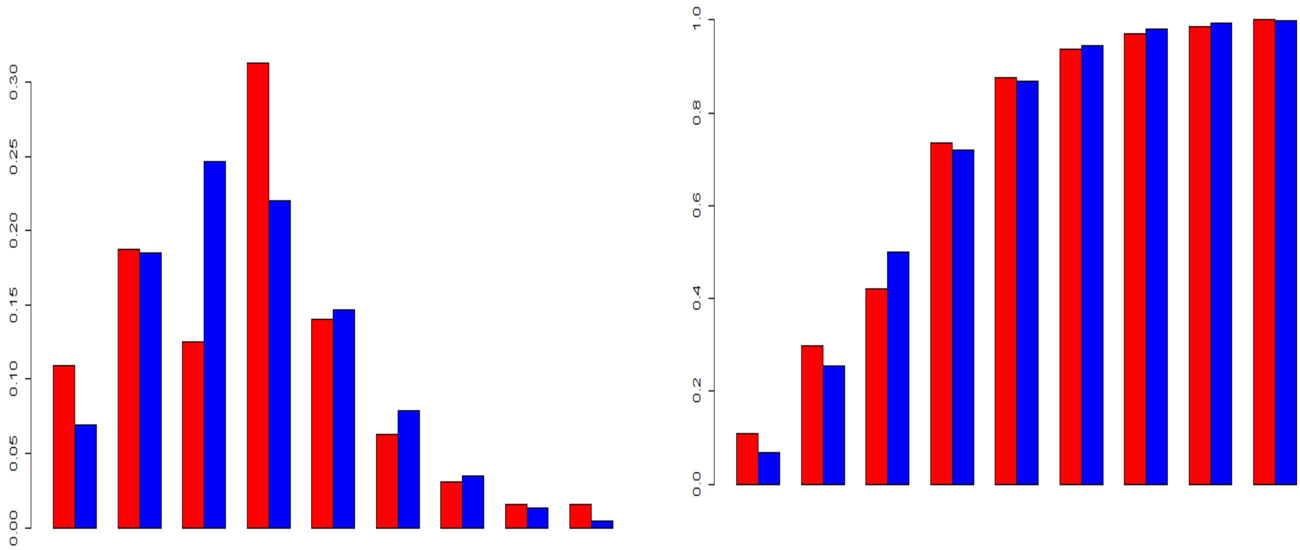
Para un nivel del 5 % rechazan la hipótesis nula los estadísticos $\hat{\Delta}_2^\star$, $\hat{\Delta}_\infty$, K_n y C_n y para un nivel del 10 % se rechaza además para T_n, $T_{n,1}$, R_n, $\hat{\Delta}_{2,1}^\star$ y $\hat{\Delta}_{2,3}^\star$. Los p-valores más altos corresponden a $\hat{\Delta}_1$ y $\hat{\Delta}_{1,1}$.

Vemos que las frecuencias empíricas más altas son mayores que las que se corresponderían a una distribución $P(2.64)$, mientras que las más bajas, menores. Las frecuencias empíricas acumuladas se ajustan bastante mejor a una distribución $Pois(2.64)$.

9.3.3. Mundial 2014

Cuadro 9.6: p-valores de algunos de los test estudiados para los datos del Mundial 2014, basados en 10000 muestras bootstrap

T_n	$T_{n,1}$	$T_{n,3}$	$T_{n,5}$	R_n	$R_{n,1}$	$R_{n,3}$	$R_{n,5}$	u
0.1617	0.1637	0.222	0.2575	0.1596	0.1816	0.221	0.2537	0.1707
$\hat{\Delta}_1$	$\hat{\Delta}_{1,1}$	$\hat{\Delta}_{1,3}$	$\hat{\Delta}_{1,5}$	$\hat{\Delta}^{\star}_2$	$\hat{\Delta}^{\star}_{2,1}$	$\hat{\Delta}^{\star}_{2,3}$	$\hat{\Delta}^{\star}_{2,5}$	$\hat{\Delta}_\infty$
0.0659	0.0749	0.0912	0.1062	0.1395	0.1448	0.1772	0.2037	0.1832
K_n	C_n	C_n*	L_n					
0.443	0.745	0.2109	0.0267					



(a) Frecuencias empíricas relativas de una $Pois(2.67)$ (azul) y de los datos del Mundial 2014 (rojo)

(b) Frecuencias empíricas relativas acumuladas de una $Pois(2.67)$ (azul) y de los datos del Mundial 2014 (rojo)

Figura 9.2

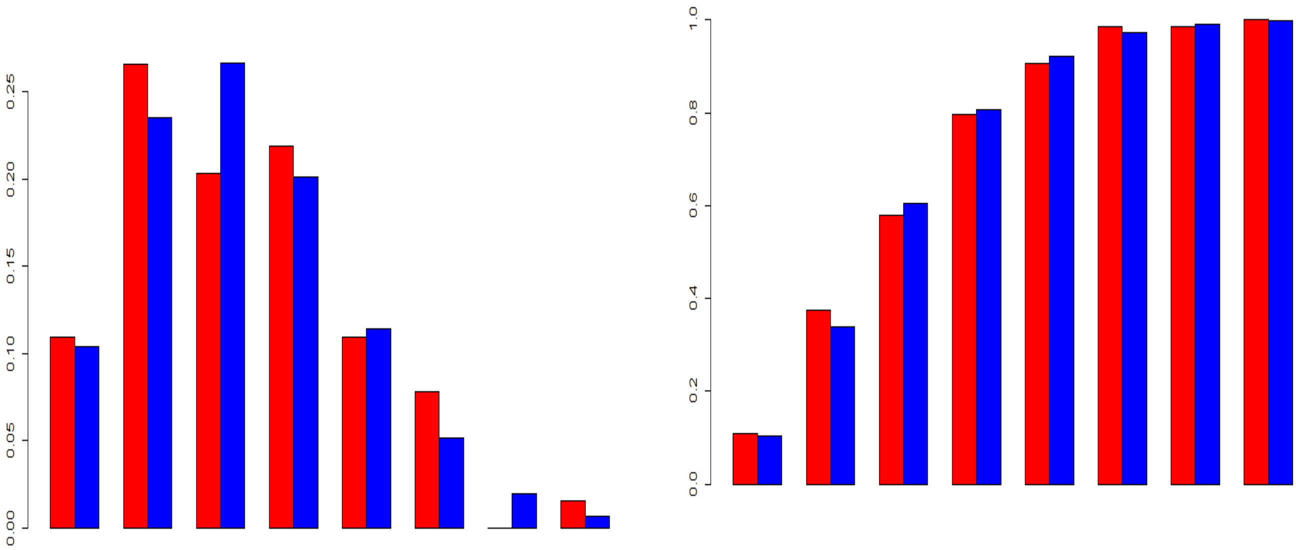
Para un nivel del 5 % el único estadístico que rechaza la hipótesis nula es L_n. Para un nivel del 10 % la rechazan, además, $\hat{\Delta}_1$, $\hat{\Delta}_{1,1}$ y $\hat{\Delta}_{1,3}$. Los p-valores más altos corresponden a $T_{n,5}$ y $R_{n,5}$.

Las frecuencias empíricas se ajustan a las de una distribución $Pois(2.67)$, exceptuando el número de partidos con 2 goles que es bastante menor al que se cabría esperar y el número de partidos con 3 goles, que es bastante mayor. Las frecuencias empíricas acumuladas se ajustan bastante bien a las una distribución $Pois(2.67)$ a pesar de estos desajustes en las frecuencias empíricas sin diferencias notables.

9.3.4. Mundial 2010

Cuadro 9.7: p-valores de algunos de los test estudiados para los datos del Mundial 2010, basados en 10000 muestras bootstrap

T_n	$T_{n,1}$	$T_{n,3}$	$T_{n,5}$	R_n	$R_{n,1}$	$R_{n,3}$	$R_{n,5}$	u
0.8743	0.7968	0.7345	0.7156	0.776	0.7331	0.7006	0.7055	0.3237
$\hat{\Delta}_1$	$\hat{\Delta}_{1,1}$	$\hat{\Delta}_{1,3}$	$\hat{\Delta}_{1,5}$	$\hat{\Delta}_2^\star$	$\hat{\Delta}_{2,1}^\star$	$\hat{\Delta}_{2,3}^\star$	$\hat{\Delta}_{2,5}^\star$	$\hat{\Delta}_\infty$
0.3504	0.3132	0.2947	0.2999	0.848	0.785	0.7159	0.7014	0.823
K_n	C_n	C_n*	L_n					
0.5568	0.5034	0.2922	0.5209					



(a) Frecuencias empíricas relativas de una $Pois(2.265)$ (azul) y de los datos del Mundial 2010 (rojo)

(b) Frecuencias empíricas relativas acumuladas de una $Pois(2.265)$ (azul) y de los datos del Mundial 2010 (rojo)

Figura 9.3

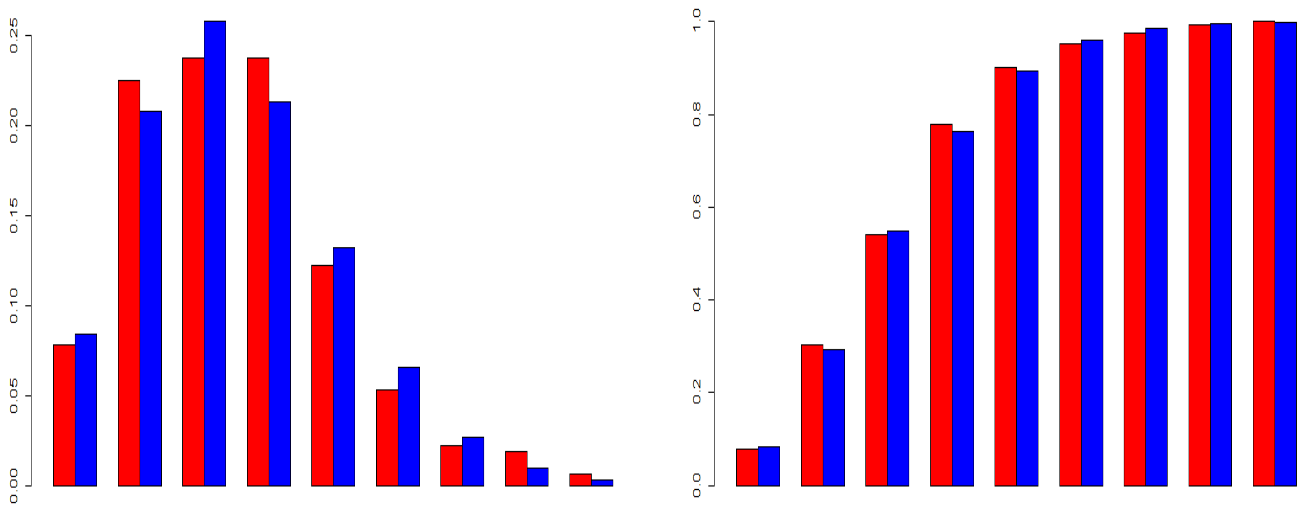
Ni para un nivel del 5 % ni del 10 % hay estadístico que rechace la hipótesis nula. Los p-valores más altos corresponden a T_n y $\hat{\Delta}_2^\star$.

Las frecuencias empíricas relativas se ajustan bien a las de una distribución $P(2.265)$, exceptuando el número de partidos con un gol, que es algo superior a lo esperado, y el número de partidos con dos goles, que es menor a lo esperado. Nuevamente, vemos que las frecuencias empíricas relativas acumuladas se ajustan bastante bien a una $Pois(2.265)$.

9.3.5. Mundiales entre 1990 y 2018

Cuadro 9.8: p-valores de algunos de los test estudiados para los datos de los Mundiales entre 1990
y 2018, basados en 10000 muestras bootstrap

T_n	$T_{n,1}$	$T_{n,3}$	$T_{n,5}$	R_n	$R_{n,1}$	$R_{n,3}$	$R_{n,5}$	u
0.79	0.8871	0.9363	0.9184	0.8608	0.923	0.9468	0.9303	0.3147
$\hat{\Delta}_1$	$\hat{\Delta}_{1,1}$	$\hat{\Delta}_{1,3}$	$\hat{\Delta}_{1,5}$	$\hat{\Delta}_2^{\star}$	$\hat{\Delta}_{2,1}^{\star}$	$\hat{\Delta}_{2,3}^{\star}$	$\hat{\Delta}_{2,5}^{\star}$	$\hat{\Delta}_{\infty}$
0.6136	0.5681	0.5039	0.4583	0.7179	0.8041	0.8906	0.9307	0.5762
K_n	C_n	C_n*	L_n					
0.5996	0.1936	0.5502	0.5212					



(a) Frecuencias empíricas relativas de una $Pois(2.478)$ (azul) y de los datos de los Mundiales entre 1990 y 2018 (rojo)

(b) Frecuencias empíricas relativas acumuladas de una $Pois(2.478)$ (azul) y de los datos de los Mundiales entre 1990 y 2018 (rojo)

Figura 9.4

Ni para un nivel del 5 % ni del 10 % hay estadístico que rechace la hipótesis nula. Los p-valores más altos corresponden a T_n y $\hat{\Delta}_2^{\star}$.

En este caso, las frecuencias empíricas relativas se ajustan bastante bien a las de una distribución $Pois(2.478)$ sin diferencias notables. Nuevamente, las frecuencias empíricas relativas acumuladas se ajustan bastante bien a las de una distribución $Pois(2.478)$.

Capítulo 10

Paquete en R Studio

10.1. Introducción

Ni en R ni en Python existe ningún paquete concreto ni ninguna librería que realicen test para determinar la Poissonidad de una muestra. No ocurre lo mismo, por ejemplo, con los test de normalidad, ya que tanto en R como en Python existe al menos una librería enfocada exclusivamente en ello. Por ejemplo, en R está el paquete *nortest* dedicada exclusivamente a los test de normalidad o algunas funciones 'sueltas' incluidas en otras librerías como *shapiro.test* incluida en el paquete *stats* que aplica el test de Shapiro-Wilk para las distribuciones normales. En Python, dentro de la librería *scipy.stats* se encuentran varias funciones para test de normalidad como *scipy.stats.normaltest*, *scipy.stats.shapiro* o *scipy.stats.kstest*.

No es esto lo que ocurre en los test de Poissonidad, tan solo existe en R la función *poisson.mtest* dentro del paquete *energy* que aplica un test de Poissonidad, concretamente, el test de Székely y Rizzo que hemos visto en (3.8), pero ningún paquete enfocado a estos test. En Python, no hay ni siquiera una función dedicada a los test de Poissonidad.

Es por esto que hemos decidido llevar a cabo la creación y el desarrollo de un paquete propio en R en el que aparezcan la mayoría de los test vistos en este trabajo (exceptuando, lógicamente, el test de Székely y Rizzo que ya está) y su aplicación.

En este capítulo va a explicarse en primer lugar el modelo seguido a la hora de programar los test y unos detalles a tener en cuenta sobre algunos test y, finalmente, un anexo en el que veremos la programación de todos los test.

10.2. Paquete TestPoissonity

El paquete creado tiene el nombre de **TestPoissonity**, se encuentra disponible para descargar en el repositorio github.com/MMH1997/TestPoissonity. El paquete consiste en el desarrollo de catorce test vistos en este trabajo cuya programación será descrita en el 10.3. A continuación vamos a ver en un R Markdown una descripción del modelo seguido para la programación de los test. Como todos tienen una estructura semejante, para la descripción nos centraremos exclusivamente en un test, K_n, visto en (3.6).

```r
#' Kn test

testKn = function(x,n.boot){ #Iniciamos la función.
  options(warn = -1) #Se eliminan posibles warnings
  Kn=function(z) {if (sum(z) == 0) {0} #Se crea el estadístico en cuestión, en este caso, de Kn.
    else {ecdfz=ecdf(z)
    f1=function(y) {sqrt(length(z)) * (ecdfz(y) - ppois(y, mean(z)))}
    resf11=f1(min(z):max(z))
    Kn=max(resf11)
    return(Kn)
    }
  } #Aquí acaba la creación del estadístico.
  pv = 0 #Se asigna un valor inicial previo al bucle de 0 al pvalor.
  n = length(x) #Se nombra como n al tamaño de la muestra
  lambda.hat=mean(x) #Se nombra como lambda.hat a la media de la muestra
  T.obs=Kn(x) #Se nombra como T.obs al valor del estadístico aplicado a la muestra.
  for (i in 1:n.boot){ #Se inicializa un bucle de n.boot repeticiones.
    x.boot=rpois(n,lambda.hat) #Se genera una distribución Poisson con mismo tamaño
#y media que la muestra original.
    T.boot=Kn(x.boot) #Se aplica la función del estadístico a la distribución Poisson generada.
    pv=pv+as.numeric(T.boot>T.obs) #Si la función aplicada al estadístico en la distribución Poisson
#generada es mayor que la función aplicada al estadístico en la #muestra original se suma 1 a pv
#(que se va actualizando en cada iteración).
  }
  #Finaliza el bucle.
  names(T.obs) <- "test statistic" #Se nombra al estadístico de la muestra original
  media = mean(x)
  names(media) <- "mean" #Se nombra a la media de la muestra original.
  e = #Se crea una lista que será lo que devuelva el test.
    list(method = paste("Kn poissonity test", #El nombre del test
                        sep = ""),
         statistic = T.obs #El estadístico de la muestra original.
         , p.value = pv/n.boot, #El pvalor, el cual se obtiene dividiendo el número de veces que
#la función del estadístico aplicada a las distintas distribuciones Poisson generadas en el bucle
#es mayor que la función del #estadístico aplicada a la muestra original entre el número total
#de iteraciones realizadas (n.boot).
         data.name = paste("sample size ", n, ", replicates ",
                           n.boot, sep = ""), #El tamaño de la muestra y el número de replicaciones.
         estimate = media) #El valor de la media.
         class(e) <- "htest" #Se transforma la lista en una lista de clase "htest".
         e #Se devuelve la lista.
}
```

A continuación, vamos a ver la aplicación de este test para una muestra obtenida a partir de una distribución Poisson de parámetro lambda = 5 y de tamaño n = 20, con 2000 replicaciones.

```r
Muestra = rpois(20,5)
testKn(Muestra,2000)
```

```
##
##  Kn poissonity test
##
## data:  sample size 20, replicates 2000
## test statistic = 0.53257, p-value = 0.1595
## sample estimates:
```

```
## mean
##  5.2
```

En la salida de la función vemos el nombre del test, la descripción de los datos (tamaño de muestra y número de replicaciones), el valor del estadístico, el valor del p-valor y la media de la muestra.

En el anexo que veremos en la siguiente sección está la programación de todas las funciones incluidas en el paquete, pero antes, es necesario hacer algunas valoraciones sobre cómo se han implementado algunas funciones.

- Como vimos en el Capítulo 8, para algunas funciones existen, al menos, dos maneras de programarlas, dependiendo la eficiencia del tamaño muestral. Esto se ha tenido en cuenta a la hora de llevar a cabo la programación de los estadísticos $\hat{\Delta}_1$, T_n y $T_{n,a}$.
- En los estadísticos R_n y $R_{n,a}$ vimos en 8.6 y 8.8 que para tamaños muestrales menores que 10, existía una forma de programación más eficiente que para tamaños muestrales mayores que 10. Sin embargo, debido a que muestras menores a 10 en la práctica no son muy comunes y que la diferencia de eficiencia entre los dos métodos es minúscula y prácticamente inapreciable, se ha decidido mantener la misma forma de programación para todos los tamaños muestrales, ya sean mayores o menores que 10, esto es, el método de integración numérica (3.9) y (3.10).
- Para el test u, vimos en 3.2 que existen dos métodos para calcular sus puntos críticos, o bien mediante una aproximación a la normal, o bien, como hemos hecho con el resto de test, mediante el método de aproximación boostrap paramétrico. En este caso, para llevar a cabo el método de aproximación a la normal, hemos incluido la opción de emplear ambos métodos para el cálculo del p-valor. Caso de asignar al parámetro 'n.boot' el valor 0, se aplicará el método de aproximaci'on a la normal, mientras que si se asigna otros número entero positivo, se aplicará el método de aproximación boostrap paramétrico con dicho número de repeticiones.

10.3. Anexo: Test incluidos en el paquete TestPoissonity

Test basados en la función de distribución empirica

```r
#' Kn test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable

testKn = function(x,n.boot){
  options(warn = -1)
  Kn=function(z) {if (sum(z) == 0) {0}
    else {ecdfz=ecdf(z)
    f1=function(y) {sqrt(length(z)) * (ecdfz(y) - ppois(y, mean(z)))}
    resf11=f1(min(z):max(z))
    Kn=max(resf11)
    return(Kn)
    }
  }
  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=Kn(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=Kn(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list("method" = paste("Kn poissonity test",
                            sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e
}
```

```r
#' Cn test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable

testCn = function(x,n.boot){
  Cn=function(z) {if (sum(z) == 0) {0}
    else {ecdfz=ecdf(z)
    f1=function(y) {(ecdfz(y) - ppois(y, mean(z)))^2*
        dpois(y,mean(z))}
    resf11=sum(f1(min(z):max(z)))
    Cn=length(z) * resf11
    return(Cn)
    }
  }
  pv = 0
```

```r
  n = length(x)
  lambda.hat=mean(x)
  T.obs=Cn(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=Cn(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Cn poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  return(e)

}
```

```r
#' Cn* test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable

testCn_ = function(x,n.boot){
  options(warn = -1)
    Cn_=function(z) {if (sum(z) == 0) {0}
    else {ecdfz=ecdf(z)
    f1=function(y) {(ecdfz(y) - ppois(y, mean(z)))^2*
    (sum(z == y)/length(z))}
    resf11=sum(f1(min(z):max(z)))
    Cn_=length(z) * resf11
    }
  }
  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=Cn_(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=Cn_(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Cn* poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
```

```
  return(e)

}
```

```
#' Ln test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable

testLn = function(x,n.boot){
  options(warn = -1)
  Ln=function(z) {if (sum(z) == 0) {0}
    else {ecdfz=ecdf(z)
    f1=function(y) {sqrt(z) * (ecdfz(y) - ppois(y, mean(z)))}
    resf11=sum(f1(min(z):max(z)))
    Ln = return(resf11)
    }
  }
  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=Ln(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=Ln(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Ln poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  return(e)

}
```

Test basados en la función generatriz de probabilidad

```
#' Delta1 test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable


testdelta1 = function(x,n.boot){
testdelta1b = function(x,n.boot){
    fdelta1b=function(z) {
      aux=outer(z,z,"+")
      num= 2-2^(-aux)
```

```
    den=aux+1
    coc=num/den
    mean(coc)
    Delta1Clas=mean(1/(z+1))-mean(coc)
    Delta1Clas
  }
  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=fdelta1b(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=fdelta1b(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Delta 1 poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e
 }
testdelta1c = function(x,n.boot){
  fdelta1c=function(z){
    fdelta1c1 =function(y,t) {
      (1/length(y)) * sum(t^y) -
        ((1/length(y)) * sum(((t+1)/2)^y))^2
    }
    i =integrate(f = Vectorize(fdelta1c1,vectorize.args = 't'), lower = 0, upper = 1,
                 y = z)
    i$value
  }
  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=fdelta1c(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=fdelta1c(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Delta 1 poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
```

```r
    class(e) <- "htest"
    e
  }
if (length(x) > 60) {
  return(testdelta1b(x,n.boot))
}
else {return(testdelta1c(x,n.boot))}
}
```

```r
#' Delta1,a test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable

testdelta1_a = function(x,a,n.boot){
  fdelta1_a=function(z, c){
    fdelta1b1 =function(y,t, b) {
      ((1/length(y)) * sum(t^y) -
        ((1/length(y)) * sum(((t+1)/2)^y))^2) * t^b
    }
    i =integrate(f = Vectorize(fdelta1b1,vectorize.args = 't'), lower = 0, upper = 1,
                 y = z, b = c)
    i$value
  }
  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=fdelta1_a(x,a)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=fdelta1_a(x.boot,a)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Delta 1,a poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e
}
```

```r
#' Delta2 test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable
#

testdelta2 = function(x,n.boot){
  fdelta2b=function(z){
    fdelta2b1 =function(y,t) {
```

```r
      ((1/length(y)) * sum(t^y) -
         ((1/length(y)) * sum(((t+1)/2)^y))^2)^2
    }
    i =integrate(f = Vectorize(fdelta2b1,vectorize.args = 't'), lower = 0, upper = 1,
                 y = z)
    i$value
  }
  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=fdelta2b(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=fdelta2b(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Delta2 poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e
}
```

```r
#' Delta2,a test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable

testdelta2_a = function(x,n.boot,a){
  fdelta2_a=function(z,c){
    fdelta2b1 =function(y,t,b) {
      (((1/length(y)) * sum(t^y) -
         ((1/length(y)) * sum(((t+1)/2)^y))^2)^2) * t^b
    }
    i =integrate(f = Vectorize(fdelta2b1,vectorize.args = 't'), lower = 0, upper = 1,
                 y = z, b = c)
    i$value
  }
  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=fdelta2_a(x,a)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=fdelta2_a(x.boot,a)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
```

```
  names(media) <- "mean"
  e = list(method = paste("Delta2,a poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e
}
```

```
#' Delta inf test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable

testdeltainf = function(x,n.boot){
  fdeltainf=function(z) {
    if (sum(z) == 0) {
      0
    }
    else {
      fdeltainfabs= function(t) abs((mean(t^z))-(mean((((t+1)/2)^z)^2))
      OptDeltaInf=optimize(fdeltainfabs, lower = 0, upper = 1,
                           maximum = T)
      return(as.numeric(OptDeltaInf[2]))
    }
  }

  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=fdeltainf(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=fdeltainf(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Delta infinity poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e

}
```

```
#' Rn test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable
```

```r
# testRn = function(x,n.boot) {
testRn = function(x,n.boot){
Rn = function(z){
  Gn = function(t,y){
  (sqrt(length(y))*((1/length(y)) * sum(t^y) -
     exp(mean(y)*(t-1))))^2
  }
    i = integrate(f = Vectorize(Gn,vectorize.args = 't'), lower = 0, upper = 1,
             y = z)
    i$value
}
pv = 0
n = length(x)
lambda.hat=mean(x)
T.obs=Rn(x)
for (i in 1:n.boot){
  x.boot=rpois(n,lambda.hat)
  T.boot=Rn(x.boot)
  pv=pv+as.numeric(T.boot>T.obs)
}
names(T.obs) <- "test statistic"
media = mean(x)
names(media) <- "mean"
e = list(method = paste("Rn poissonity test",
                        sep = ""),
         statistic = T.obs, p.value = pv/n.boot,
         data.name = paste("sample size ", n, ", replicates ",
                           n.boot, sep = ""), estimate = media)
class(e) <- "htest"
e
}
```

```r
#' Rn,a test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable

testRna = function(x,a,n.boot){
  Rna = function(z,c){
    Gna = function(t,y,b){
      (sqrt(length(y))*((1/length(y)) * sum(t^y) -
                          exp(mean(y)*(t-1))))^2 * t^b
    }
    i = integrate(f = Vectorize(Gna,vectorize.args = 't'), lower = 0, upper = 1,
                  y = z, b = c)
    i$value
  }
  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=Rna(x,a)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
```

```r
    T.boot=Rna(x.boot,a)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Rn,a poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e
}
```

```r
#' Tn test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable

testTn = function(x,n.boot) {
  options(warn = -1)
testTn1 = function(x,n.boot){
  Tn1=function(z) {auxsuma=outer(z,z,"+")
  auxmult=outer(z,z,"*")
  tt=table(z)

  dd=as.numeric(names(tt))+1
  ni=rep(0,(max(z)+1))
  ni[dd]=as.numeric(tt)



  mauxsum=auxsuma[auxmult>0]
  mauxmult=auxmult[auxmult>0]



  TN=(1/length(z))*(sum(mean(z)^2/(auxsuma+1))+sum(mauxmult/(mauxsum-1))) - mean(z)*(length(z)-(1/length
  return(TN)
  }

  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=Tn1(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=Tn1(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
```

```r
  e = list(method = paste("Tn poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e
}



testTn2 = function(x,n.boot){
  Tn2=function(z) {
    f2noder= function(t) {mean(z) * mean(sum(t^z))}
    f2der= function(t) {mean(sum(t^z*z/t))}
    f3_a=function(t,b) {((f2noder(t) - f2der(t)) ^2)}
    b_a= integrate(Vectorize(f3_a, vectorize.args = 't'), lower=0, upper=1)
    b_aSol=as.numeric(b_a[1])
    return(b_aSol/length(z))
  }

  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=Tn2(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=Tn2(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Tn poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e
}

if (x > 140) {
  return(testTn2(x, n.boot))
}
else {return(testTn1(x,n.boot))}
}
```

```r
#' Tn,a test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable

testTn_a = function(x,a,n.boot){
```

```
testTn_a2 = function(x,a,n.boot){
  Tn_a2=function(z,a) {
    f2noder= function(t) {mean(z) * mean(sum(t^z))}
    f2der= function(t) {mean(sum(t^z*z/t))}
    f3_a=function(t,b) {((f2noder(t) - f2der(t)) ^2) * t^b}
    b_a= integrate(Vectorize(f3_a, vectorize.args = 't'), lower=0, upper=1, b = a)
    b_aSol=as.numeric(b_a[1])
    return(b_aSol/length(z))
  }

  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=Tn_a2(x,a)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=Tn_a2(x.boot,a)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Tn,a poissonity test",
                      sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
       data.name = paste("sample size ", n, ", replicates ",
                         n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e
}
testTn_a1 = function(x,a,n.boot){
  Tna1 = function(x,a){
    aux = outer(x,x,"+")
    aux1 = outer(x,x,"*")
    num1 = mean(x)^2
    den1 = aux + a + 1
    coc1 = sum(num1/den1)
    num2 = mean(x)*(aux)
    den2 = aux + a
    coc2 = sum(num2/den2)
    num3 = aux1
    den3 = aux + a - 1
    coc3 = sum(num3/den3)
    (coc1 - coc2 + coc3)/length(x)
  }

  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=Tna1(x,a)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=Tna1(x.boot,a)
```

```
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
  media = mean(x)
  names(media) <- "mean"
  e = list(method = paste("Tn,a poissonity test",
                          sep = ""),
           statistic = T.obs, p.value = pv/n.boot,
           data.name = paste("sample size ", n, ", replicates ",
                             n.boot, sep = ""), estimate = media)
  class(e) <- "htest"
  e
}
if (length(x) > 200) {
  return(testTn_a2(x,a,n.boot))
}
else {return(testTn_a1(x,a,n.boot))}
}
```

Test basado en momentos

```
#' U test
#' @export
#' @param x numeric variable
#' @param n.boot numeric variable
testu = function(x,n.boot){
  u <- function(z) {
    if (sum(z) == 0) {
      0
    }
    else if (sum(z) == 1) {
      0
    }
    else {
      b =var(z)
      c = mean(z)
      d = (b/c - 1)
      e = length(z)-1
      f = 2*(1-1/sum(z))
      g = sqrt(e/f)
      return (d * g)
    }
  }
  pv = 0
  n = length(x)
  lambda.hat=mean(x)
  T.obs=u(x)
  for (i in 1:n.boot){
    x.boot=rpois(n,lambda.hat)
    T.boot=u(x.boot)
    pv=pv+as.numeric(T.boot>T.obs)
  }
  names(T.obs) <- "test statistic"
```

```
  media = mean(x)
  names(media) <- "mean"

  if (n.boot > 1) {
    e = list(method = paste("U poissonity test",
                            sep = ""),
             statistic = T.obs, p.value = pv/n.boot,
             data.name =  paste("sample size ", n, ", replicates ",
                                n.boot, sep = ""), estimate = media)
    class(e) <- "htest"
    e
  }

  else {e = list(method = paste("U poissonity test",
                                sep = ""),
                 statistic = T.obs, p.value = 2*(1-pnorm(abs(T.obs))),
                 data.name = paste("sample size ", n, ", normal approximation ",
                                   sep = ""), estimate = media)
  class(e) <- "htest"
  e
  }
}
```

Capítulo 11

Conclusiones y futura líneas de investigación

Los test de bondad de ajuste para distribuciones diferentes a la normal son un campo no muy desarrollado dentro de la Estadística. En este trabajo nos hemos enfocado en detallar las bases de veinte test de bondad de ajuste de la distribución Poisson y un análisis mucho más exhaustivo de tres artículos en los cuales se proponen y desarrollan ocho de los veinte test mencionados. Hemos podido corroborar que existen numerosas formas de caracterizar la distribución Poisson y, por tanto, numerosos métodos de creación de test diferentes.

También hemos comprobado en el Capítulo 9 la importancia que tiene este tipo de test en la práctica, ya que permiten descartar o no Poissonidad en una muestra determinada que a la larga puede ser clave para otros temas estadísticos como hallar la probabilidad de un suceso ya sea a partir de la función de densidad o bien a través de replicaciones usando métodos como el bootstrap o el Montecarlo.

Algunas líneas de investigación que puede seguirse a partir de este trabajo son:

1. Publicación de un paquete o una librería como la creada en R que hemos visto en el Capítulo 10 en otros lenguajes de programación como Python.

2. Desarrollo programático para añadir al paquete TestPoissonity de algunos test como I_n o V_n 3.11, 3.7, los cuales no hemos podido llevar a cabo en el paquete creado en este trabajo ya que no hemos encontrado la forma de escribir las funciones de estos test para que sean ejecutado en un tiempo computacional razonable.

Bibliografía

[1] Baringhaus L., Gürtel, N., Henze, N. (2000) Weighted L^2-Statistics and components of smooth tests of fit. *Austral. New Zeal J. Statist., 42 (2), 179-192*

[2] Baringhaus, L., Henze, N. (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function. *Statististics and Probability Letters*, 13, 269–274.

[3] Billingsley, P. (1968), Convergence of probability mesures. Wiley

[4] Bosq, D. (2000). Linear process in function spaces: theory and applications.

[5] Bosq, D. (2007). Inference and Predictions in Large Dimension.

[6] Dixon, M., Coles, S. (1997) Modelling Association Football Scores and Inefficiencies in the Football Betting Market *Applied Statistics, Volume 46, Issue 2, 265-280*

[7] Gurtel N., Henze, N. (2000) Recent and classical goodness-of-fit test for the Poisson distribution. *Journal of Statistic Planning and Inference 90, (2000), 207-225*

[8] Henze, N. (1996) Empirical-distribution-function goodness-of-fit tests for discrete models. *Canad. J. Statist. 24(1), 81-93*

[9] Klar, B. (2000) Goodness-of-fit tests for discrete models based on the integrated distribution function. *Metrika, 49: 53-69*

[10] Kocherlakota S., Kocherlakota K. (1986) Goodness of fit test for discrete distributions. *Commum Statist. - Theory Methods 15, 815-838*

[11] Kokoszka, P., Lajos, H. (1992) Inference for Functional Data with Applications.

[12] Mahrer, M. J. (1982) Modelling association football scores. *Statist. Neerland, 36, 109-118*

[13] Nakamura, M., Pérez Abreu, V. (1993) Use of an empirical probability generating function for testing a Poisson model. *Canad J. Statist. 21 (2) 149-156*

[14] Novoa Muñoz, F., Jiménez-Gamero, M.D. (2014). Testing for the bivariate Poisson distribution. *Metrika*, 77, 771–793.

[15] Puig P., Kokonendji, C.C. (2018) Non-parametric Estimation of the Number of Zeros in Truncated Count Distributions. *Scand J. Statist. 45, 347-365*

[16] Puig, P, Weib C (2020). Some goodness-of-fit test for the Poisson distribution with the applications in Biodosimetry. *Computational Statistic and Data Analysis, Vol 144, Article 106878*

[17] Rohagi, V K., Ehsanes Saleh, A.K.Md. An introduction to probability and statistics.

[18] Rueda, R., Pérez Abreu V., O'Reilly, F. (1991) Goodness of fit for the Poisson distribution based on the probability generating function. *Commum. Statist - Theory Methods 20 (10), 3093-3110*

[19] Serfling, R J. (1980). Approximation Theorems of Mathematical Statistics.

[20] Székeley, G., Rizzo, M. (2004) Mean distance test of Poisson distribution. *Statistics and Probability Letters 67, 241-247*

[21] Treutler, B. (1995) Test for the Poisson distribution. *Diploma Tesis, University of Karlsruhe*

[22] Wang, D. (2010) Soccer tournament simulation and analysis for South Africa World Cup with Poisson model of goal probability. *2010 Chinese Control and Decision Conference*