



UNIVERSIDAD NACIONAL DE COLOMBIA

Estimación del parámetro λ y del número de ceros n_0 de la distribución Poisson Truncada en cero

Estimation of the parameter λ and the number of zeros n_0 of the Zero Truncated Poisson distribution.

Cristian Daniel Obando Arbeláez

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2020

Estimación del parámetro λ y del número de ceros n_0 de la distribución Poisson Truncada en cero

Estimation of the parameter λ and the number of zeros n_0 of the Zero Truncated Poisson distribution.

Cristian Daniel Obando Arbeláez

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Magister en Ciencias-Estadísticas

Director(a):
(Ph.D.) Juan Carlos Correa Morales

Línea de Investigación:
Estadística Bayesiana

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia

2020

A mi padre que siempre está en mis proyectos y
a mí madre que lo hace desde el cielo.

Agradecimientos

Agradezco a todo el equipo de profesores de la Escuela de Estadística de la facultad de ciencias de la Universidad Nacional por todas las enseñanzas impartidas. Hago una mención especial al profesor Juan Carlos Correa Morales quien es mi tutor y siempre ha estado atento a mis necesidades en la elaboración del proyecto y en los diferentes procesos de aprendizaje.

Resumen

La distribución Poisson Truncada en cero tiene múltiples aplicaciones en problemas de conteo. Por ejemplo, cuando se desea estimar el número de personas que han tenido, o tienen, problemas de adicción, se cuenta únicamente con información del número de ingresos de cada individuo y se desconoce el número de consumidores que no han ingresado a los centros de rehabilitación. En este trabajo se proponen diferentes estimadores puntuales y por intervalos para el parámetro λ y el número de ceros n_0 de la distribución Poisson Truncada en cero. Los estimadores puntuales y los intervalos son construidos mediante técnicas propias de la estadística clásica y bayesiana. Estos estimadores son comparados en conjunto con los encontrados en la literatura mediante simulación utilizando el software estadístico *R*. Se encontró que entre los estimadores puntuales el mejor es el de máxima verosimilitud modificada. En cuanto a los estimadores por intervalo el que tiene mayor probabilidad de cobertura fue el propuesto por Vélez and Correa (2013), no obstante el algoritmo para calcular este estimador fracasa con tamaños de muestra grandes, en este caso se prefieren el intervalo de confianza exacto o el de verosimilitud.

Palabras clave: Distribución Poisson Truncada, Datos censurados, Distribución Poisson Positiva, Estimación Bayesiana.

Abstract

The Poisson Truncated at Zero distribution has multiple applications in counting problems. For example, when you want to estimate the number of people who have had, or are having, addiction problems, you count only information on the number of incomes of each individual and you do not know the number of consumers who have not been admitted to rehabilitation centers. This paper proposes different point and interval estimators for the λ parameter and the number of zeros n_0 in the Poisson Truncated at Zero distribution. The point estimators and the intervals are constructed using classical and Bayesian statistical techniques. These estimators are compared in conjunction with those found in the literature through simulation using *R* statistical software. It was found that among the point estimators the best is the modified maximum likelihood one. As for the interval estimators, the one with the highest coverage probability was proposed by Vélez and Correa (2013), but the algorithm

for calculating this estimator fails with large sample sizes.

Keywords: Truncated Poisson Distribution, Censored Data, Positive Poisson Distribution, Bayesian Estimation

Contenido

Agradecimientos	vii
Resumen	ix
Lista de figuras	xiii
Lista de tablas	1
1 Introducción	2
2 Caracterización de la distribución Poisson estándar y Truncada en cero	5
2.1 Distribución Poisson	5
2.2 Distribución Poisson Truncada en Cero (DPTC)	6
3 Estimación Puntual de los parámetros de la distribución Poisson Truncada en Cero	8
3.1 Estimadores clásicos	8
3.1.1 Estimador de máxima verosimilitud	8
3.1.2 Estimador mínimo Chi-cuadrado	9
3.1.3 Estimación basada en la función de distribución acumulada truncada	9
3.1.4 Estimación basada en los percentiles teóricos y observados	10
3.1.5 Otros estimadores clásicos	10
3.2 Estimación puntual Bayesiana	11
3.2.1 Estimador máximo a posteriori usando el algoritmo EM	11
3.2.2 Media a posteriori	13
3.2.3 Máxima verosimilitud no restringida	13
3.2.4 Estimador de máxima verosimilitud modificada	14
3.2.5 Distribución Poisson sobre n_0	14
4 Estimación por intervalos para el parámetro λ de la distribución Poisson Truncada en Cero	16
4.1 Intervalos clásicos	16
4.1.1 Intervalos de confianza de wald	16
4.1.2 Intervalo de confianza exacto	17
4.1.3 Intervalo de verosimilitud	19

4.2	Intervalos bayesianos	19
4.2.1	Estimador por intervalos MAP	19
4.2.2	Estimador por intervalos con distribución Poisson sobre n_0	20
5	Estudio de simulación	21
5.1	Comparación de los estimadores puntuales	21
5.1.1	Resultados	23
5.2	Comparación de los intervalos de confianza	47
5.2.1	Resultados	47
6	Aplicaciones	54
6.1	Estimación del número de unidades de proteína presentes en las células	54
6.2	Estimación del número de adictos de marihuana en 2006 en Tailandia	55
6.3	Estimación del número de trabajadores en una empresa	56
6.4	Estimación del número total de hogares	57
7	Conclusiones	59
	Bibliografía	60

Lista de Figuras

4-1	$\log(p_1)$ y $\log(p_2)$ según tamaño de muestra	18
5-1	Probabilidad de obtener valores de cero en la muestra aleatoria Poisson . . .	22
5-2	<i>ECM</i> de los estimadores puntuales de λ para combinaciones de λ y $n(1)$. .	23
5-3	<i>ECM</i> de los estimadores puntuales de λ para combinaciones de λ y $n(2)$. .	24
5-4	<i>ECM</i> de los estimadores puntuales de λ para combinaciones de λ y $n(2)$. .	25
5-5	<i>ECM</i> de los estimadores puntuales de λ para combinaciones de λ y n - restringida(1)	26
5-6	<i>ECM</i> de los estimadores puntuales de λ para combinaciones de λ y n - restringida(2)	27
5-7	<i>ECM</i> de los estimadores puntuales de λ para combinaciones de λ y n - restringida(2)	28
5-8	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y $n(1)$	29
5-9	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y $n(2)$	30
5-10	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y $n(3)$	31
5-11	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y n - restringida(1)	32
5-12	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y n - restringida(2)	33
5-13	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y n - restringida(3)	34
5-14	<i>ECM</i> de los estimadores puntuales de n_0 para combinaciones de λ y $N(1)$. .	35
5-15	<i>ECM</i> de los estimadores puntuales de n_0 para combinaciones de λ y $N(2)$. .	36
5-16	<i>ECM</i> de los estimadores puntuales de n_0 para combinaciones de λ y $N(3)$. .	37
5-17	<i>ECM</i> de los estimadores puntuales de n_0 para combinaciones de λ y N . - restringida(1)	38
5-18	<i>ECM</i> de los estimadores puntuales de n_0 para combinaciones de λ y N . - restringida(2)	39
5-19	<i>ECM</i> de los estimadores puntuales de n_0 para combinaciones de λ y N . - restringida(3)	40
5-20	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y $n(1)$	41
5-21	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y $n(2)$	42
5-22	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y $n(3)$	43
5-23	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y n - restringida(1)	44
5-24	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y n - restringida(2)	45
5-25	<i>SP</i> de los estimadores puntuales para diferentes valores de λ y n - restringida(3)	46

5-26 Longitud de los intervalos de confianza para combinaciones de λ y $n(1)$. . .	48
5-27 Longitud de los intervalos de confianza para combinaciones de λ y $n(2)$. . .	49
5-28 Longitud de los intervalos de confianza para combinaciones de λ y $n(3)$. . .	50
5-29 Probabilidad de cobertura de los intervalos de confianza para combinaciones de λ y n	51
5-30 Probabilidad de cobertura de los intervalos de confianza para combinaciones de λ y n	52
5-31 Probabilidad de cobertura de los intervalos de confianza para combinaciones de λ y n	53

Lista de Tablas

6-1	Frecuencia con que se encuentran i etiquetas en un epítope de proteína(Cullen et al., 1990)	54
6-2	Estimaciones del promedio de etiquetas en un epítope (λ) y el número de epítopes sin etiquetas (n_0)	55
6-3	Estimaciones por intervalo para el promedio de etiquetas en un epítope . . .	55
6-4	Frecuencia con la cual un usuario de marihuana ha ingresado a tratamiento de adicción i veces (Tailandia-2006)	55
6-5	Estimaciones del promedio de ingresos a tratamiento (λ) y número de usuarios de marihuana que no han ingresado a tratamiento(n_0)	56
6-6	Intervalos de confianza para el promedio de ingresos a tratamiento (λ)	56
6-7	Frecuencia con que se encuentra un empleado que ha sufrido i accidentes laborales (Plackett, 1953)	56
6-8	Estimaciones del promedio de accidentes padecidos por un empleado (λ) y el número de empleados que no han padecido accidentes (n_0)	57
6-9	Intervalo de confianza del promedio de accidentes padecidos por un empleado (λ).	57
6-10	Frecuencia con la cual se presentan i casos de cólera en un hogar en una villa de la India.	57
6-11	Estimación del promedio de personas que presentan casos de cólera por hogar (λ) y el número de hogares sin casos (n_0)	58
6-12	Intervalos de confianza para el promedio de personas que presentan casos de cólera por hogar	58

1 Introducción

La Distribución Poisson Truncada en Cero (DPTC) surge de un proceso de conteo en el cual no es posible observar la variable aleatoria Poisson cuando toma el valor de cero. Esta distribución tiene aplicaciones en biología, medicina, industria entre otras áreas. Específicamente, algunas situaciones en las que aparece esta distribución son:

- En el conteo de las personas por hogar que padecen una enfermedad infecciosa, cuando no se registra el número de hogares sin enfermos (David and Johnson, 1952; Cohen, 1960). Para conocer el número total de hogares visitados se estima el número de hogares sin enfermos. Por ejemplo, Dahiya and Gross (1973) y Blumental et al. (1978) estimaron el número de hogares que no presentaron enfermos de cólera en una aldea de la India. También puede ser de interés estimar el promedio de enfermos por hogar λ .
- En el conteo del número de accidentes por trabajador en una fábrica, cuando no se tiene registro de los trabajadores que no sufrieron accidentes (David and Johnson, 1952; Cohen, 1960). Por ejemplo, Plackett (1953) encontró un estimador para el promedio de accidentes por trabajador λ . Por otro lado, si el objetivo es calcular la proporción de trabajadores que han sufrido un accidente se debe estimar el número de trabajadores que no presentaron.
- En algunas aplicaciones de las ciencias médicas. Por ejemplo, Matthews and Appleton (1993) estimaron el número de unidades (epítopes) de cierta proteína presentes en las fibras musculares. En este caso el truncamiento ocurre porque las unidades solo se identifican al estar etiquetadas, y cada unidad puede tener una, varias o no tener etiquetas.
- En algunas aplicaciones de la biología. Finney and Varley (1955) estimaron el número de flores en las que no se habían depositado huevos de un insecto, para posteriormente estimar la tasa de mortalidad de dichos insectos.
- Con base en información del número de entradas de personas adictas a la marihuana a los centros de rehabilitación, Viwatwongkasem (2016) estimó el número de adictos en Tailandia.

La amplia aplicabilidad de la distribución Poisson Truncada llevó a que varios autores se ocuparan de proponer estimadores puntuales para el promedio de eventos ocurridos por

unidad de tiempo y el número de unidades de tiempo en las que no ocurrieron eventos denotados por λ y n_0 respectivamente. Bajo el enfoque clásico se tienen los estimadores contruidos por Plackett (1953), Rider (1953), Tate and Goen (1958), Cohen (1960).

Suponiendo que el tamaño de muestra truncada n es fijo, David and Johnson (1952), Irwin (1959), Cohen (1960) y Viwatwongkasem (2016) encontraron el estimador de máxima verosimilitud para el parámetro λ . Asimismo, Plackett (1953) encontró un estimador para λ mediante el método de momentos. Este autor mostró que el cociente de varianzas entre el estimador de máxima verosimilitud y el suyo nunca cae por debajo de 0,95 y tiende a 1 cuando $\lambda \rightarrow 0$ o $\lambda \rightarrow \infty$. Adicionalmente, Finney and Varley (1955) en un ejercicio aplicado encontraron estimaciones aproximadamente iguales utilizando estos dos estimadores.

El procedimiento de estimación implementado por Plackett (1953) es generalizado por Moore (1954). Este autor encontró el estimador de λ en distribuciones truncadas a la derecha o izquierda de un valor arbitrario. Otro procedimiento general usado para estimar λ cuando la distribución Poisson es truncada a la izquierda fue propuesto por Rider (1953). En un ejercicio de simulación para diferentes tamaños de muestras y valores de λ , se encontró en muchos casos que este estimador presenta un menor Error Cuadrático Medio (ECM) que el de máxima verosimilitud.

Tate and Goen (1958) estimaron el parámetro λ cuando los datos están truncados a la izquierda de un valor arbitrario. Estos autores encontraron un estimador insesgado de varianza mínima partiendo de las propiedades de los estadísticos suficientes. Matthews and Appleton (1993) compararon este estimador con el de máxima verosimilitud y el propuesto por Plackett (1953) mediante un ejercicio de simulación, obteniendo en promedio un valor más cercano al verdadero valor del parámetro con el estimador de Tate and Goen (1958), y la menor desviación estándar con el estimador de máxima verosimilitud.

El enfoque bayesiano fue implementado por Dahiya and Gross (1973), Blumental et al. (1978), Mir (2011) y Vélez and Correa (2013). Dahiya and Gross (1973) determinan los estimadores de máxima verosimilitud sin restringir, para los parámetros λ y n_0 de la DPTC. A su vez, Blumental et al. (1978) encuentran los estimadores de los parámetros λ y n_0 mediante máxima verosimilitud no restringida y modificada. Mir (2011) asumió una distribución a priorí gamma para el parámetro λ y propone como estimador para λ la media a posterior. Este autor mediante un ejercicio de simulación obtuvo que el estimador bayesiano tiene un menor ECM que el de máxima verosimilitud. Adicionalmente, Vélez and Correa (2013) asume una distribución gamma sobre λ y Poisson sobre n_0 .

Con respecto a la estimación por intervalos bajo el enfoque clásico Daidoji and Iwasaki (2012) propuso un intervalo de confianza tipo wald para estimar el parámetro λ . Y bajo el enfoque

bayesiano Vélez and Correa (2013) proponen realizar bootstrap a la distribución a posteriori resultante al asumir una a priori gamma sobre λ y Poisson sobre n_0 . Otra forma para construir intervalos de confianza es utilizar los estimadores de n_0 y utilizar intervalos de confianza propuestos para la distribución Poisson. No obstante, este enfoque no es considerado pues se tendrían que analizar una más amplia gama de estimadores.

En este trabajo de grado se proponen nuevos estimadores puntuales y por intervalos para el parámetro λ y n_0 de la DPTC. Los estimadores puntuales propuestos para el parámetro λ son el mínimo Chi-cuadrado, uno basado en la función de distribución acumulada truncada, uno basado en la comparación de los percentiles teóricos con los observados y estimador máximo a posteriori (MAP). Para n_0 se propuso el MAP. Con respecto a la estimación por intervalos de λ se propusieron el intervalo de verosimilitud, el exacto y el MAP. Los estimadores propuestos son comparados en conjunto con los encontrados en la literatura vía simulación, implementando medidas tales como el ECM, el estimador promedio y el sesgo promedio, asumiendo distintas combinaciones de tamaños de muestra truncada (n) y valores del parámetro λ . La simulación se llevará a cabo mediante el software estadístico *R* (R Core Team, 2018) y los nuevos estimadores se obtendrán usando técnicas frecuentistas y bayesianas.

2 Caracterización de la distribución Poisson estándar y Truncada en cero

2.1. Distribución Poisson

La distribución Poisson es utilizada en diferentes problemas de conteo. La forma funcional de esta distribución es:

$$p_x(\lambda) = p_x = P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!} \text{ para } x = 0, 1, \dots, \quad (2-1)$$

donde λ es el promedio de eventos ocurridos por unidad de tiempo.

La función generatriz de probabilidad para esta distribución es:

$$P_1(t) = \exp(\lambda(t - 1)) , \quad (2-2)$$

La expresión (2-3) puede ser utilizada para calcular probabilidades a partir de la siguiente relación:

$$P(X = r) = \frac{P_1^{(r)}(0)}{r!} \text{ para } r = 0, 1, 2, \dots, \quad (2-3)$$

donde:

$P_1^{(r)}(0)$: es la r -ésima derivada de la función generatriz de probabilidad evaluada en cero.

$$P_1^{(0)}(0) = P_1(0)$$

Adicionalmente la función generatriz de momentos es:

$$M_1(t) = E[\exp(tx)] = \exp(\lambda(\exp(t) - 1)) , \quad (2-4)$$

$M_1(t)$ puede ser utilizada para encontrar el r -ésimo momento alrededor de la media para la variable aleatoria x de la siguiente manera:

$$E[X^r] = M_1^{(r)}(0) , \quad (2-5)$$

donde:

$M_1^{(r)}(0)$: Es la r -ésima derivada de la función generatriz de momentos evaluada en $t=0$.

A partir de (2-5) se puede mostrar que la media y la varianza para la distribución Poisson están dadas por:

$$E[X] = Var[X] = \lambda , \quad (2-6)$$

Ahora considerando una muestra aleatoria X_1, X_2, \dots, X_N de tamaño N , de una distribución Poisson y denotando n_x a los valores dentro de la muestra iguales a x para $x = 0, 1, 2, \dots$. En este caso, la función de verosimilitud tiene la forma:

$$L_1(\lambda) = \frac{\lambda^{\sum_{x=1}^k x n_x} \exp\left(-\sum_{x=0}^k n_x \lambda\right)}{\prod_{x=1}^k (x!)^{n_x}}, \quad (2-7)$$

donde k es el máximo valor alcanzado en la muestra.

Relación entre la Distribución Poisson y la Multinomial Negativa

Considere el conjunto de variables aleatorias N_0, N_1, \dots, N_{k+1} donde N_x es el número de valores iguales a x observados en una muestra Poisson, para $x = 0, 1, 2, \dots, k$ y $N_k + 1$ son los valores mayores o iguales a $k + 1$. Asumiendo que $N_k = n_k$ es conocido y que $N = \sum_{x=0}^{k-1} N_x + n_k + N_{k+1}$ se tiene que $N_0, N_1, \dots, N_{k-1}, N_{k+1}$, siguen una distribución multinomial negativa con parámetros n_k y p_0, p_1, \dots, p_{k+1} :

$$\binom{N-1}{N_0, N_1, \dots, n_k-1, N_{k+1}} \prod_{x=0}^{k-1} p_x^{N_x} p_k^{n_k} p_{k+1}^{N_{k+1}}, \quad (2-8)$$

donde:

$$p_{k+1} = 1 - \sum_{x=0}^k p_x$$

$$N_x \in \{0, 1, \dots\} \text{ para } x = 0, 1, \dots, k-1, k+1$$

Esta relación se utilizará en la implementación del algoritmo EM (Expectation-Maximization). Como su nombre lo indica este algoritmo consiste en estimar los valores faltantes mediante su valor esperado y posteriormente maximizar la función de verosimilitud utilizando los datos iniciales y los estimados. Para más información ver Dempster et al. (1977); Rizzo (2007)

2.2. Distribución Poisson Truncada en Cero (DPTC)

Si asumimos que en la muestra aleatoria Poisson los valores de $x = 0$ no son observables la función de distribución (2-1) no representa en forma adecuada el comportamiento de los datos. En este caso la función de distribución es:

$$\pi_x(\lambda) = \pi_x = P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x! (1 - \exp(-\lambda))} \text{ para } x = 1, 2, \dots, \quad (2-9)$$

La expresión (2-9) es conocida como la función DPTC. La función generatriz de probabilidad para esta distribución está dada por:

$$P_2(t) = \frac{\exp(\lambda t) - 1}{\exp(\lambda) - 1}, \quad (2-10)$$

Y la función generatriz de momentos tiene la forma:

$$M_2(t) = \frac{\exp(\lambda \exp(t)) - 1}{\exp(\lambda) - 1}, \quad (2-11)$$

A partir de $M_2(t)$ se obtiene la media:

$$\mu = E(X) = \frac{\lambda}{1 - \exp(-\lambda)}, \quad (2-12)$$

Y varianza:

$$\sigma^2 = Var(X) = \frac{\lambda + \lambda^2}{1 - \exp(-\lambda)} - \frac{\lambda^2}{(1 - \exp(-\lambda))^2}, \quad (2-13)$$

La función de verosimilitud está dada por

$$L_2(\lambda) = \frac{\lambda^{\sum_{x=1}^k n_x x} \exp(-n\lambda)}{(1 - \exp(-\lambda))^n \prod_{x=1}^k (x!)^{n_x}}, \quad (2-14)$$

donde $n = \sum_{x=1}^k n_x$ es el tamaño de la muestra truncada en cero.

3 Estimación Puntual de los parámetros de la distribución Poisson Truncada en Cero

3.1. Estimadores clásicos

La estimación puntual clásica hace referencia a los métodos que utilizan la información disponible para obtener un valor aproximado de un parámetro o vector de parámetros. Algunos de los estimadores presentados en esta sección son el de máxima verosimilitud, el mínimo chi-cuadrado, el de varianza mínima, uno basado en la función de distribución acumulada, uno basado en la comparación de los deciles teóricos y observados, entre otros.

3.1.1. Estimador de máxima verosimilitud

El estimador de máxima verosimilitud es el valor de λ que maximiza la función dada en (2-14). Este estimador se obtiene al resolver la siguiente ecuación mediante métodos numéricos (David and Johnson, 1952; Irwin, 1959; Cohen, 1960):

$$\frac{\lambda}{(1 - e^{-\lambda})} = \frac{\sum_{x=1}^k x n_x}{n}, \quad (3-1)$$

Otra forma de encontrar el estimador de máxima verosimilitud es utilizando el algoritmo EM. Para la implementación de este algoritmo se utilizan las siguientes dos ecuaciones (Vivatwongkasem, 2016):

$$n_0^{(r)} = \frac{n \exp(-\lambda^{(r)})}{1 - \exp(-\lambda^{(r)})}, \quad (3-2)$$

$$\hat{\lambda}^{(r)} = \sum_{i=1}^k \frac{i n_i}{\hat{n}_0^{(r)} + n}, \quad (3-3)$$

Posteriormente se realizan los siguientes pasos (Dempster et al., 1977; Rizzo, 2007):

1. Se parte de un valor arbitrario $\hat{\lambda}^{(0)}$. Se comienza en la iteración $r = 0$.

2. Se calcula $n_0^{(r)}$ reemplazando λ_0 en la ecuación (3-2).
3. Con el $n_0^{(r)}$ obtenido en el paso anterior se estima $\lambda^{(r+1)}$ usando la ecuación (3-3).
4. Comienza la iteración $r = r + 1$ y se repiten los pasos 2 y 3 hasta la convergencia.

Los estimadores de máxima verosimilitud del parámetro λ y n_0 se denotan por $\hat{\lambda}_1$ y \hat{n}_{01} respectivamente.

3.1.2. Estimador mínimo Chi-cuadrado

El primer estimador para el parámetro λ propuesto en este trabajo es el mínimo Chi-cuadrado (EMCC). El EMCC para la DPTC es el valor $\hat{\lambda}_2$ que cumple la condición:

$$\sum_{x=1}^k \frac{((n_x - n\pi_x(\hat{\lambda}_2)))^2}{n\pi_x(\hat{\lambda}_2)} \leq \sum_{x=1}^k \frac{((n_x - n\pi_x(\lambda)))^2}{n\pi_x(\lambda)} \quad (3-4)$$

Para todo $\lambda \in \Omega$. Donde Ω es el espacio parametral de λ

La condición (3-4) es equivalente a:

$$\sum_{x=1}^k \left[\left((n_x - n\pi_x(\lambda)) + \frac{(n_x - n\pi_x(\lambda))^2}{2n\pi_x(\lambda)} \right) \left(\frac{x}{\lambda} - 1 - \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \right) \right] = 0, \quad (3-5)$$

La ecuación (3-5) debe ser resuelta numéricamente. Para una descripción general del procedimiento de estimación mínimo chi-cuadrado ver Harris and Kanji (1983); Mood et al. (1974). Berkson (1980) señala que el EMCC se puede obtener para los parámetros de cualquier distribución y en muchos casos es igual al de máxima verosimilitud. A su vez este autor muestra que tener un estimador de máxima verosimilitud no garantiza que se cumpla el principio de suficiencia para las diferentes muestras.

3.1.3. Estimación basada en la función de distribución acumulada truncada

El segundo estimador para el parámetro λ propuesto se basa en la función de distribución acumulada truncada. Este enfoque consiste en encontrar el parámetro λ que minimiza las diferencias al cuadrado entre la función DPTC y la frecuencia relativa acumulada observada. Sea $F_\lambda(x)$ la función de distribución de probabilidad acumulada para una variable aleatoria Poisson truncada en cero:

$$F_\lambda(x) = P(X \leq x) = \sum_{t=1}^x \frac{e^{-\lambda} \lambda^t}{x!(1 - e^{-\lambda})}, \quad (3-6)$$

Para $x = 1, 2, \dots$,

La frecuencia relativa acumulada hasta x está dada por:

$$F_x = \sum_{t=1}^x \frac{n_t}{n}, \quad (3-7)$$

A partir de $F_\lambda(x)$ y F_x se puede construir la función $Q(\lambda)$ dado por:

$$Q(\lambda) = \sum_{x=1}^{\infty} (F_\lambda(x) - F_x)^2, \quad (3-8)$$

Un estimador para el parámetro λ es un $\hat{\lambda}_3$ tal que:

$$Q(\hat{\lambda}_3) \leq Q(\lambda), \quad (3-9)$$

Para todo $\lambda \in \Omega$.

3.1.4. Estimación basada en los percentiles teóricos y observados

El tercer estimador para el parámetro λ propuesto en este trabajo es el basado en los percentiles teóricos y observados. Bajo este enfoque se escoge el valor del parámetro λ que minimiza la diferencia entre los percentiles observados y los teóricos. Sea $q(\lambda, \alpha)$ el percentil del $\alpha \times 100\%$ de la DPTC y sea $\hat{q}(\alpha)$ el percentil observado del $\alpha \times 100\%$. Haciendo $\beta_t = pt$ donde $p = 0.01, 0.02, \dots, 0.1$ se define la función $K(\lambda)$ como:

$$K(\lambda) = \sum_{t=1}^h (q(\lambda, \beta_t) - \hat{q}(\beta_t))^2, \quad (3-10)$$

donde $h = 1/p - 1$. Un estimador para el parámetro λ es un $\hat{\lambda}_4$ tal que:

$$k(\hat{\lambda}_4) \leq k(\lambda), \quad (3-11)$$

para todo $\lambda \in \Omega$

3.1.5. Otros estimadores clásicos

En los años 50 se presentaron diferentes propuestas para encontrar estimadores del parámetro λ . Debido a la dificultad que representaba encontrar el estimador de máxima verosimilitud por el escaso desarrollo de los computadores, Plackett (1953) propuso el siguiente estimador:

$$\hat{\lambda}_5 = \sum_{x=2}^k \frac{n_x x}{n}, \quad (3-12)$$

Atacando un problema más general Rider (1953) propone un estimador del parámetro λ cuando los datos están truncados a la izquierda de un valor arbitrario. En el caso en que los datos están truncados en cero se tiene:

$$\hat{\lambda}_6 = \frac{\sum_{x=1}^k x^2 n_x - \sum_{x=1}^k x n_x}{\sum_{x=1}^k x n_x}, \quad (3-13)$$

Del mismo modo, Tate and Goen (1958) encontró un estimador insesgado de varianza mínima para el parámetro λ en el caso en que los datos están truncados a la izquierda de un valor arbitrario. En el caso en que los datos están truncados en cero el estimador está dado por (Mattews and Appleton, 1993) :

$$\hat{\lambda}_7 = \begin{cases} \sum_{x=1}^k x n_x \frac{S_t^{n-1}}{S_t^n} & \text{para } t > n, \\ 0 & \text{en otro caso,} \end{cases} \quad (3-14)$$

donde:

$S_t^n = \sum_{q=0}^n (-1)^{n-q} \binom{n}{q} \frac{q^t}{n!}$ es conocido como el número de Stirling de segundo tipo y $t = \sum_{x=1}^k x x$

3.2. Estimación puntual Bayesiana

En esta sección se presentan estimadores bayesianos puntuales asumiendo diferentes ditribuciones a priori sobre n_0 y λ , y diferentes funciones de pérdida.

3.2.1. Estimador máximo a posteriori usando el algoritmo EM

Otro estimador propuesto para el parámetro λ en este trabajo es el estimador máximo a posteriori denotado como MAP. La obtención de este estimador se hace mediante la maximización de la función de distribución a posterior (Rangan et al., 2012). Dado que n_0 es desconocido este proceso de maximización se realiza utilizando el algoritmo EM. La implementación de este algoritmo requiere encontrar el valor esperado de las observaciones desconocidas. Para lograr este objetivo es necesario utilizar la ecuación (2-8) y encontrar la función de distribución de probabilidad condicional de N_0 dado $N_1 = n_1, N_2 = n_2, \dots, N_{k+1} = n_{k+1}$ y λ :

$$P(N_0 | n_1, n_2, \dots, n_{k+1}) = \binom{N_0 + n - 1}{N_0} p_0^{N_0} (1 - p_0)^n, \quad (3-15)$$

donde:

$$N_0 \in \{0, 1, \dots, \}$$

$n = \sum_{x=1}^{k+1} n_x$ Dado que la ecuación (3-15) es la función de distribución binomial negativa con párametro p_0 entonces la esperanza condicional de N_0 dado $N_1 = n_1, N_2 = n_2, \dots, N_{k+1} = n_{k+1}, \lambda$ está dada por:

$$\hat{n}_{08} = E [N_0 | n_1, n_2, \dots, n_{k+1}, \lambda] = n \frac{p_0}{1 - p_0}, \quad (3-16)$$

Ahora suponiendo que se tiene una muestra aleatoria en la cual N_0 no es observable, $n_{k+1} = 0$ y en la que el parámetro λ tiene una distribución a priori $gamma(\alpha, \beta)$. La función de distribución a posteriori de λ está dada por:

$$\epsilon(\lambda|N_0, n_1, n_2, \dots, n_k) \propto \lambda^{\sum_{x=1}^k x n_x + \alpha - 1} \exp(-(\beta + n + n_0 \lambda)), \quad (3-17)$$

Aplicando logaritmo natural a la ecuación (3-17) se obtiene:

$$\log(\epsilon(\lambda|N_0, n_1, n_2, \dots, n_k)) \propto C + \left(\sum_{x=1}^k x n_x + \alpha - 1 \right) \log \lambda - \left(\sum_{x=1}^k n_x + n_0 + \beta \right) \lambda, \quad (3-18)$$

donde:

C es un término que no depende del parámetro λ .

Para encontrar el estimador máximo a posteriori mediante el algoritmo EM primero se calcula el valor esperado de la ecuación (3-18) dado n_1, n_2, \dots, n_k y λ . Es decir, el valor esperado bajo la función de distribución (3-15), por tanto la variable aleatoria es N_0 :

$$E(\log(\epsilon(\lambda|N_0, n_1, n_2, \dots, n_k))|n_1, n_2, \dots, n_k, \lambda) \propto C + \left(\sum_{x=1}^k x n_x + \alpha - 1 \right) \log \lambda - \left(\hat{n}_{08} - \sum_{x=1}^k n_x + \beta \right) \lambda, \quad (3-19)$$

donde \hat{n}_{08} está dado por la ecuación (3-16).

Derivando la ecuación (3-19) con respecto a λ se obtiene:

$$\frac{dE(\log(\epsilon(\lambda|N_0, n_1, n_2, \dots, n_k))|n_1, n_2, \dots, n_k, \lambda)}{d\lambda} = \frac{\left(\sum_{x=1}^k x n_x + \alpha - 1 \right)}{\lambda} - \left(\hat{n}_{08} - \sum_{x=1}^k n_x + \beta \right), \quad (3-20)$$

Igualando la ecuación (3-20) a cero y despejando λ se obtiene:

$$\hat{\lambda}_8 = \frac{\sum_{x=1}^k x n_x + \alpha - 1}{\beta + \hat{n}_{08} + n}, \quad (3-21)$$

utilizando las ecuaciones (3-16) y (3-21) se puede aplicar el algoritmo EM mediante los pasos descritos en la sección 3.1.1. Note que el estimador de n_0 es denotado como \hat{n}_{08} esto es para mantener congruencia con la notación usada para el estimador propuesto para λ a saber $\hat{\lambda}_8$

3.2.2. Media a posteriori

En este enfoque a diferencia del anterior no se utiliza la estimación de \hat{n}_{08} . Mir (2011) y Hassan et al. (2008) asumen una distribución Gamma a priori para el parámetro λ y proponen estimar este parámetro de la media de la distribución a posteriorí:

$$\hat{\lambda}_9 = \frac{\int_0^{\infty} \exp(-(\beta+n)\lambda) \lambda^{\sum_{i=1}^n x_i + \alpha} (1 - \exp(-\lambda))^{-n} d\lambda}{\int_0^{\infty} \exp(-(\beta+n)\lambda) \lambda^{\sum_{i=1}^n x_i + \alpha - 1} (1 - \exp(-\lambda))^{-n} d\lambda}, \quad (3-22)$$

Mir (2011) y Hassan et al. (2008) reescriben:

$$\hat{\lambda}_9 = \frac{S + \alpha}{(n + \beta) + n B(S, n)}, \quad (3-23)$$

donde:

$$S = \sum_{i=1}^n x_i \quad B(S, n) = \frac{M(S, n+1)}{M(S, n)}$$

Los términos $M(S, n+1)$ y $M(S, n)$ son obtenidos mediante la relación recursiva:

$$M(S, n+1) = \frac{M(S-1, n)}{n} - \frac{(n+\beta)M(S, n)}{n}, \quad (3-24)$$

con condición inicial:

$$M(S, 1) = \sum_{k=1}^{\infty} \frac{1}{(\beta+k)^{S+\alpha+1}},$$

Otra forma de encontrar este estimador es utilizando integración Monte Carlo.

3.2.3. Máxima verosimilitud no restringida

Sanathanan (1977) presenta el procedimiento general para encontrar los estimadores de máxima verosimilitud no restringida. Estos estimadores pueden ser considerados dentro del marco bayesiano pues suponen que el tamaño de la muestra truncada (n) es una variable aleatoria con distribución binomial con parámetros N y $p(\lambda)$:

$$B(n; N, \lambda) = \binom{N}{n} p_0(\lambda)^n (1 - p_0(\lambda))^{N-n} \quad \text{para } n = 1, 2, \dots, N, \quad (3-25)$$

La función de máxima verosimilitud para la muestra truncada está dada por la ecuación (2-14). De este modo la función de máxima verosimilitud conjunta de N y λ es:

$$L_3(N, \lambda) = B(n; N, p_0(\lambda)) L_2(\lambda), \quad (3-26)$$

Para obtener los estimadores de máxima verosimilitud no restringida se maximiza (3-26) con respecto N y λ , estos estimadores se obtienen a partir de las ecuaciones (Blumental et al., 1978):

$$\hat{\lambda}_{10} = \frac{\sum_{i=1}^n x_i}{N}, \quad (3-27)$$

$$\frac{\log\left(\frac{\hat{N}+1}{\hat{N}+1-n}\right)}{\log\left(\frac{\hat{N}+1}{\hat{N}}\right)} < \sum_{i=1}^n x_i \leq \frac{\log\left(\frac{\hat{N}}{\hat{N}-n}\right)}{\log\left(\frac{\hat{N}}{\hat{N}-1}\right)}, \quad (3-28)$$

Usando la ecuación (3-28) se obtiene \hat{N}_{10} que es un estimador para el tamaño de muestra completa N y por tanto $\hat{n}_{010} = \hat{N}_{10} - n$ donde n es el tamaño de muestra truncada como se mencionó anteriormente.

3.2.4. Estimador de máxima verosimilitud modificada

Para obtener los estimadores de máxima verosimilitud modificada, Blumental et al. (1978) impuso densidades a priori sobre λ y N :

$$b(\lambda) = \frac{r^{\rho+1} \lambda^{\rho} e^{-r\lambda}}{\Gamma(\rho+1)}; \text{ para } \rho > -1 \text{ y } r > 0, \quad (3-29)$$

$$P(N = j) = 1/K_0; \text{ para } j = 1, 2, \dots, N_0, \quad (3-30)$$

donde:

K_0 es el máximo valor que puede alcanzar N .

La función (3-29) corresponde a una re-parametrización de la distribución gamma. Las ecuaciones para obtener los estimadores para N y λ son:

$$\frac{\log\left(1 - \frac{n}{\hat{N}+1}\right)}{\log\left(1 - \frac{1}{\hat{N}+1+r}\right)} < \sum_{x=1}^k x n_x + \rho \leq \frac{\log\left(1 - \frac{n}{\hat{N}}\right)}{\log\left(1 - \frac{1}{\hat{N}+r}\right)}, \quad (3-31)$$

En este caso los estimadores se denotan \hat{N}_{11} y \hat{n}_{011} . El estimador para λ está dado por:

$$\hat{\lambda}_{11} = \frac{\sum_{i=1}^n x_i + \rho}{\hat{N}_{11} + r} \quad (3-32)$$

3.2.5. Distribución Poisson sobre n_0

Asumiendo una distribución $Gamma(\alpha, \beta)$ a priori sobre λ y $Poisson(\lambda_0)$ sobre n_0 Vélez and Correa (2013) encuentran estimadores bayesianos puntuales y por intervalos utilizando el muestreador de Gibbs. En este caso las distribuciones condicionales de λ y n_0 están dadas por:

$$\epsilon(\lambda | \text{Datos}, n_0) \propto \lambda^{\alpha+s-1} \exp(-\lambda(n + \beta + n_0)), \quad (3-33)$$

$$\epsilon(n_0 | \text{Datos}, \lambda) \propto \frac{(n + n_0)!}{n_0!^2} \lambda_0^{n_0} \exp(-n_0 \lambda), \quad (3-34)$$

Para implementar el muestreador de Gibbs se realizan los siguientes pasos:

- Se selecciona un punto arbitrario $(\lambda^{(0)}, n_0^{(0)})$
- Genere $\lambda^{(i)}$ de (3-33) asumiendo $n_0 = n_0^{(i-1)}$ (inicialmente $n_0^{(i-1)} = n_0^{(0)}$) y $n_0^{(i)}$ de (3-34) asumiendo $\lambda = \lambda^{(i)}$.
- Repita el paso anterior 5.000 veces.
- Elimine los primeros 1.000 λ y n_0 generados.

A partir de la muestra generada se obtiene los estimadores de λ y n_0 como la media y mediana respectivamente. En este caso los estimadores se denotan $\hat{\lambda}_{12}$ y \hat{n}_{012} .

4 Estimación por intervalos para el parámetro λ de la distribución Poisson Truncada en Cero

4.1. Intervalos clásicos

Un estimador por intervalos clásico utiliza la información muestral para estimar un conjunto de valores entre los cuales probablemente está el verdadero valor del parámetro. Los estimadores por intervalos clásicos considerados son los de Wald, razón de verosimilitud y el exacto.

4.1.1. Intervalos de confianza de wald

Daidoji and Iwasaki (2012) encuentra un intervalo de confianza de Wald (IC_1) despejando λ de la ecuación:

$$\frac{\hat{\lambda}_1 - \lambda}{\sqrt{\lambda(1 - \exp(-\lambda))^2 / (n(1 - (1 + \lambda)\exp(-\lambda)))}} = \pm Z_{\alpha/2}, \quad (4-1)$$

donde:

$Z_{\frac{\alpha}{2}}$ es el percentil del $(1 - \frac{\alpha}{2}) \times 100$ de la distribución normal

Una variación del IC_1 obtenido a partir de la ecuación (4-1) se puede hacer aplicando la transformación logarítmica al parámetro λ (IC_2):

$$\frac{\log(\hat{\lambda}_1) - \log(\lambda)}{\sqrt{(1 - \exp(-\lambda))^2 / (n\lambda(1 - (1 + \lambda)\exp(-\lambda)))}} = \pm Z_{\alpha/2} \quad (4-2)$$

Para una explicación de como se construyen intervalos de confianza tipo Wald para diferentes transformaciones del parámetro de interés ver Pawitan (2001, cap. 2).

La transformación a los parámetros de interés en la construcción de los intervalos tipo wald “mejoran” el desempeño de los estimadores por intervalos, siempre y cuando haya una mejor aproximación al logaritmo de la verosimilitud por medio de la siguiente expresión (Pawitan,

2001, cap. 2).

$$\log \left(\frac{L_2(\lambda)}{L_2(\hat{\lambda}_1)} \right) \approx -\frac{1}{2} I(\hat{\lambda}_1) (\lambda - \hat{\lambda}_1)^2, \quad (4-3)$$

donde $I(\hat{\lambda}_1) = -d^2 \log L_2(\lambda) / d\lambda^2 |_{\lambda=\hat{\lambda}_1}$.

Considerando la transformación $\psi = \log(\lambda)$, la aproximación cuadrática está dada por:

$$\log \left(\frac{L_2(\psi)}{L_2(\hat{\psi}_1)} \right) \approx -\frac{1}{2} I(\hat{\psi}_1) (\psi - \hat{\psi}_1)^2, \quad (4-4)$$

donde $\hat{\psi}_1 = \exp(\hat{\lambda}_1)$ y $I(\hat{\psi}_1) = -d^2 \log L_2(\psi) / d\psi^2 |_{\psi=\hat{\psi}_1}$.

4.1.2. Intervalo de confianza exacto

El primer estimador por intervalo propuesto en este trabajo es el exacto. Para la obtención del intervalo de confianza exacto (IC_3) es necesario conocer la distribución de $S = \sum_{i=1}^n x_i$. Lograr esto conlleva a utilizar la función generatriz de probabilidad de esta variable:

$$P_s(t) = \left(\frac{\exp(\lambda t) - 1}{\exp(\lambda) - 1} \right)^n, \quad (4-5)$$

El calculo de esta función es descrito en (Sarabia et al., 2006)[Cap. 5]. De la estadística matemática se sabe que :

$$P(X = i) = \frac{P_s^i(0)}{i!}, \quad (4-6)$$

donde $P_s^i(0)$ es la i -ésima derivada de la función generatriz de probabilidad evaluada en $t=0$. Por tanto, la función de densidad de probabilidad de S está dada por:

$$P(S = s) = \frac{\sum_{r=0}^n (-1)^r \binom{n}{r} (n-r)^s \lambda^s}{s! (\exp(\lambda) - 1)^n}, \quad (4-7)$$

El límite inferior del intervalo de confianza exacto es el valor de λ_L que satisface la ecuación:

$$P(S \geq s) = \sum_{k=1}^s \frac{\sum_{r=0}^n (-1)^r \binom{n}{r} (n-r)^k \lambda_L^k}{k! (\exp(\lambda_L) - 1)^n} \leq \alpha/2, \quad (4-8)$$

Y el límite superior es λ_U que satisface:

$$P(S \leq s) = \sum_{k=s}^{\infty} \frac{\sum_{r=0}^n (-1)^r \binom{n}{r} (n-r)^k \lambda_U^k}{k! (\exp(\lambda_U) - 1)^n} \leq \alpha/2, \quad (4-9)$$

La utilización de las ecuaciones (4-8) y (4-9) para estimar $P(S \leq s)$ no es recomendable a partir de cierto tamaño de muestra, pues se presentan probabilidades negativas. Por ejemplo, en el caso que se desea calcular $P(S = n)$ se puede emplear la ecuación (4-5) o la expresión:

$$P(S = n) = P(x_1 = 1, x_2 = 1, \dots, x_n = 1) = \frac{\lambda^n}{(\exp(\lambda) - 1)^n}, \quad (4-10)$$

El cálculo de $P(S = n)$ debería ser equivalente al usar las expresiones (4-8) y (4-10), no obstante debido a las limitaciones computacionales se producen errores de cálculo obteniendo incluso probabilidades negativas. En el cálculo de $P(S = n)$ la mejor aproximación está dada por la expresión (4-10).

La Figura 4-1 muestra el logaritmo del valor absoluto de la probabilidad calculada a partir de la expresión (4-7)(p_1) y (4-10)(p_2) asumiendo $\lambda=0.5$. En la gráfica p_1+ representa las probabilidades que dieron positivas y p_1- las negativas, en este caso se utilizó el valor absoluto para aplicar logaritmo. Se puede observar que cuando el tamaño de muestra es mayor o igual a 25 las probabilidades p_1 y p_2 se comienzan a alejar de su verdadero valor y p_1 toma valores negativos.

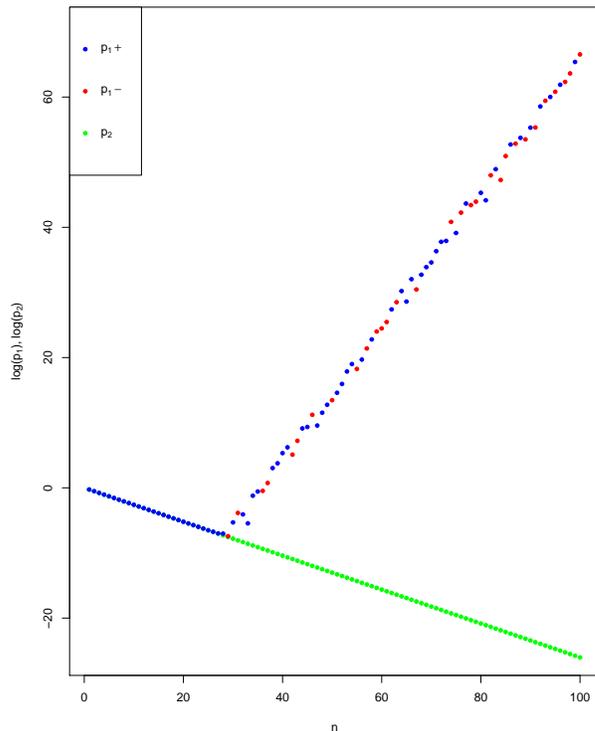


Figura 4-1: $\log(p_1)$ y $\log(p_2)$ según tamaño de muestra

Para paliar este problema se utiliza el teorema central del límite:

$$\sum_{i=1}^n x_i \sim_{aprox} N(n\mu, n\sigma^2), \quad (4-11)$$

donde μ y σ^2 están dadas en las expresiones (2-12) y (2-13) respectivamente. A partir de (4-11) es posible obtener:

$$P(S \geq s) \leq \alpha/2, \quad (4-12)$$

$$P(S \leq s) \leq \alpha/2, \quad (4-13)$$

Tsiatis et al. (1984) describe una metodología para obtener intervalos de confianza exactos.

4.1.3. Intervalo de verosimilitud

El segundo estimador por intervalo propuesto es el de razón de verosimilitud. Tomando la función de verosimilitud dada en la ecuación (2-14) y el estimador de máxima verosimilitud $\hat{\lambda}_1$ se obtiene la razón de verosimilitud:

$$R(\lambda) = \left(\frac{(1 - \exp(-\hat{\lambda}_1))}{1 - \exp(-\lambda)} \right)^n \left(\frac{\lambda}{\hat{\lambda}_1} \right)^{\sum_{x=1}^k x n_x} \exp(-n(\lambda - \hat{\lambda}_1)), \quad (4-14)$$

donde $0 < R(\lambda) < 1$. A partir de la ecuación (4-14) se define el intervalo de verosimilitud (IC_4) como aquel que satisface $R(\theta) \geq p$ donde $0 < p < 1$. Para tamaños de muestra grande $-2 \log R(\theta) \sim \chi_1^2$. El procedimiento para construir estos intervalos es descrito por Pawitan2001 y Kalbfleisch (1985)[Cap. 9]

4.2. Intervalos bayesianos

En esta sección se presentan dos intervalos bayesianos. El primero se construye a partir de la propiedad de invarianza del estimador bayesiano. El segundo se construye a partir de bootstrap, mediante la implementación del muestrador de Gibbs.

4.2.1. Estimador por intervalos MAP

El tercer estimador por intervalos propuesto es el MAP. Partiendo del estimador MAP del parámetro n_0 es posible definir el estimador por intervalos (L_i, L_s) del $(1 - \alpha)100\%$ para el parámetro λ (IC_5) (?)[Cap. 9]:

$$\int_{L_i}^{\infty} \frac{(n + \beta + n_0)^{\sum_{x=1}^k x n_x + \alpha}}{\Gamma(\sum_{x=1}^k x n_x + \alpha)} \lambda^{\sum_{x=1}^k x n_x + \alpha - 1} \exp(-\lambda(n + n_0 + \beta)) d\lambda = 1 - \frac{\alpha}{2} \quad (4-15)$$

$$\int_0^{L_s} \frac{(n + \beta + n_0)^{\sum_{x=1}^k x n_x + \alpha}}{\Gamma(\sum_{x=1}^k x n_x + \alpha)} \lambda^{\sum_{x=1}^k x n_x + \alpha - 1} \exp(-\lambda(n + n_0 + \beta)) d\lambda = 1 - \frac{\alpha}{2} \quad (4-16)$$

Debido a la propiedad de invarianza este es un estimador MAP para el parámetro λ . El estimador por intervalos para el parámetro n_0 puede ser obtenido reemplazando L_i y L_s en la ecuación (3-16).

4.2.2. Estimador por intervalos con distribución Poisson sobre n_0

Retomando el modelo bayesiano propuesto por Vélez and Correa (2013), de la muestra obtenida mediante el muestreador de Gibbs se toman los cuantiles del $\alpha/2$ y del $1 - \alpha/2$ para λ y n_0 y de este modo se obtienen intervalos del $(1 - \alpha)100\%$ (IC_6). Ver sección de estimadores puntuales bayesianos para conocer mejor dicho procedimiento.

5 Estudio de simulación

5.1. Comparación de los estimadores puntuales

La comparación de los estimadores puntuales se realiza mediante el Error Cuadrático Medio (*ECM*) y el Sesgo Promedio (*SP*). Se prefieren aquellos estimadores con menor *ECM* y *SP*. El *ECM* para los estimadores de λ se calcula a partir de la expresión:

$$ECM_{\lambda} = \sum_{i=1}^m \frac{(\hat{\lambda}_i - \lambda)^2}{m} \quad (5-1)$$

donde:

m : Es el número de muestras generadas

$\hat{\lambda}_i$: Es el estimador para el parámetro λ encontrado a partir de la i -ésima muestra.

Para n_0 el *ECM* es:

$$ECM_{n_0} = \sum_{i=1}^m \frac{(\hat{n}_{0i} - n_{0i})^2}{m} \quad (5-2)$$

donde:

n_{0i} : Es el número de ceros en la i -ésima muestra

\hat{n}_{0i} : Es el estimador del número de ceros en la i -ésima muestra

Debido al diseño de la simulación mientras el verdadero valor del parámetro en el cálculo de ECM_{λ} se mantiene fijo entre las muestras utilizadas, en el cálculo de ECM_{n_0} este valor varía de muestra en muestra.

El *SP* para λ y n_0 están dados por:

$$SP_{\lambda} = \sum_{i=1}^m \frac{\hat{\lambda}_i - \lambda}{m} \quad (5-3)$$

$$SP_{n_0} = \sum_{i=1}^m \frac{\hat{n}_{0i} - n_{0i}}{m} \quad (5-4)$$

En la comparación de los estimadores de λ y n_0 se consideran tamaños de muestra truncada n iguales a 10, 20, 30, ..., 500 y N iguales a 50, 60, 70, ..., 500, respectivamente. La elección de estos valores para n y N se da para tener una buena representación gráfica, además no

se consideraron valores mayores a 500 para ambos parámetros pues las mayores diferencias entre los estimadores se dan en los tamaños muestrales pequeños. En ambos casos se asumen valores del parámetro $\lambda = 0.5, 1, 1.5, \dots, 4.5$. La elección de estos valores de λ se basa en el hecho de que a medida que incrementa el valor de este parámetro es menor la probabilidad de obtener ceros en la muestra aleatoria. La Figura 5-1 muestra la probabilidad de obtener valores iguales a cero para diferentes niveles de λ .

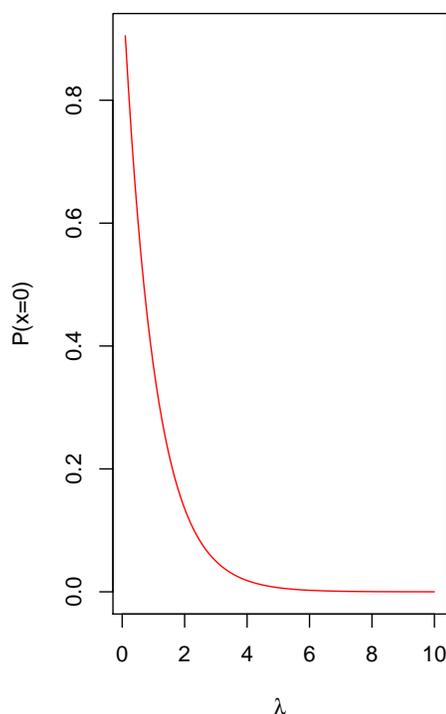


Figura 5-1: Probabilidad de obtener valores de cero en la muestra aleatoria Poisson

Para poder conocer el verdadero valor de n_0 en una muestra dada se genera una muestra aleatoria de una distribución Poisson de tamaño N y se eliminan los ceros. En cuanto a los estimadores bayesianos se asume que el valor de $\beta = 2$ y que $\alpha = 2 * \lambda$, es decir el valor esperado de la distribución a priori es igual al verdadero valor del parámetro. Dado que λ_{12} presenta problemas computacionales en su calculo, como se mencionó anteriormente, para tamaños de muestra grande, este intervalo solo se calcula hasta $n = 150$.

5.1.1. Resultados

Estimadores puntuales para λ

Las Figuras 5-2 a 5-4 muestran el ECM de los estimadores puntuales de λ para combinaciones de λ y n . Se observa que el estimador $\hat{\lambda}_{12}$ es el que tiene menor ECM , no obstante este estimador solo puede ser calculado para muestras relativamente pequeñas. Con respecto a los demás estimadores, los que tienen mayor ECM son $\hat{\lambda}_2$, $\hat{\lambda}_3$, $\hat{\lambda}_4$, $\hat{\lambda}_5$, $\hat{\lambda}_6$ y $\hat{\lambda}_7$. Adicionalmente se destaca que el ECM se hace más pequeño al incrementar n y λ .

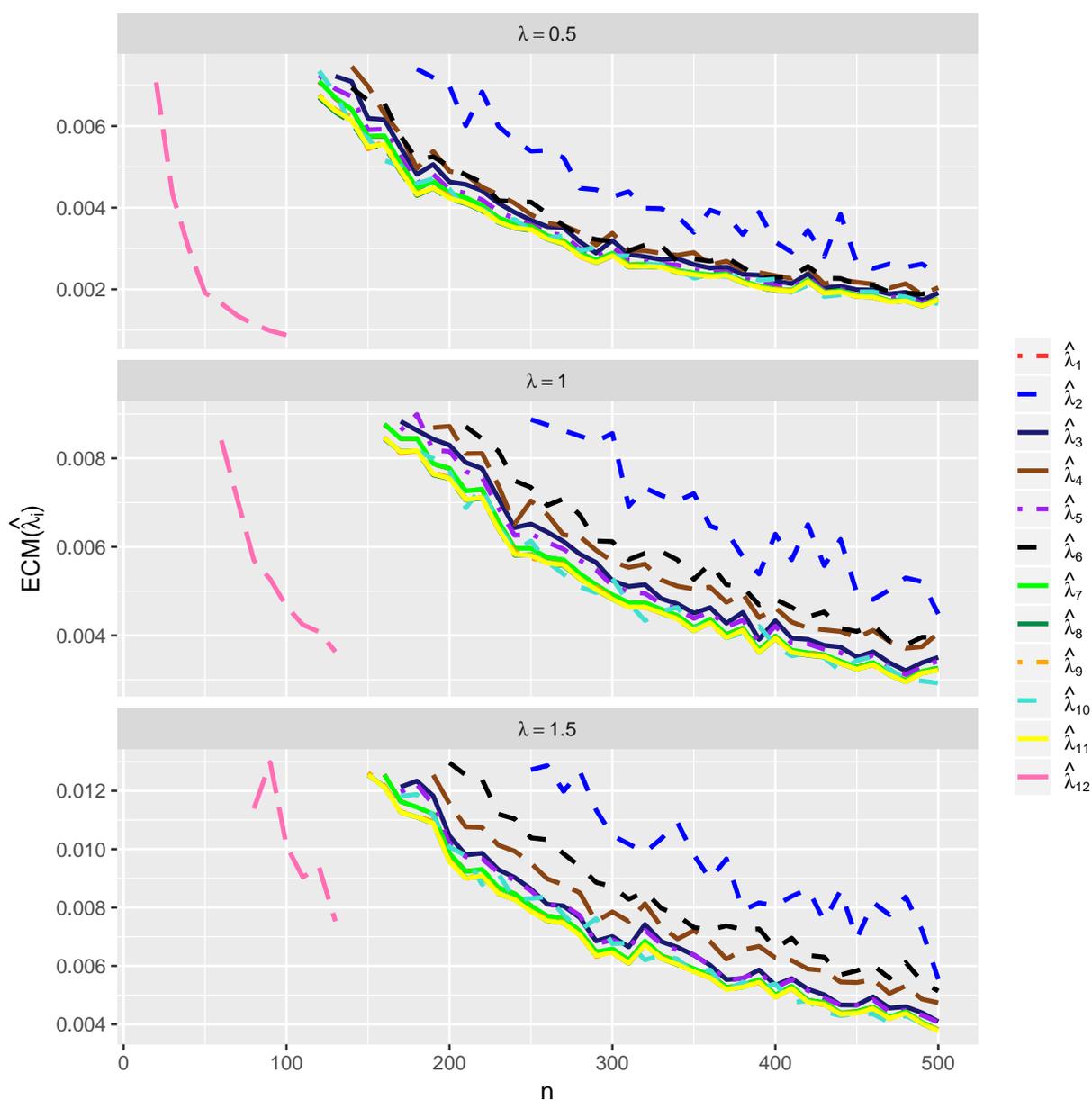


Figura 5-2: ECM de los estimadores puntuales de λ para combinaciones de λ y $n(1)$

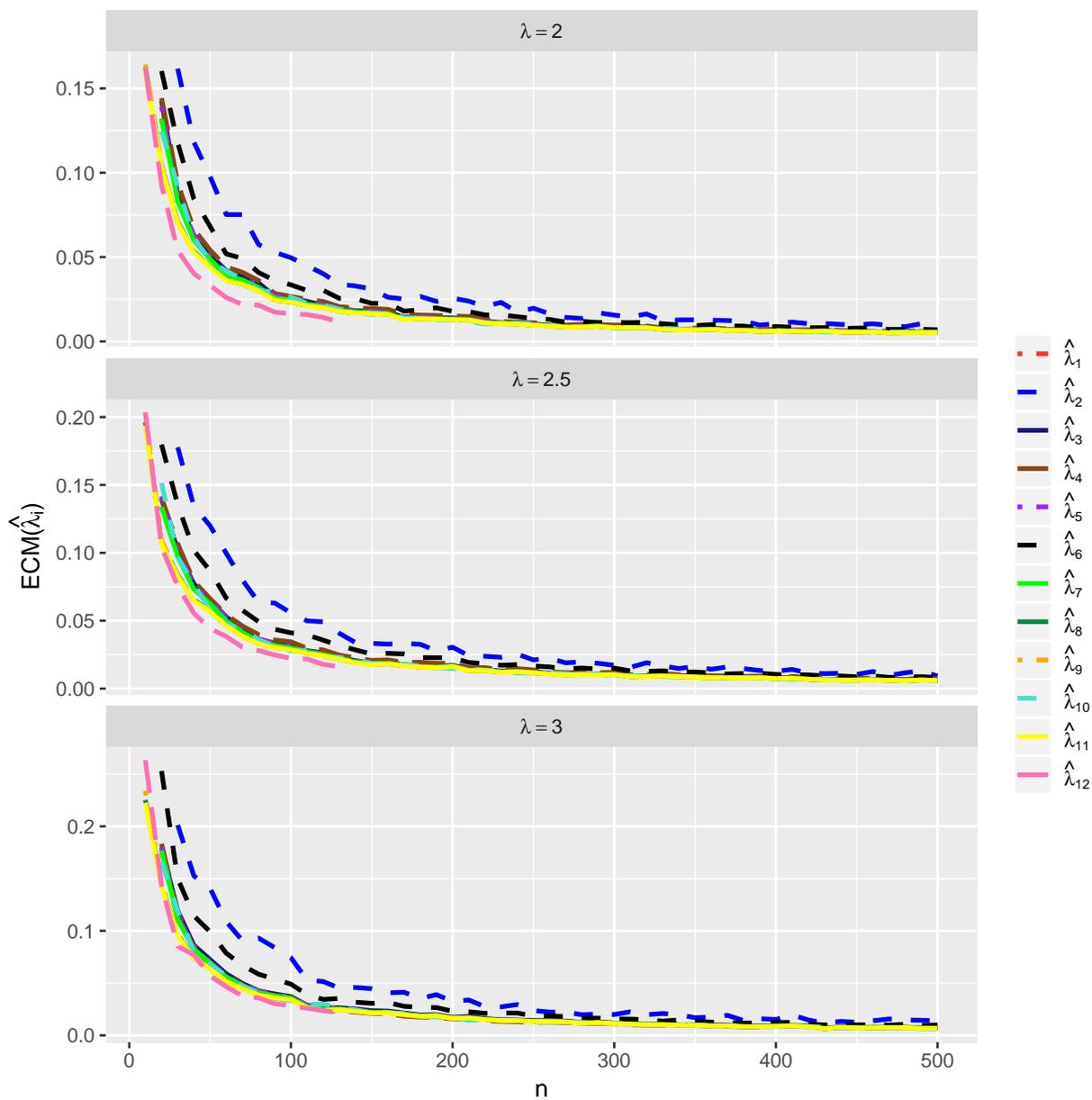


Figura 5-3: ECM de los estimadores puntuales de λ para combinaciones de λ y $n(2)$

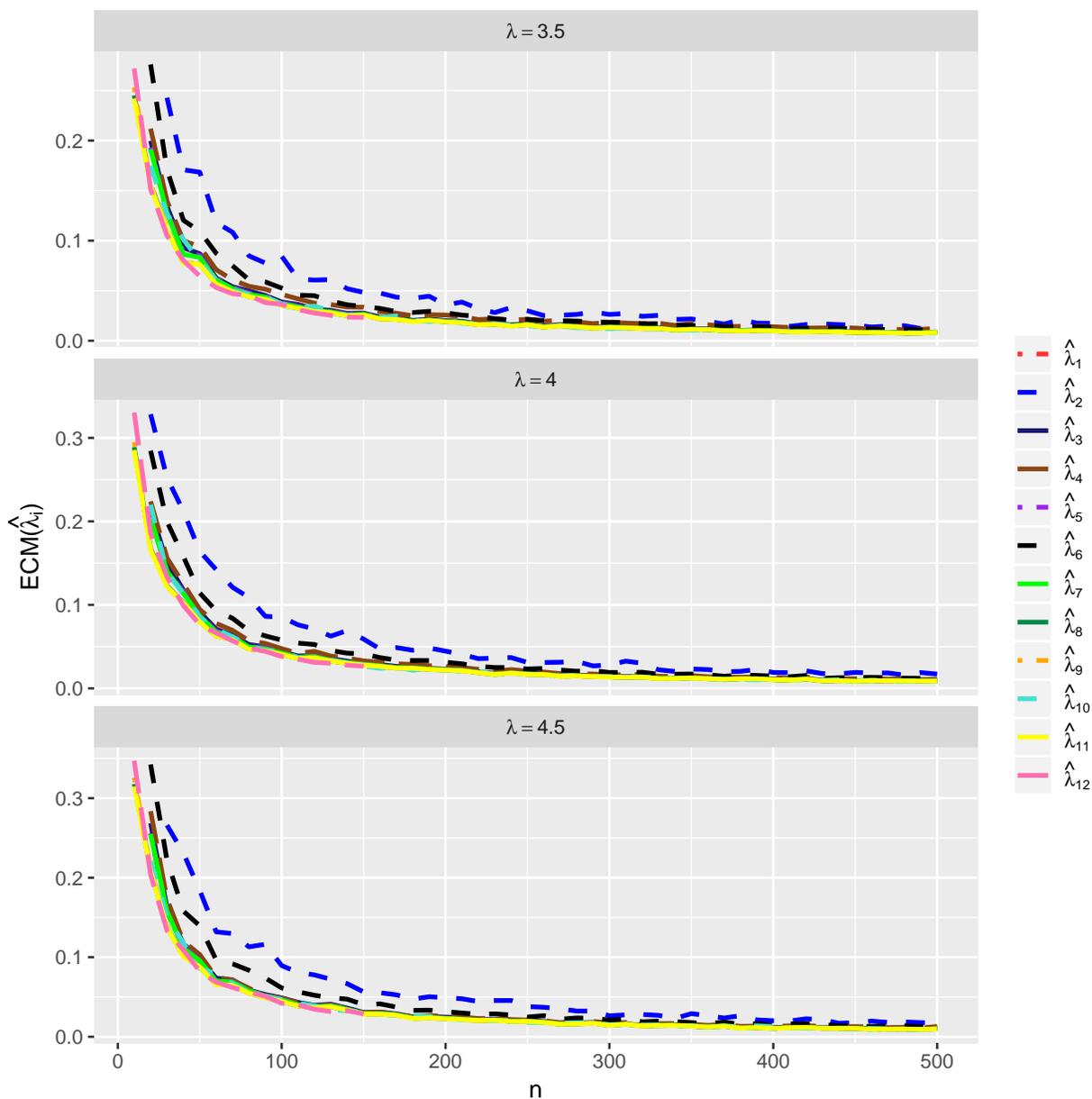


Figura 5-4: *ECM* de los estimadores puntuales de λ para combinaciones de λ y $n(2)$

Las Figuras 5-5 a 5-7 muestran el *ECM* de los estimadores puntuales de λ para combinaciones de λ y n , restringiendo n entre 150 y 300 y descartando $\hat{\lambda}_{12}$ y los estimadores *ECM*. Esta figura se restringe para observar mejor el comportamiento de los estimadores. Se observa que $\hat{\lambda}_1$ y $\hat{\lambda}_7$ son los estimadores con mayor *ECM*. En cuanto a $\hat{\lambda}_{10}$ el *ECM* puede ser mayor o menor a el de los otros estimadores para los diferentes tamaños de muestra.

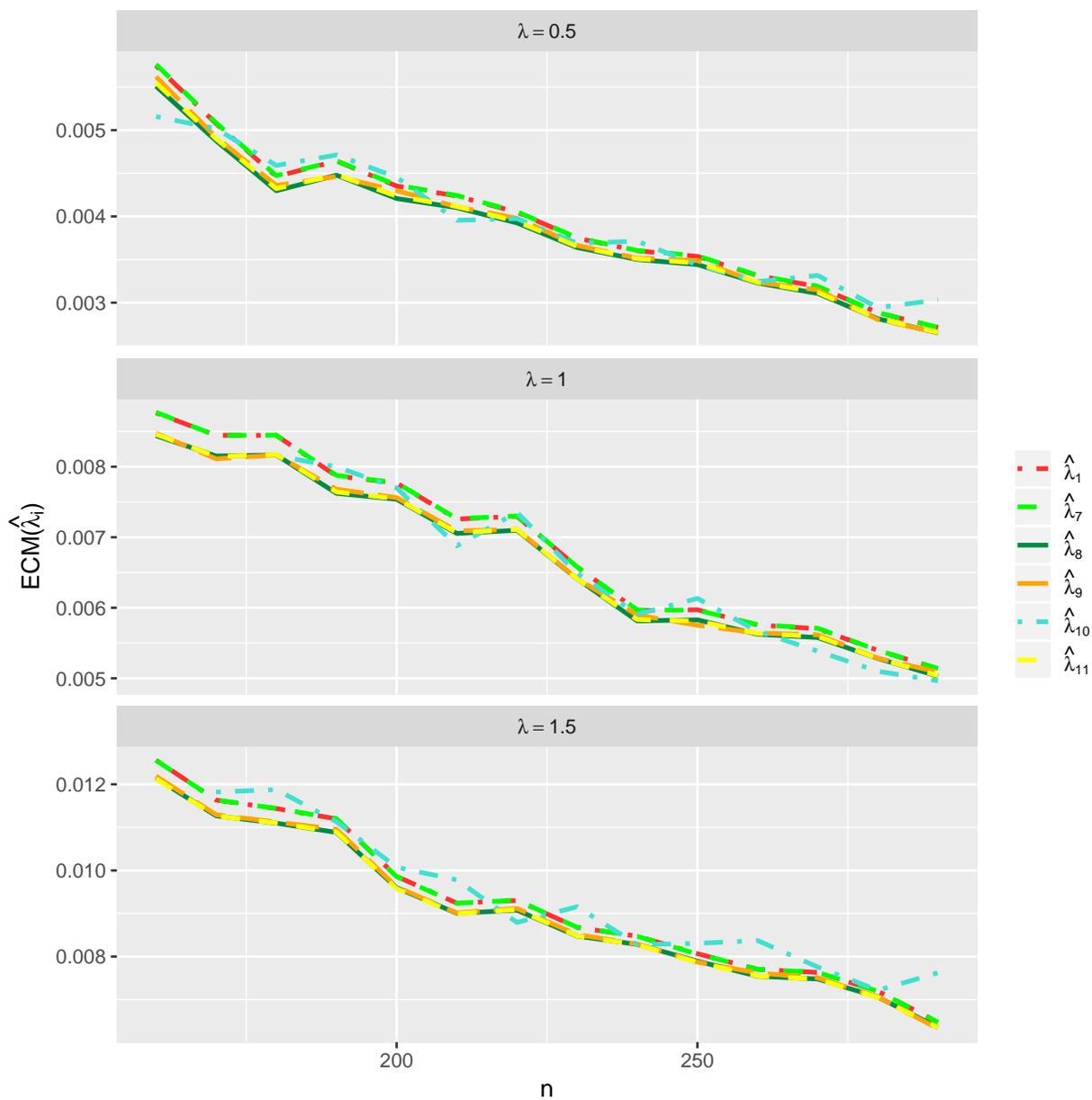


Figura 5-5: *ECM* de los estimadores puntuales de λ para combinaciones de λ y n - restringida(1)

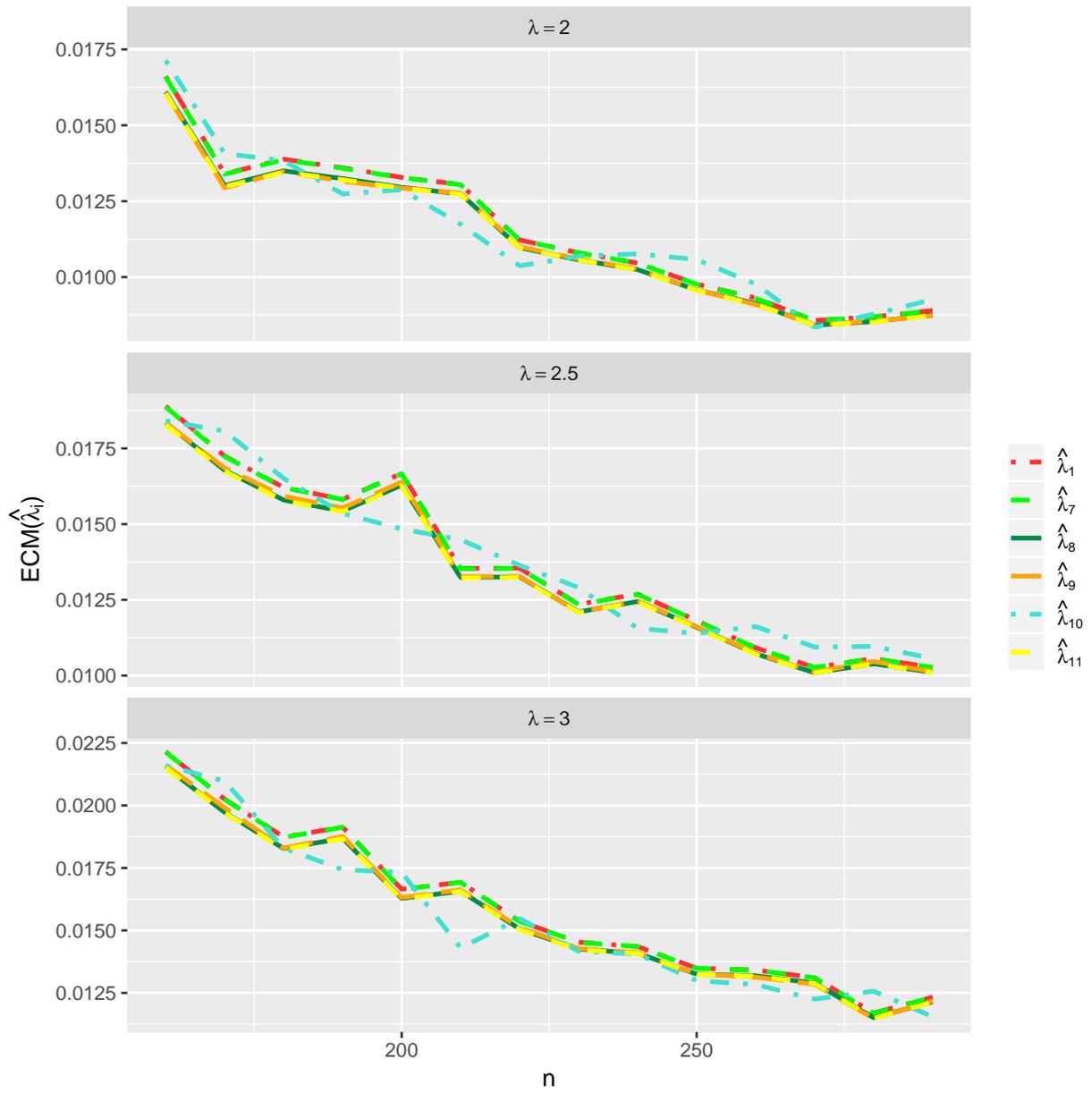


Figura 5-6: *ECM* de los estimadores puntuales de λ para combinaciones de λ y n - restringida(2)

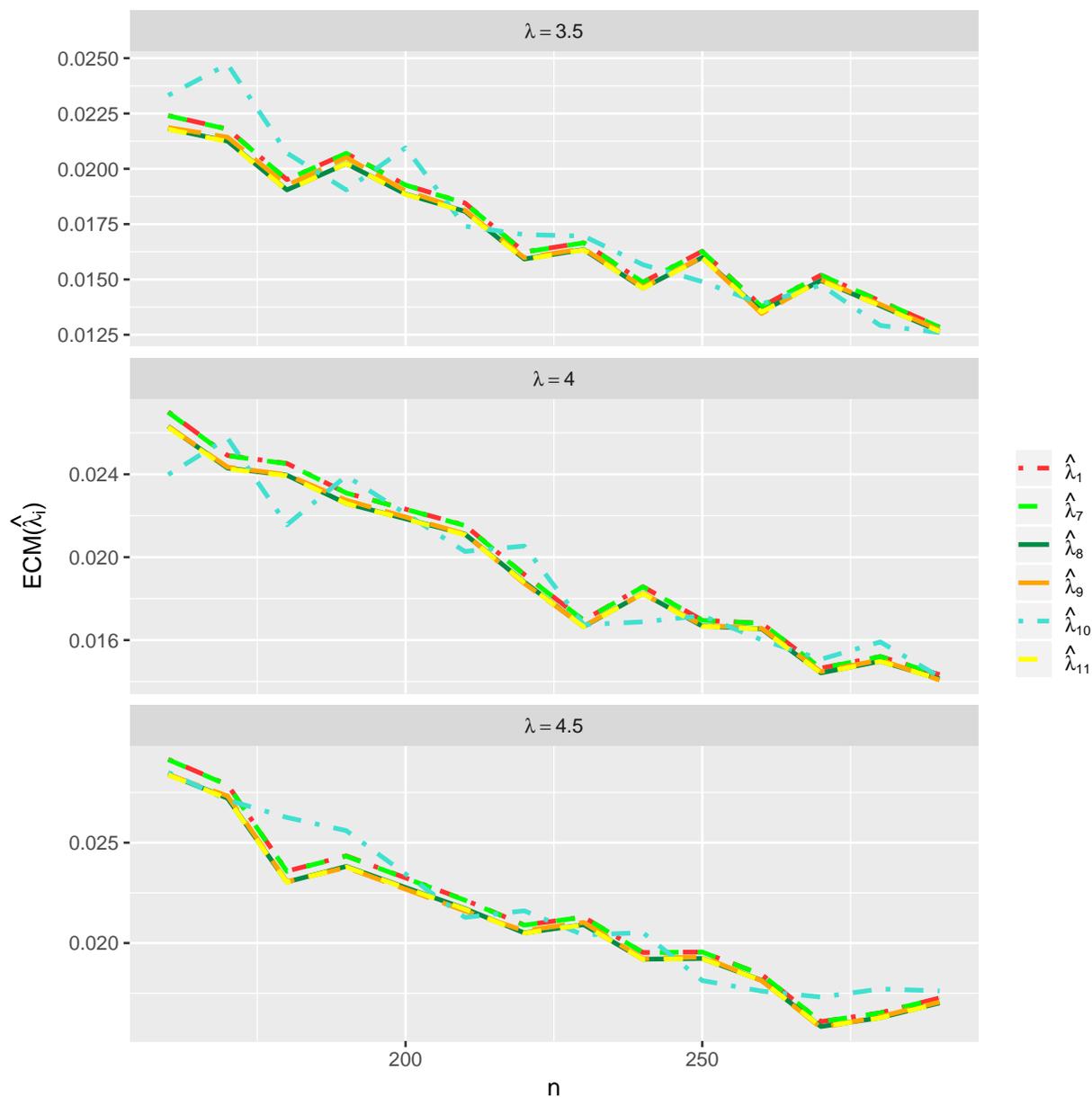


Figura 5-7: *ECM* de los estimadores puntuales de λ para combinaciones de λ y n - restringida(2)

Las Figuras 5-8 a 5-10 muestran el *SP* de los estimadores puntuales de λ para diferentes valores de λ y n . Por un lado, se observa que $\hat{\lambda}_2$, $\hat{\lambda}_9$ y $\hat{\lambda}_{12}$ sobre-estiman a λ . Por el otro, $\hat{\lambda}_4$ y $\hat{\lambda}_6$ lo subestiman. Adicionalmente se observa que el *SP* se hace más pequeño al incrementar n o λ .

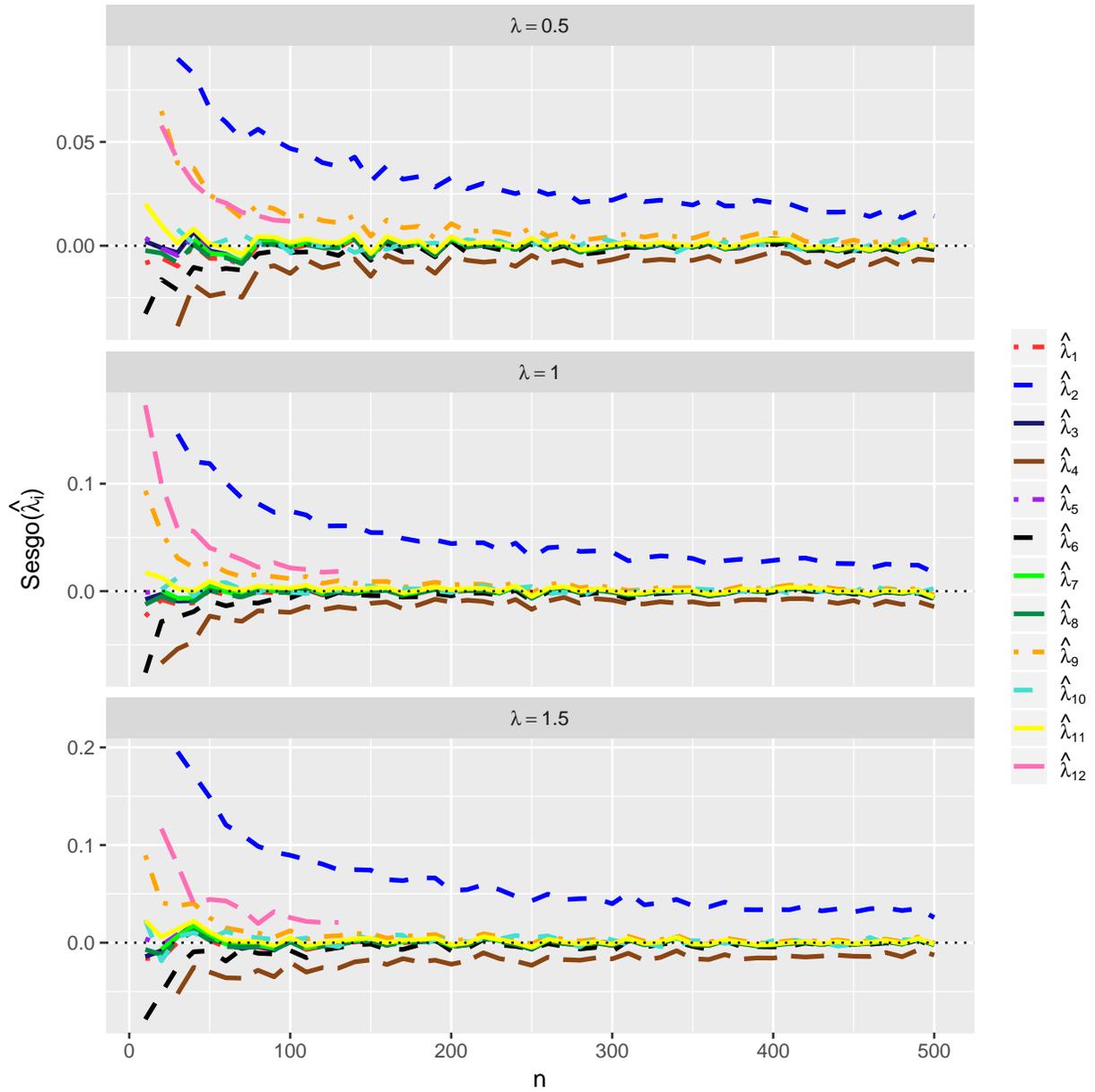


Figura 5-8: SP de los estimadores puntuales para diferentes valores de λ y $n(1)$

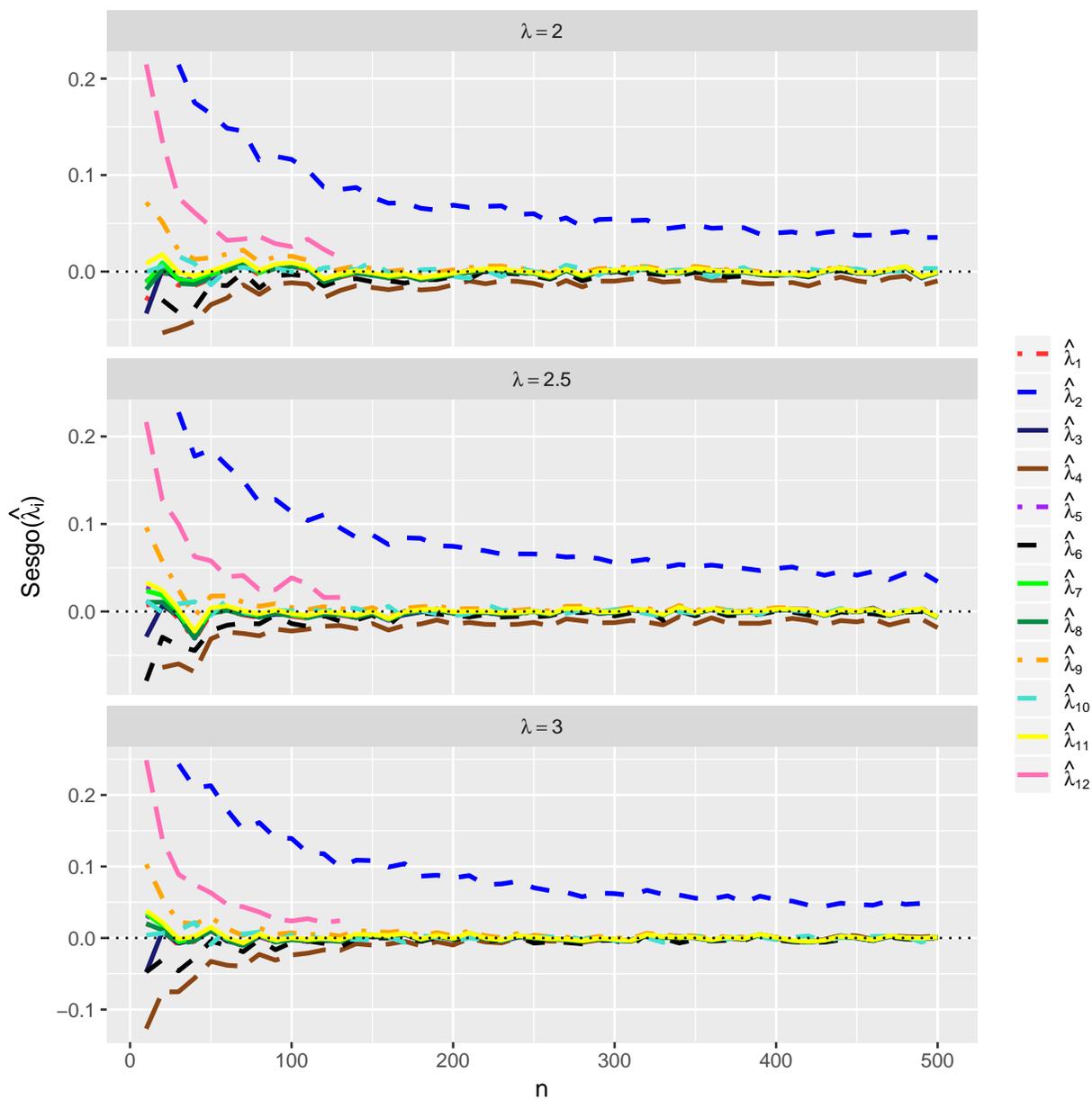


Figura 5-9: SP de los estimadores puntuales para diferentes valores de λ y $n(2)$

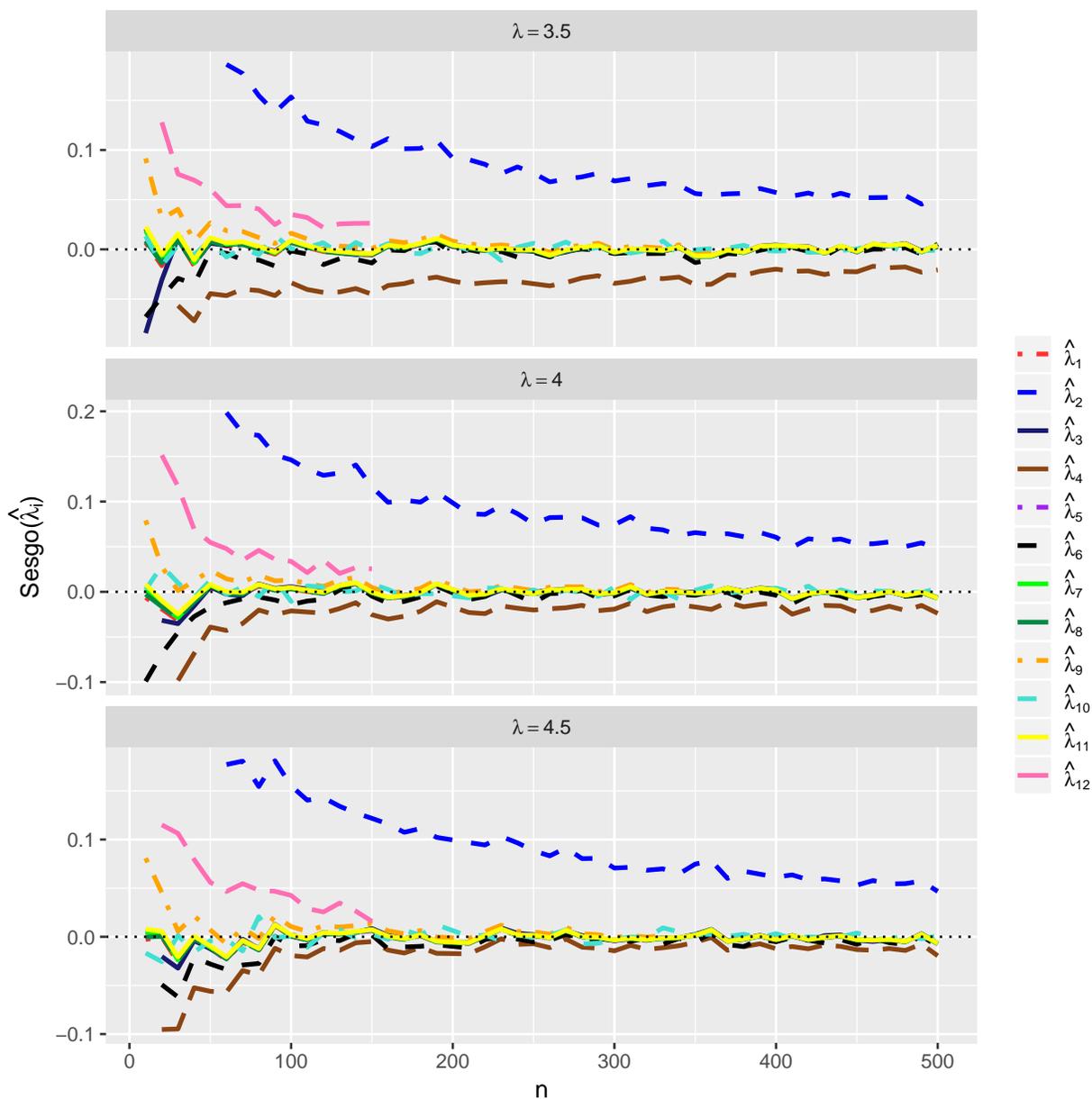


Figura 5-10: *SP* de los estimadores puntuales para diferentes valores de λ y $n(3)$

Las Figuras 5-11 a 5-13 muestran el *SP* de los estimadores puntuales de λ para diferentes valores de λ y n , considerando solo n menor a 250 y los estimadores con sesgo promedio más cerca de cero. Para estos estimadores no es posible discernir cual presenta menor sesgo promedio.

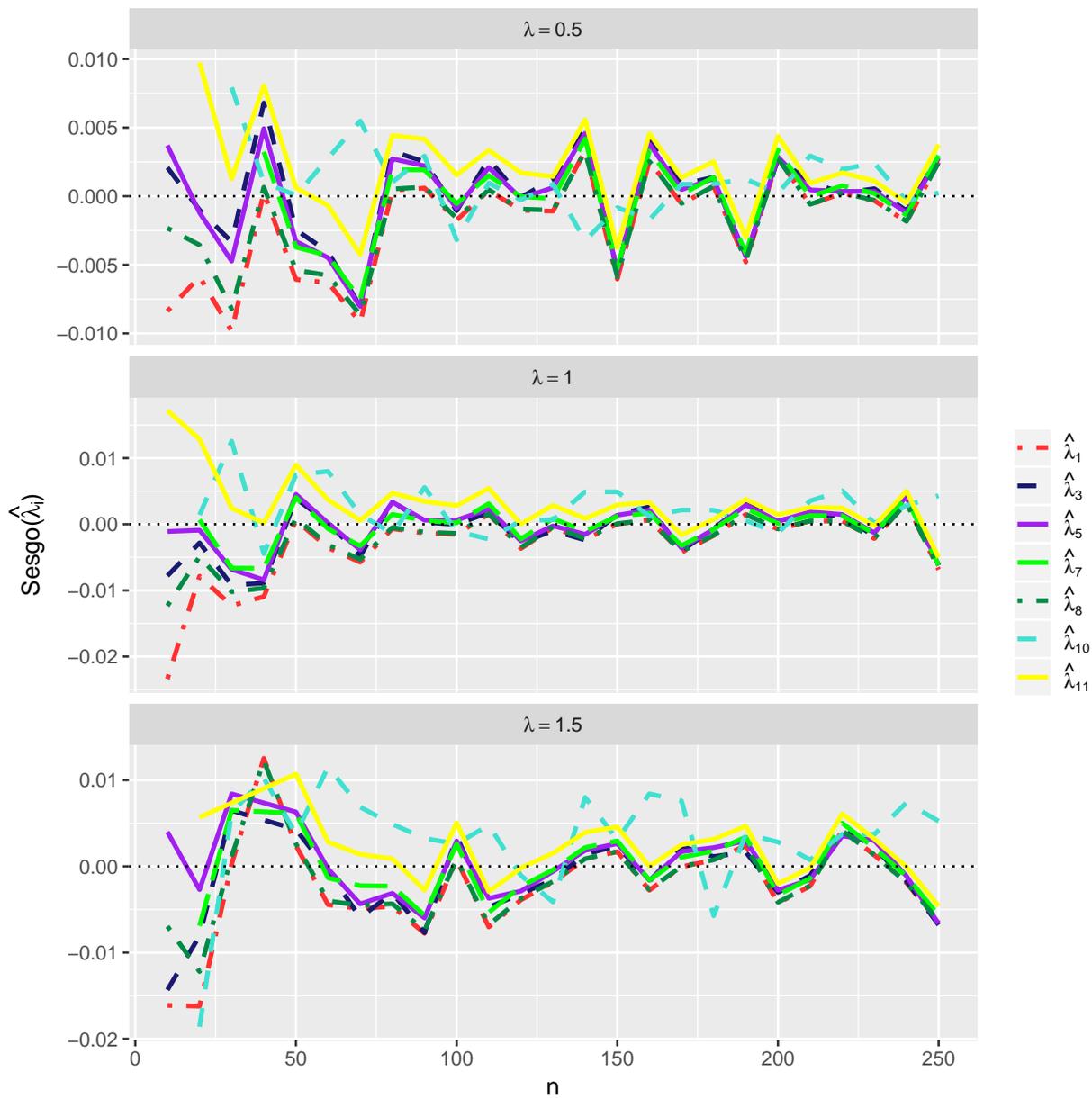


Figura 5-11: SP de los estimadores puntuales para diferentes valores de λ y n - restringida(1)

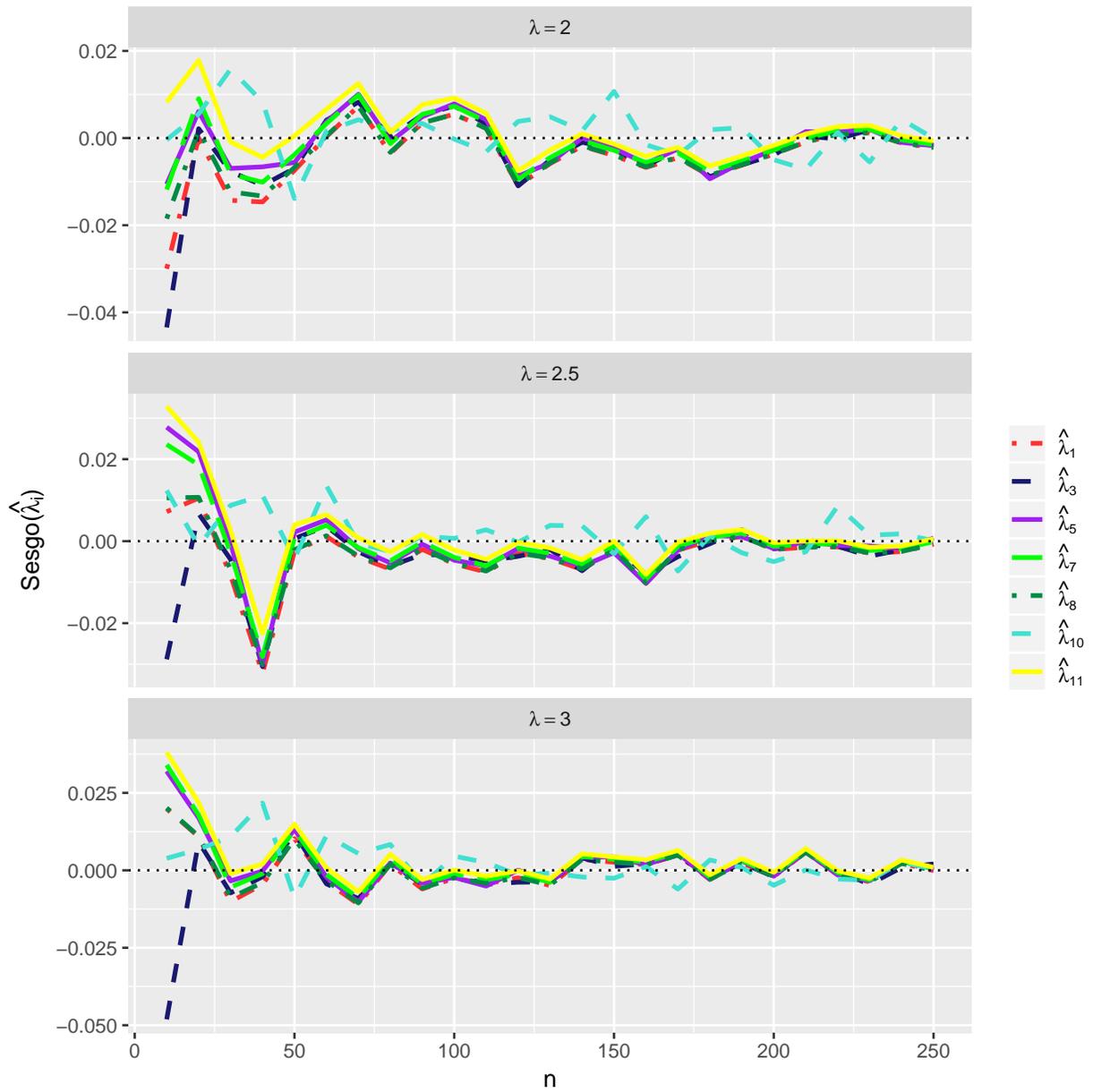


Figura 5-12: SP de los estimadores puntuales para diferentes valores de λ y n - restringida(2)

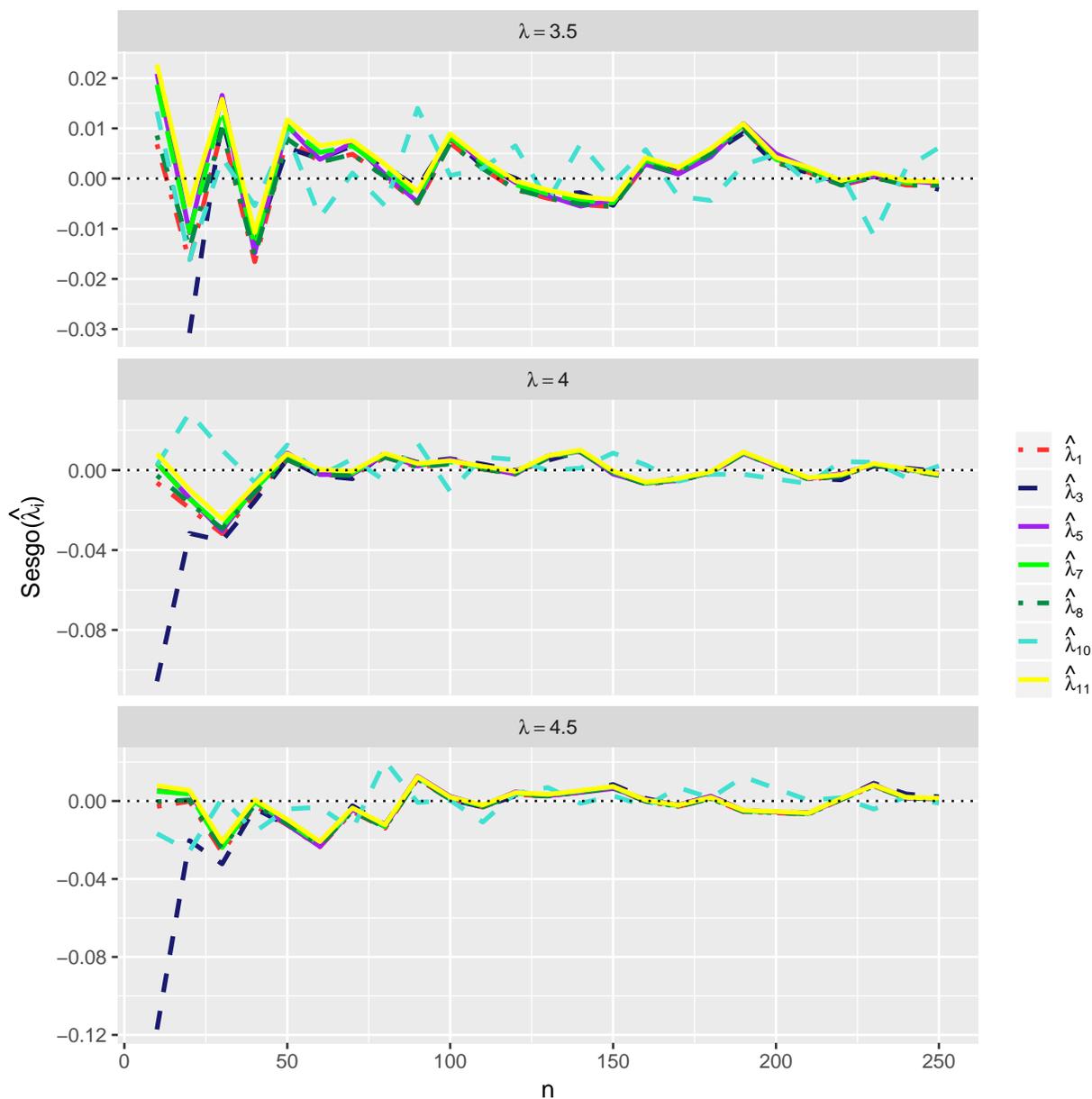


Figura 5-13: *SP* de los estimadores puntuales para diferentes valores de λ y n - restringida(3)

Estimadores puntuales para n_0

Las Figuras 5-14 a 5-16 muestran el *ECM* de los estimadores puntuales de n_0 para combinaciones de λ y N . Se destaca que para $\lambda \leq 3,5$ \hat{n}_{012} es el que tiene el menor *ECM* para N pequeño. No obstante a partir de cierto valor de N el *ECM* comienza a crecer sobrepasando el de los demás estimadores. Esto ocurre debido a problemas computacionales en el cálculo de \hat{n}_{012} . Otra observación importante es que a medida que N crece el *ECM* de todos los estimadores también lo hace, puesto que N sea grande no garantiza que n que es la muestra

conocida también lo sea.

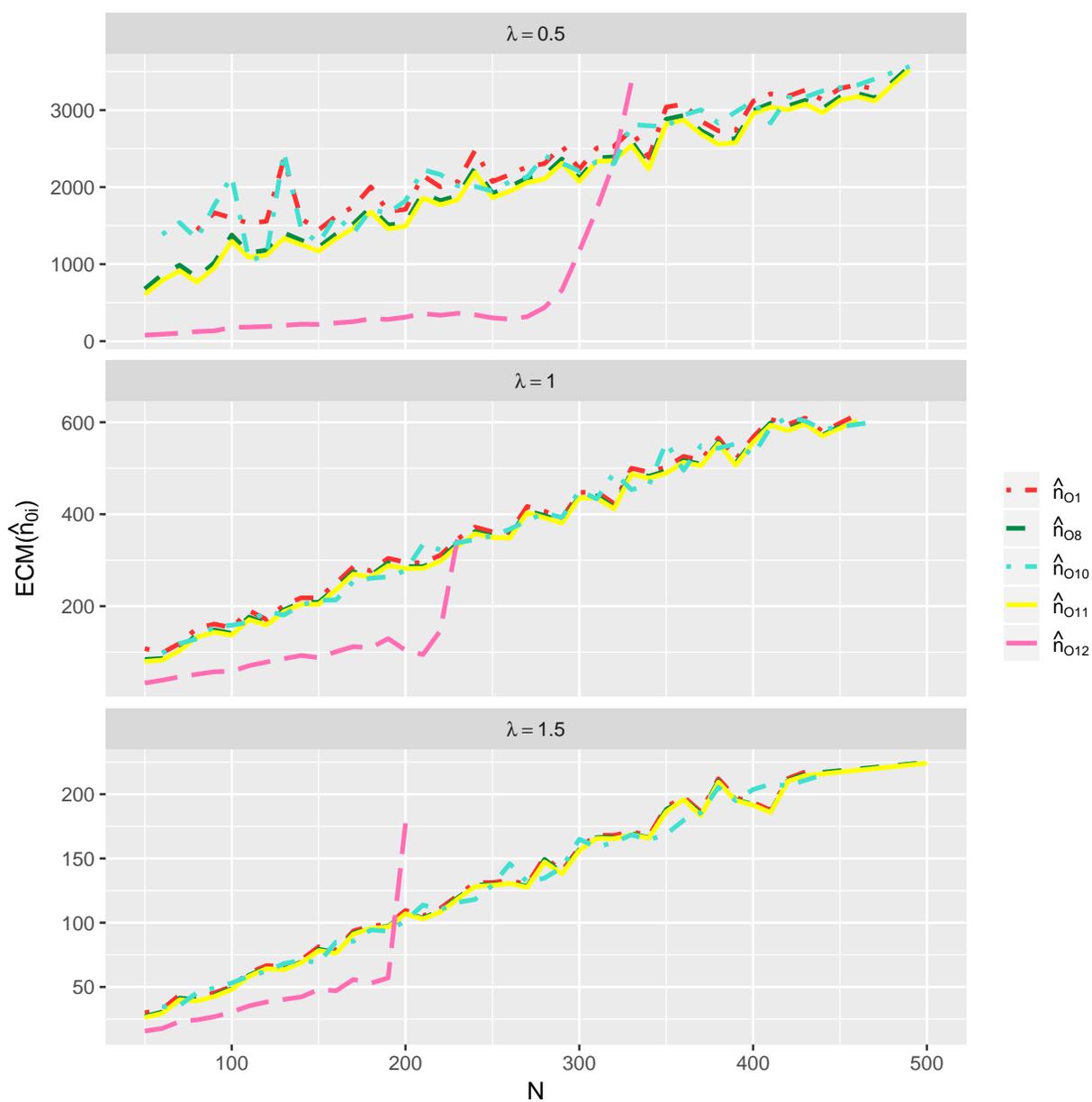


Figura 5-14: ECM de los estimadores puntuales de n_0 para combinaciones de λ y $N(1)$.

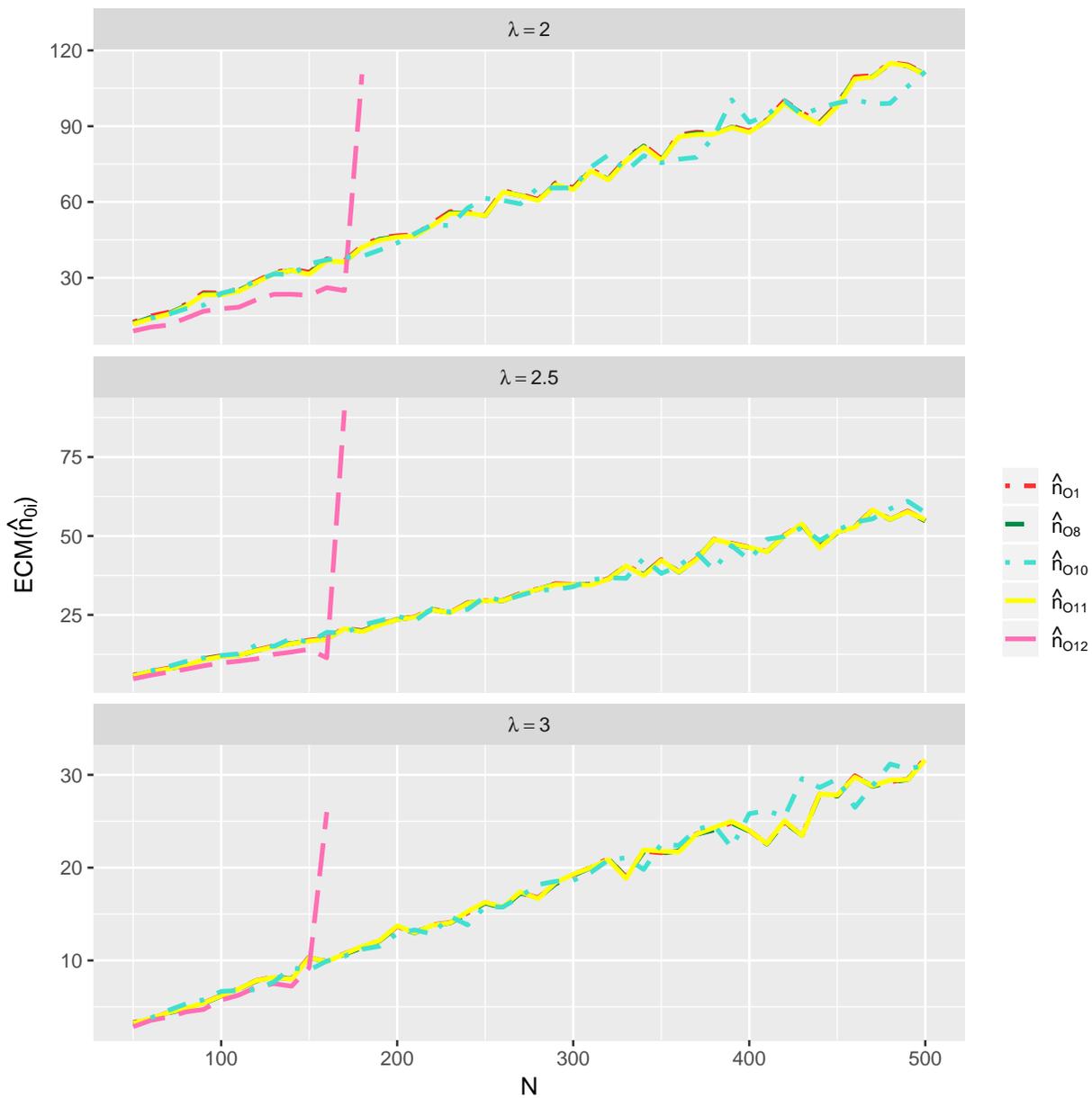


Figura 5-15: ECM de los estimadores puntuales de n_0 para combinaciones de λ y $N(2)$.

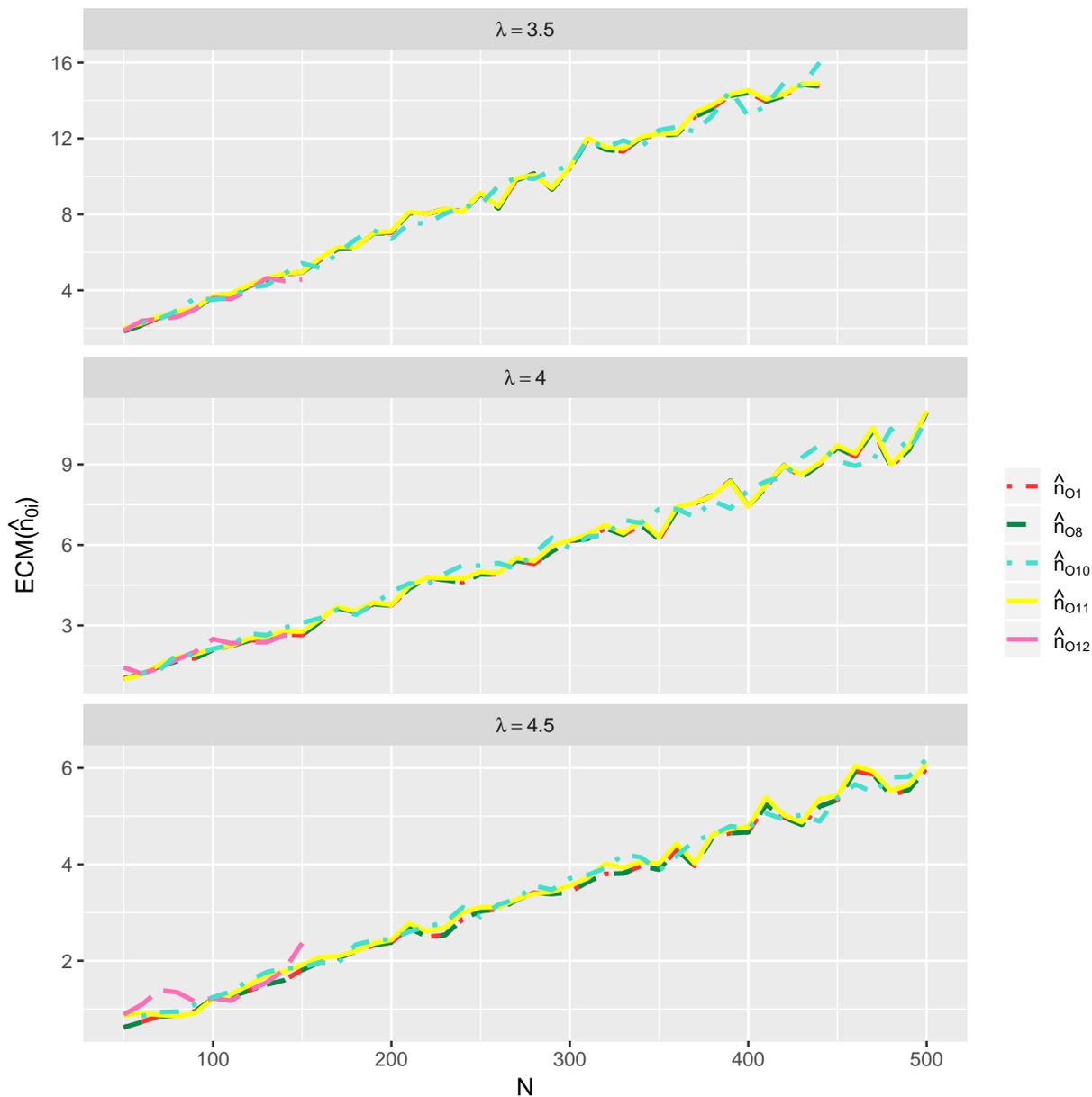


Figura 5-16: *ECM* de los estimadores puntuales de n_0 para combinaciones de λ y $N(3)$.

Con el objetivo de observar mejor el comportamiento de los estimadores se obtienen las Figuras 5-17 a 5-19. En estas figuras solo se considera $N < 200$. Para $\lambda < 3,5$, se tiene que los estimadores con menor *ECM* son \hat{n}_{011} seguido de \hat{n}_{08} . Para $\lambda \geq 3,5$ \hat{n}_{01} y \hat{n}_{08} son los que tienen menor *ECM*.

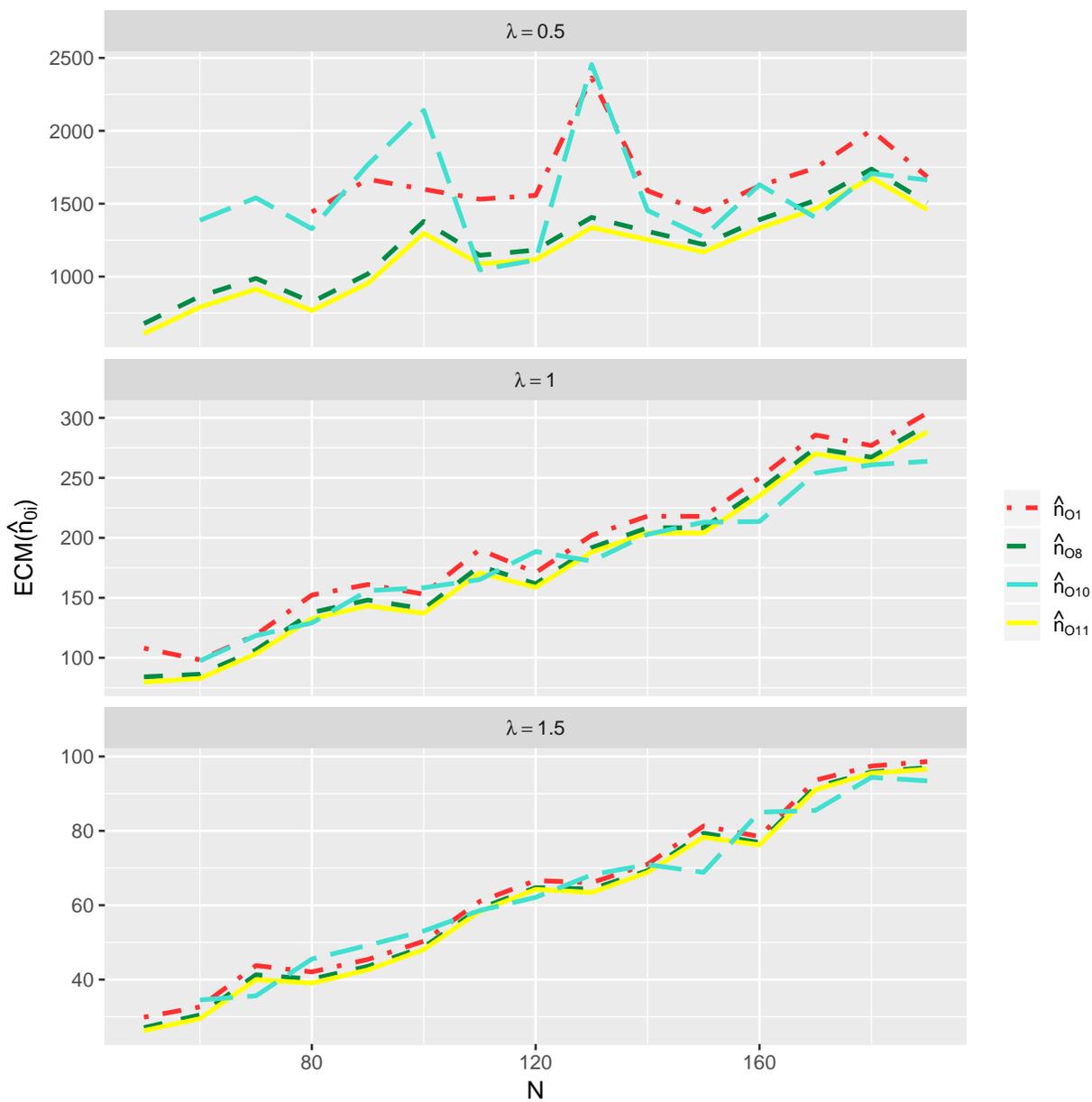


Figura 5-17: *ECM* de los estimadores puntuales de n_0 para combinaciones de λ y N . - restringida(1)

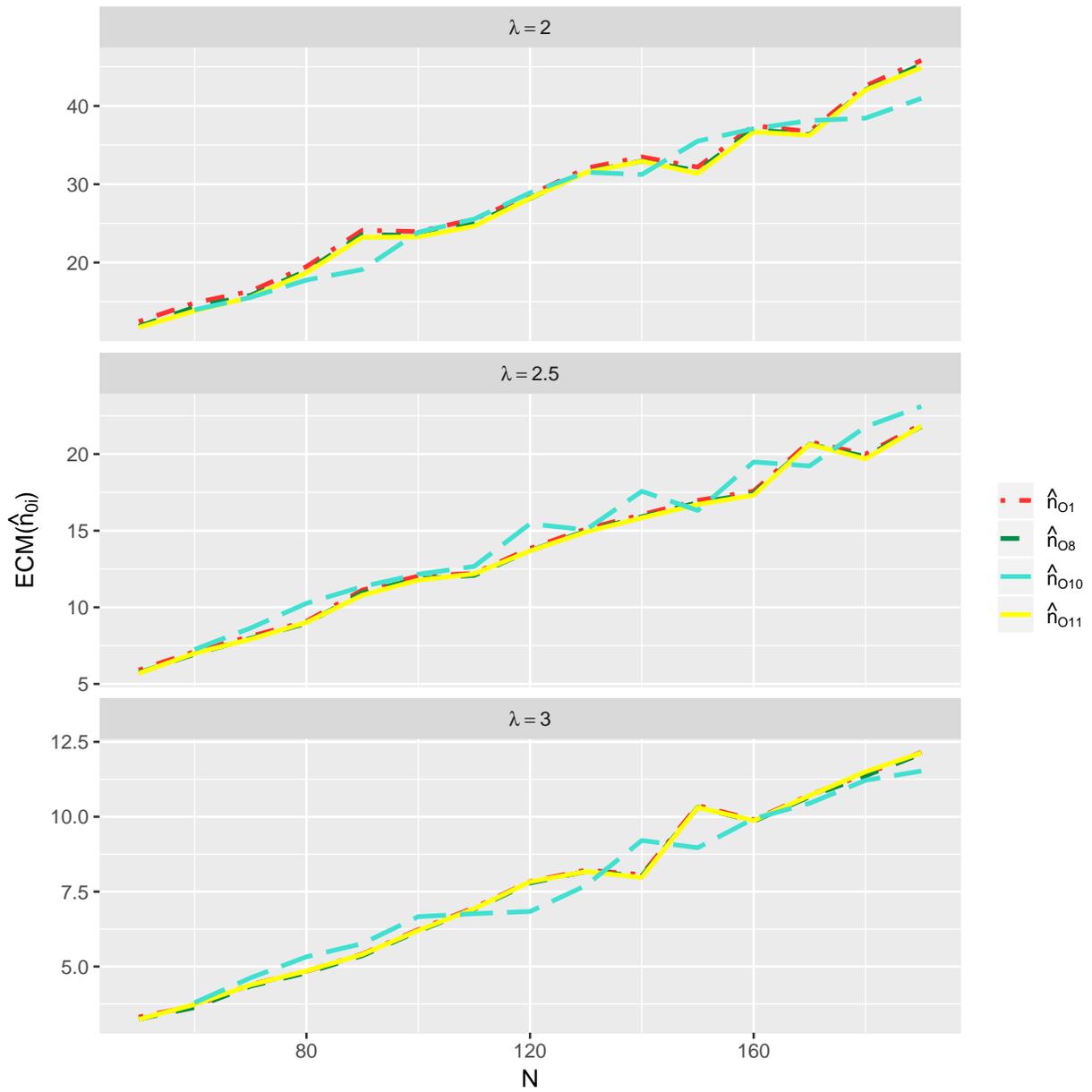


Figura 5-18: *ECM* de los estimadores puntuales de n_0 para combinaciones de λ y N . - restringida(2)

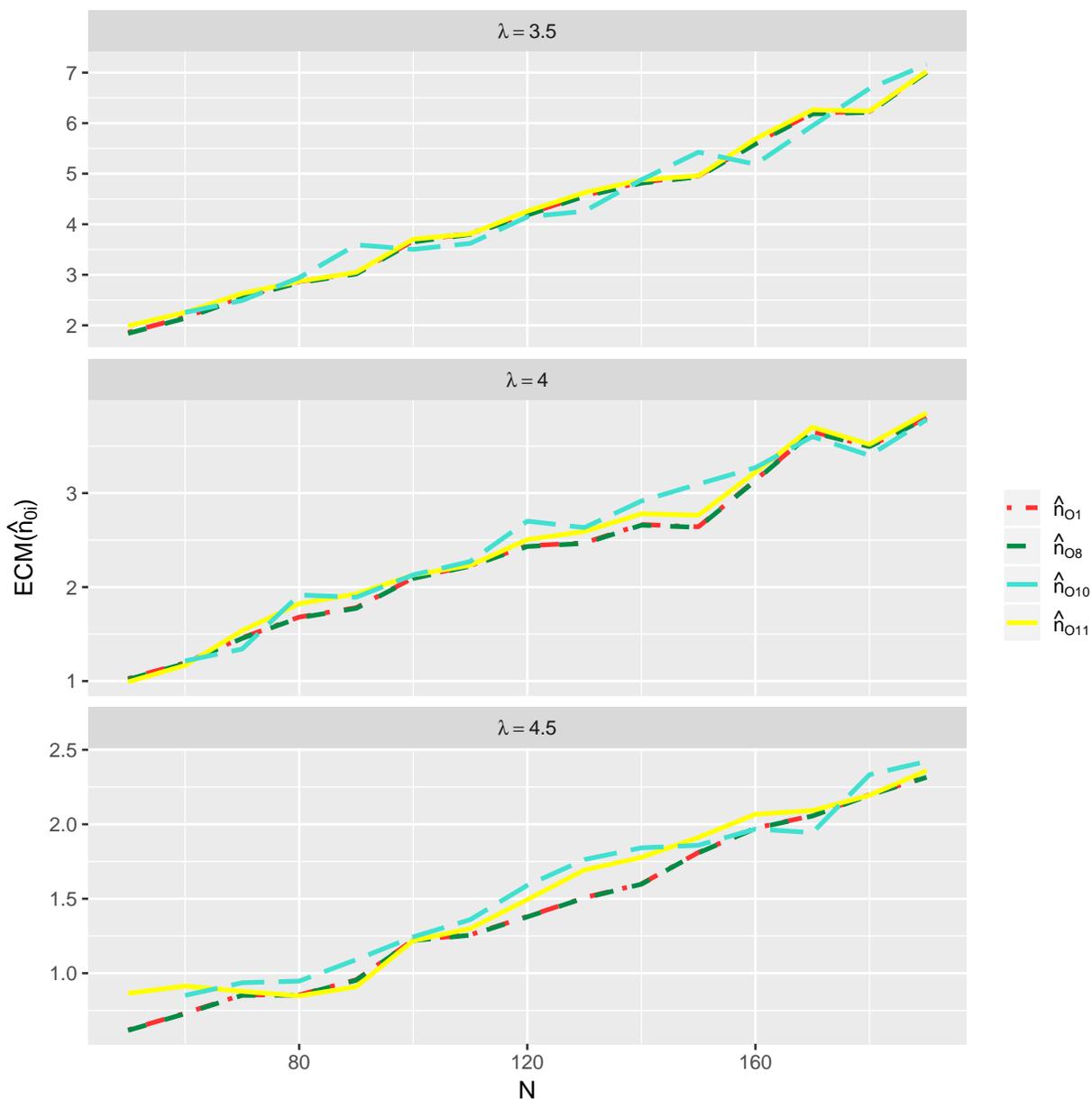


Figura 5-19: *ECM* de los estimadores puntuales de n_0 para combinaciones de λ y N . - restringida(3)

Las Figuras 5-20 a 5-21 muestran el *SP* de los estimadores puntuales de n_0 para diferentes valores de λ y N . Para $\lambda = 0,5$, \hat{n}_{012} es el estimador con sesgo más cerca del cero. No obstante, para los demás valores de λ no es así. Adicionalmente el sesgo promedio para \hat{n}_{012} tiende a crecer a partir de cierto valor de N .

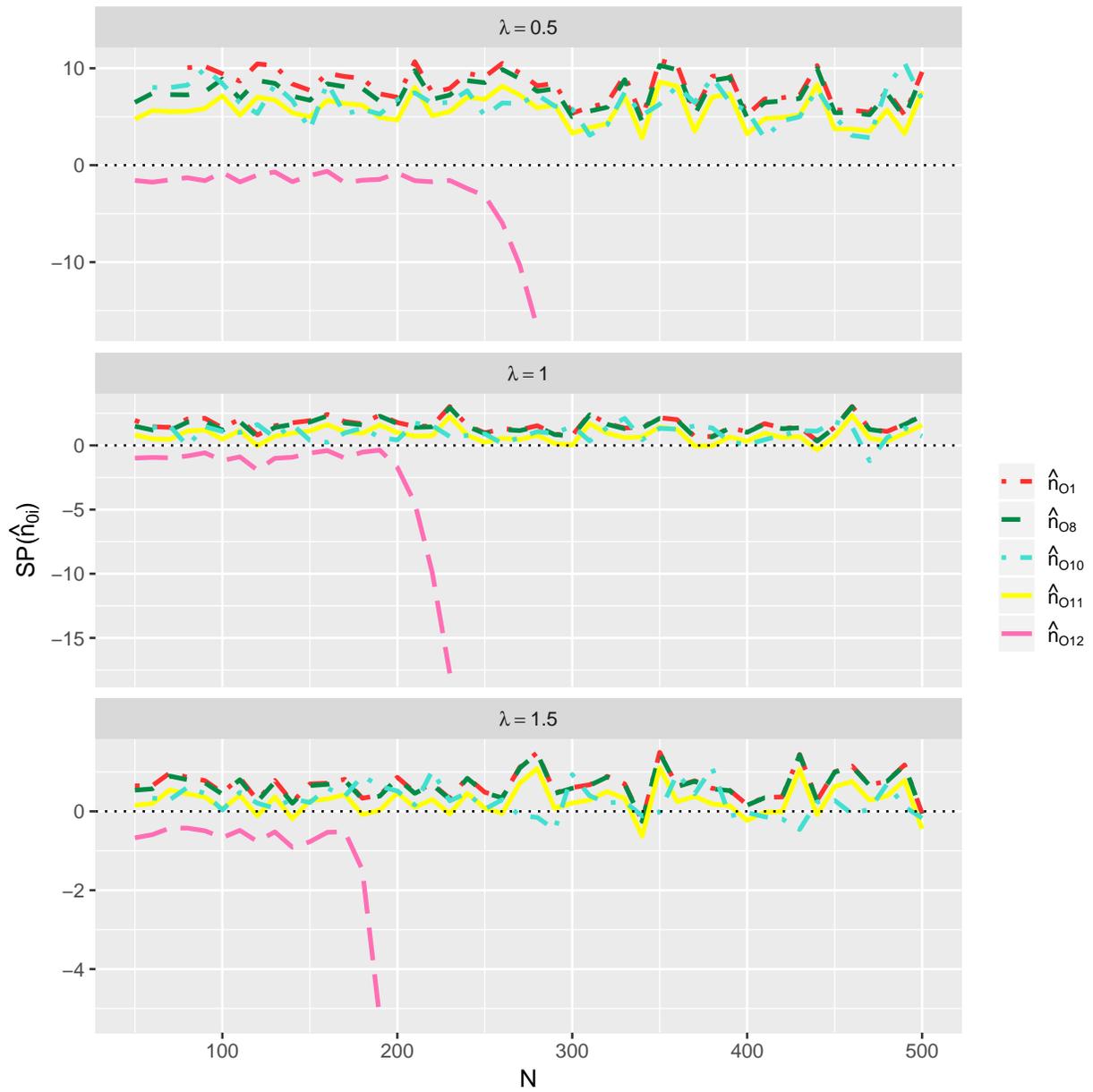


Figura 5-20: SP de los estimadores puntuales para diferentes valores de λ y $n(1)$

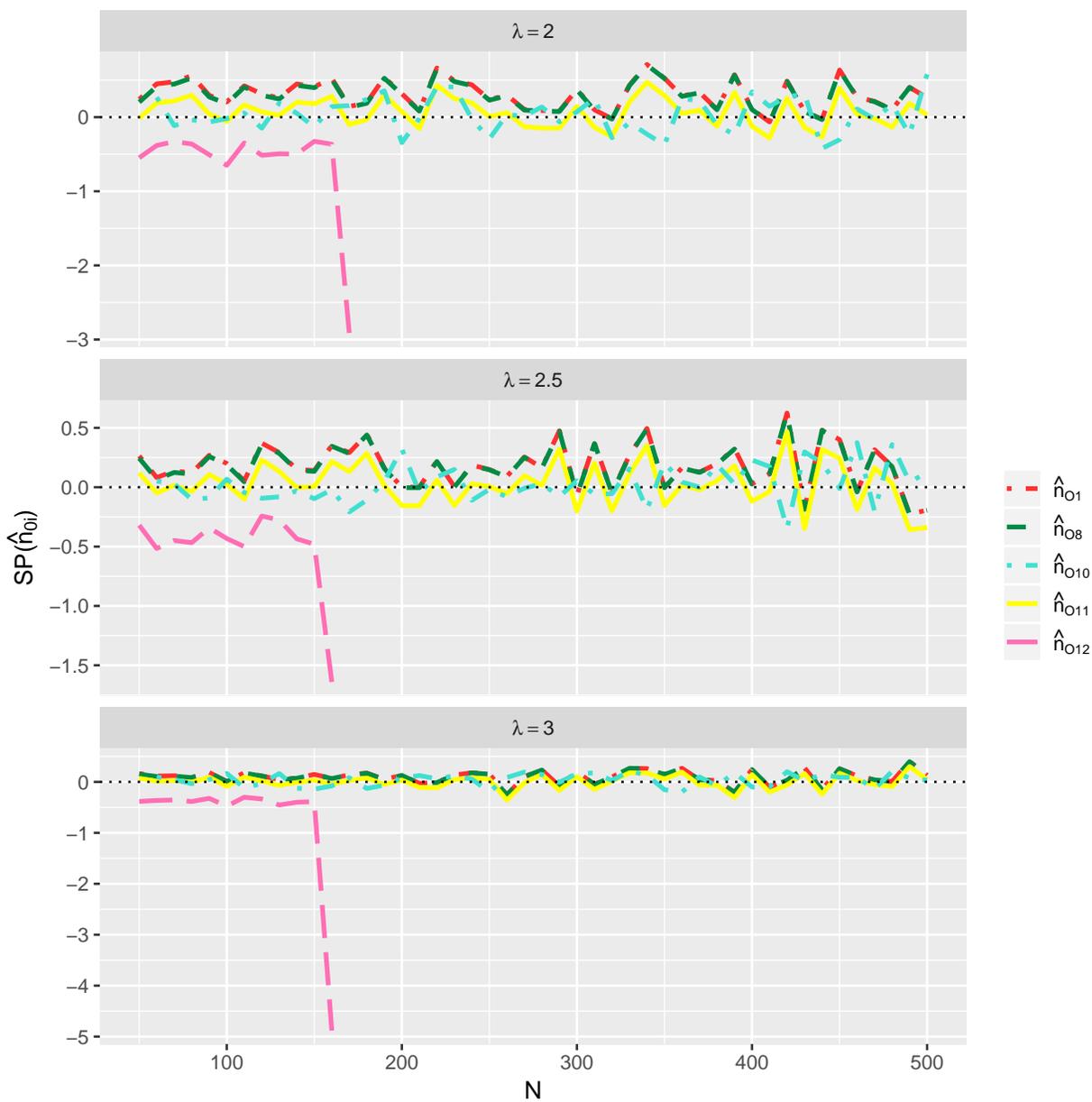


Figura 5-21: SP de los estimadores puntuales para diferentes valores de λ y $n(2)$

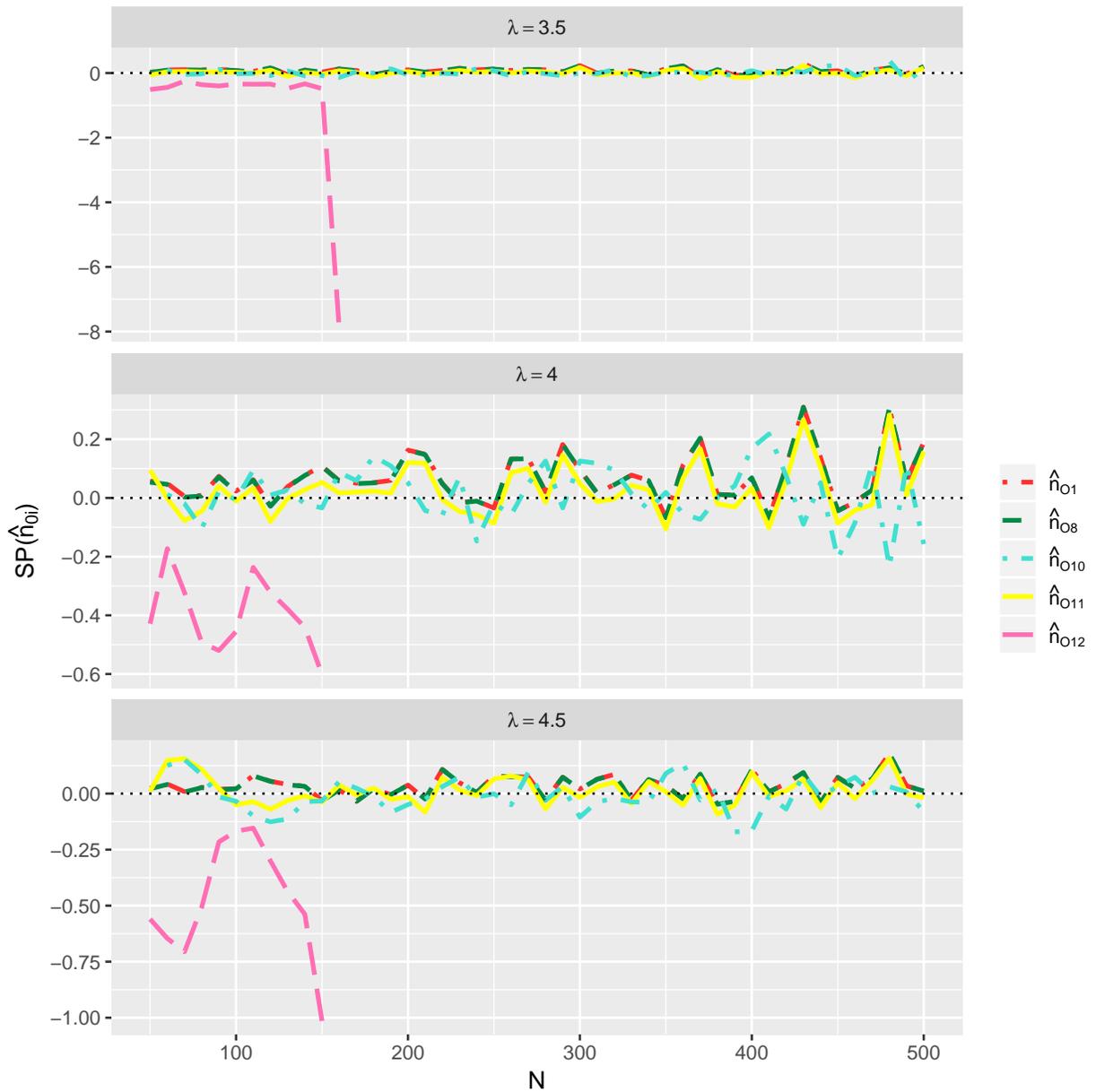


Figura 5-22: SP de los estimadores puntuales para diferentes valores de λ y $n(3)$

Las Figuras 5-23 a 5-25 muestran el SP de los estimadores puntuales de n_0 para diferentes valores de λ y N , considerando solo $N < 200$. Para $\lambda \leq 3,5$, \hat{n}_{o10} y \hat{n}_{o11} son los estimadores con sesgo promedio más cerca de cero.

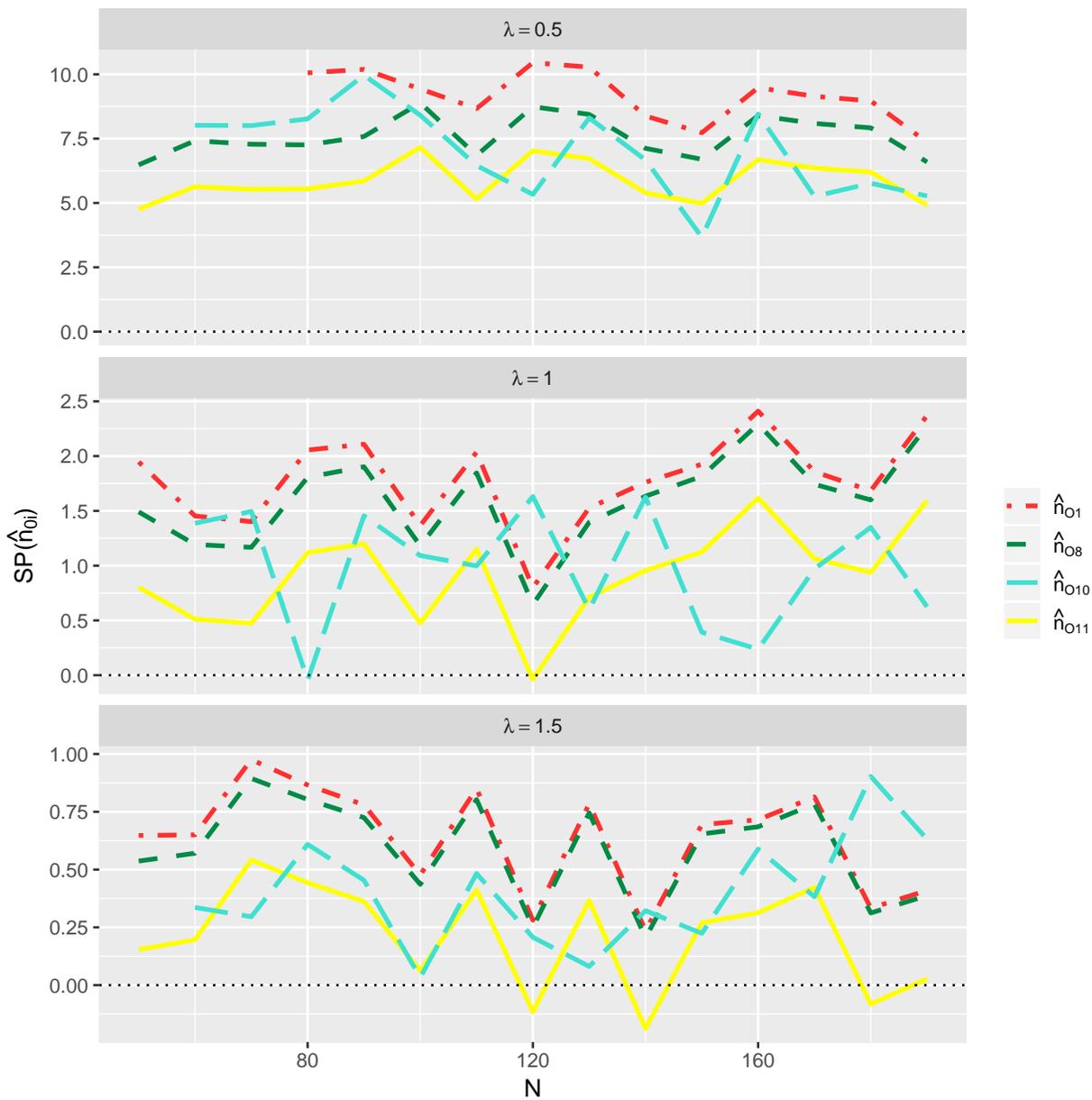


Figura 5-23: SP de los estimadores puntuales para diferentes valores de λ y n - restringida(1)

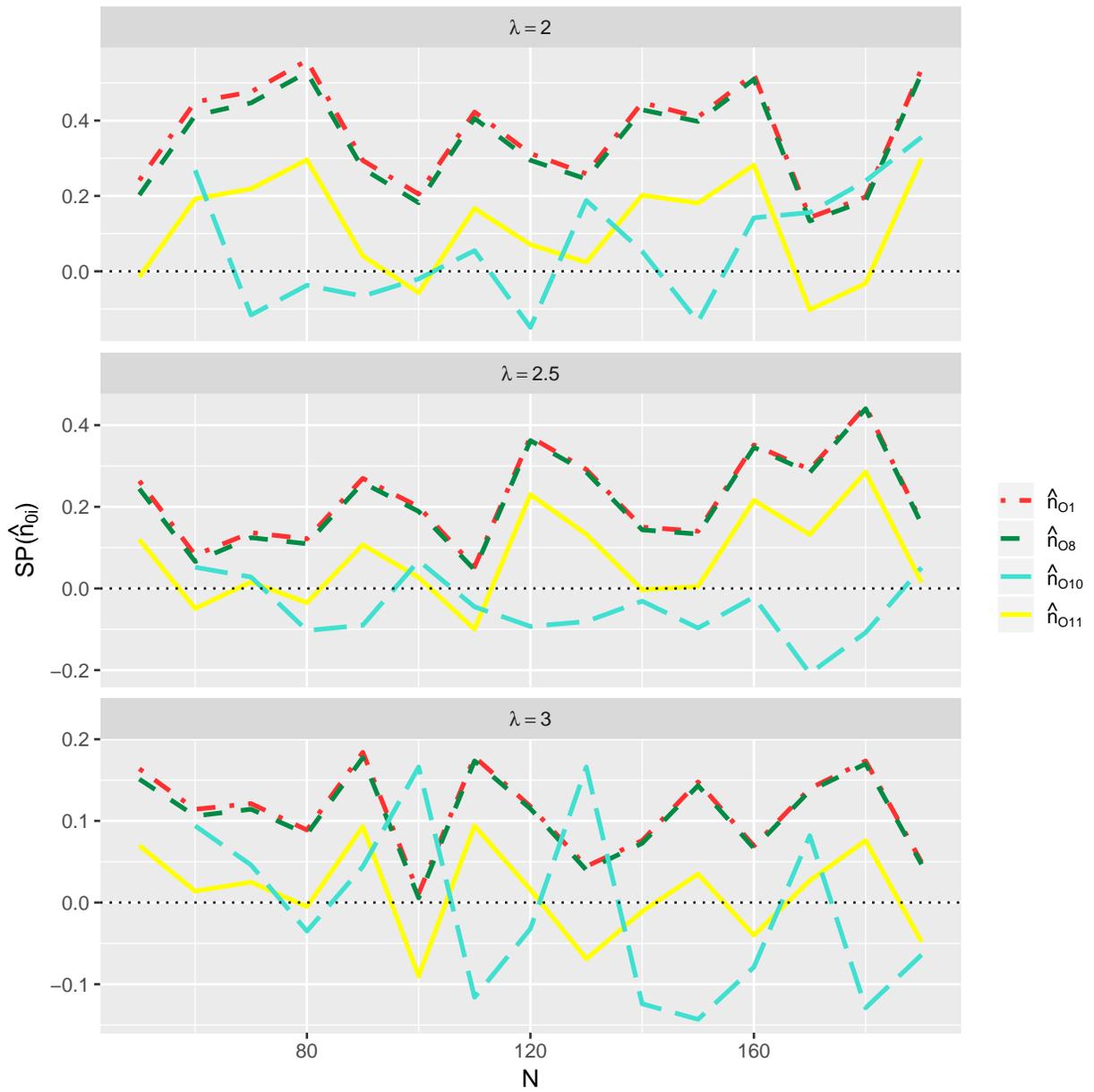


Figura 5-24: SP de los estimadores puntuales para diferentes valores de λ y n - restringida(2)

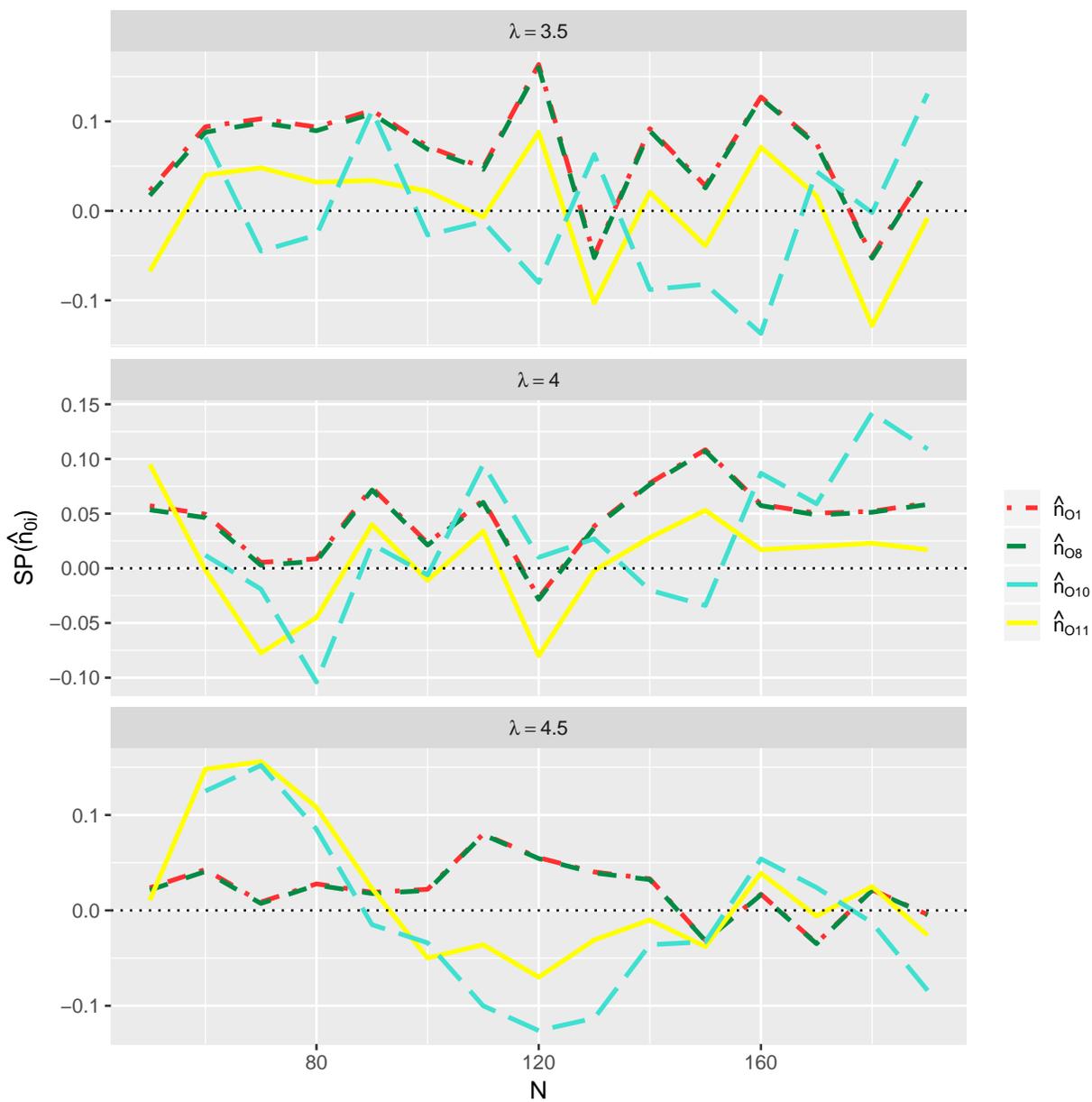


Figura 5-25: SP de los estimadores puntuales para diferentes valores de λ y n - restringida(3)

5.2. Comparación de los intervalos de confianza

La comparación de los intervalos de confianza se realiza mediante las medidas de longitud media del intervalo y probabilidad de cobertura. Se prefieren aquellos intervalos con longitud media pequeña y probabilidad de cobertura alta. La longitud media del intervalo se obtiene a partir de la ecuación:

$$\bar{L} = \bar{L}_S - \bar{L}_I \quad (5-5)$$

donde:

\bar{L} : Longitud media del intervalo.

\bar{L}_I : Límite inferior promedio.

\bar{L}_S : Límite superior promedio

Por otro lado, la Probabilidad de Cobertura (PC) se define como el porcentaje de veces que el verdadero valor del parámetro cae dentro del intervalo de confianza:

$$PC = \frac{C}{NC} \quad (5-6)$$

donde:

C : Número de intervalos que contienen el verdadero valor del parámetro.

NC : Número de intervalos construidos.

En la comparación se consideran tamaños de muestra n iguales a 10, 20, 30, . . . , 500 y valores del parámetro $\lambda = 0.5, 1, 1.5, \dots, 4.5$. En cuanto al intervalo bayesiano se asume que el valor de $\beta = 2$ y que $\alpha = 2 * \lambda$, es decir el valor esperado de la distribución a priori es igual al verdadero valor del parámetro. Dado que IC_6 presenta problemas computacionales en su calculo para el tamaños de muestra grandes, este intervalo solo se calcula hasta $n = 150$.

5.2.1. Resultados

La Figura 5-26 a 5-28 muestran la longitud de los intervalos de confianza para combinaciones de λ y n . Dos hechos son de destacar en esta figura. Primero, para los diferentes valores de λ y n se observa que los intervalos IC_4 e IC_3 son los de mayor longitud, seguidos del IC_6 . Estas diferencias en la longitud disminuyen a medida que incrementa el valor del parámetro λ . Segundo, cuando $\lambda = 0,5$ el IC_2 es el que presenta menor longitud. Sin embargo a partir de $\lambda = 1$ es el IC_1 el que comienza a mostrar las menores longitudes.

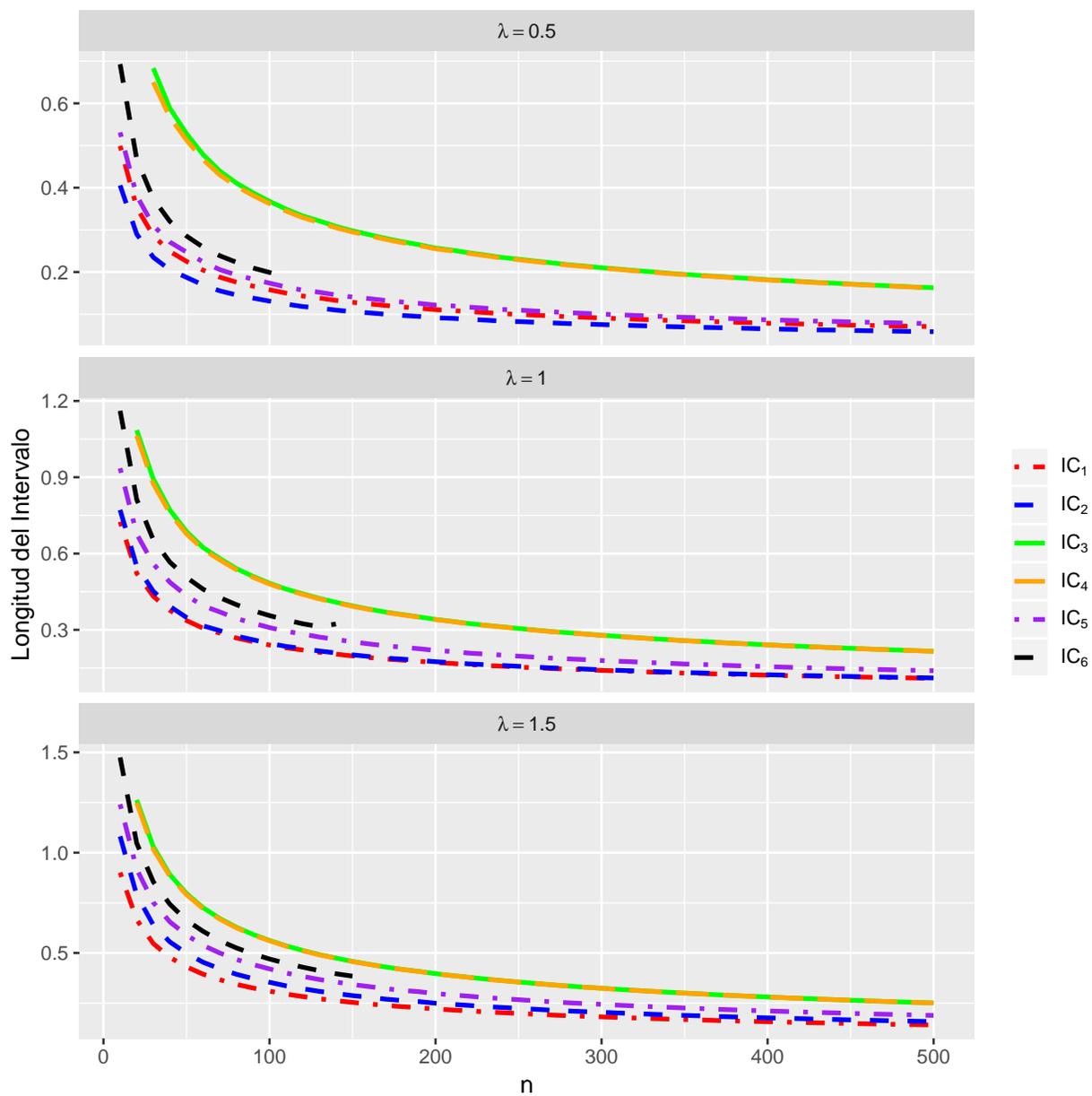


Figura 5-26: Longitud de los intervalos de confianza para combinaciones de λ y $n(1)$

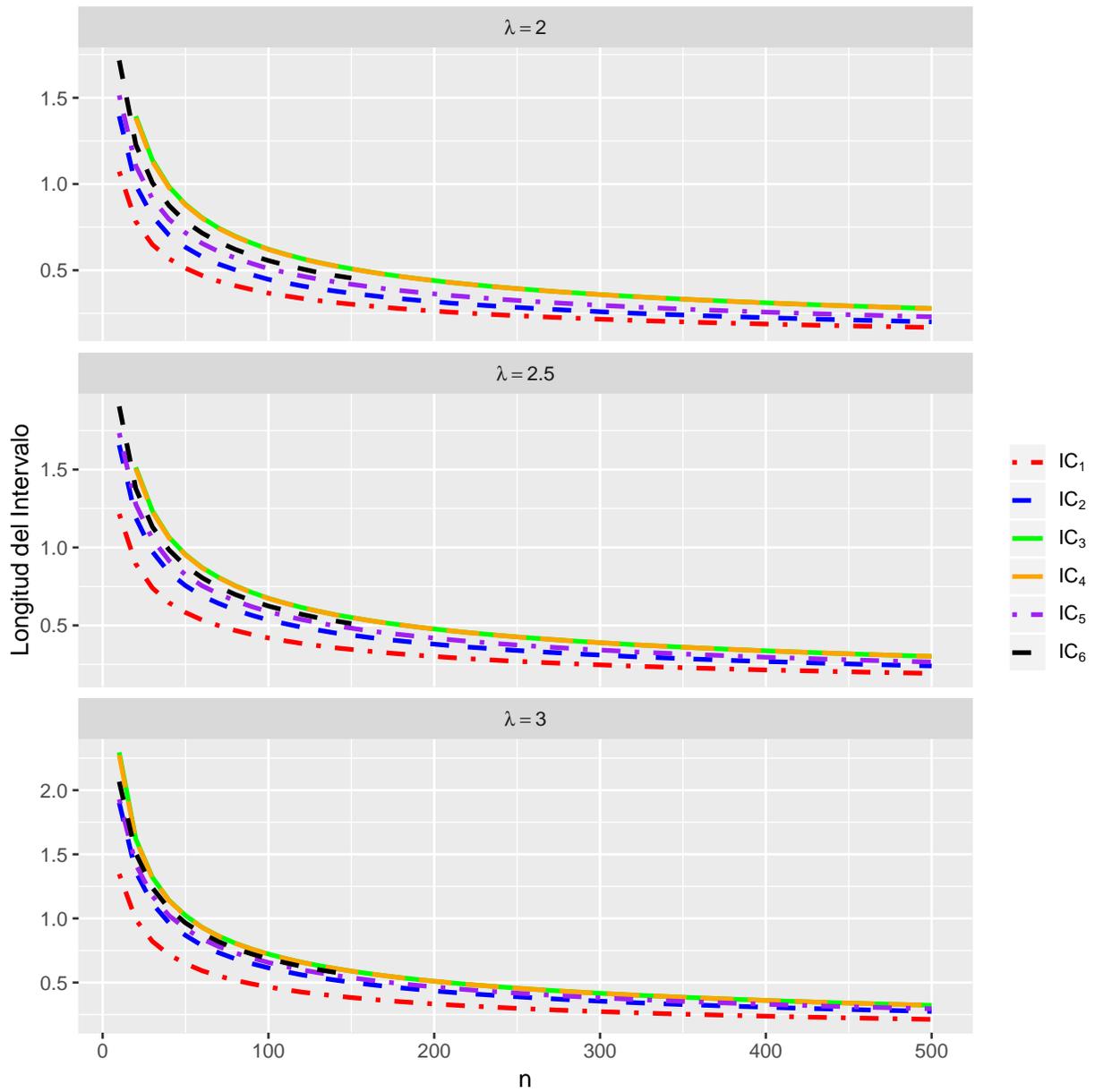


Figura 5-27: Longitud de los intervalos de confianza para combinaciones de λ y $n(2)$

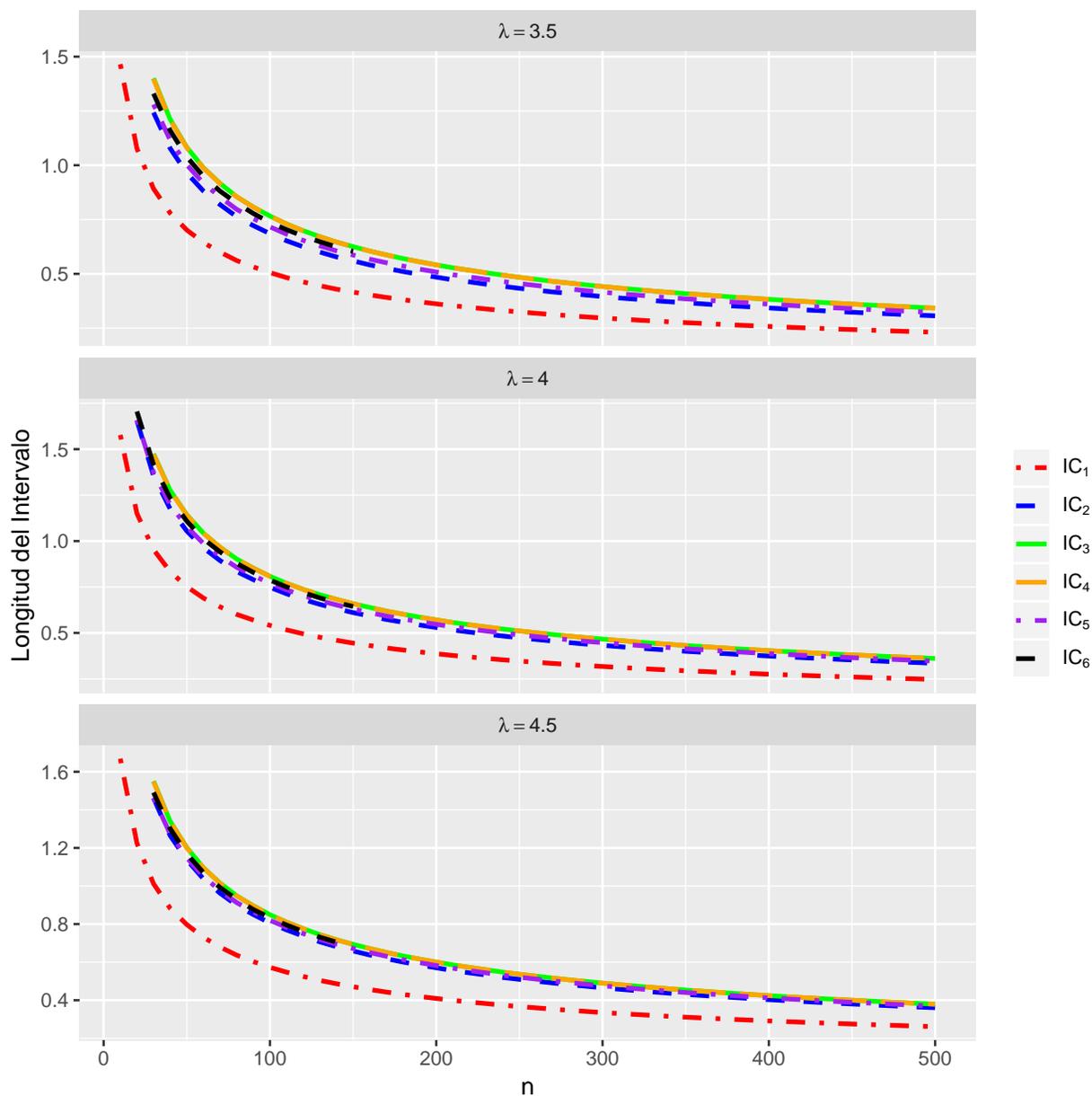


Figura 5-28: Longitud de los intervalos de confianza para combinaciones de λ y $n(3)$

Las Figuras 5-29 a 5-31 muestran la probabilidad de cobertura de los intervalos de confianza para combinaciones de λ y n . Se destaca que el intervalo IC_6 para la mayoría de las combinaciones de λ y n fue el que obtuvo la mayor probabilidad de cobertura. Aparte de este intervalo, solo IC_3 e IC_4 mantienen una probabilidad de cobertura oscilando cerca a el nivel nominal. Adicionalmente, se observa que la probabilidad de cobertura para los intervalos de confianza incrementan cuando el valor del parámetro λ se hace más grande.

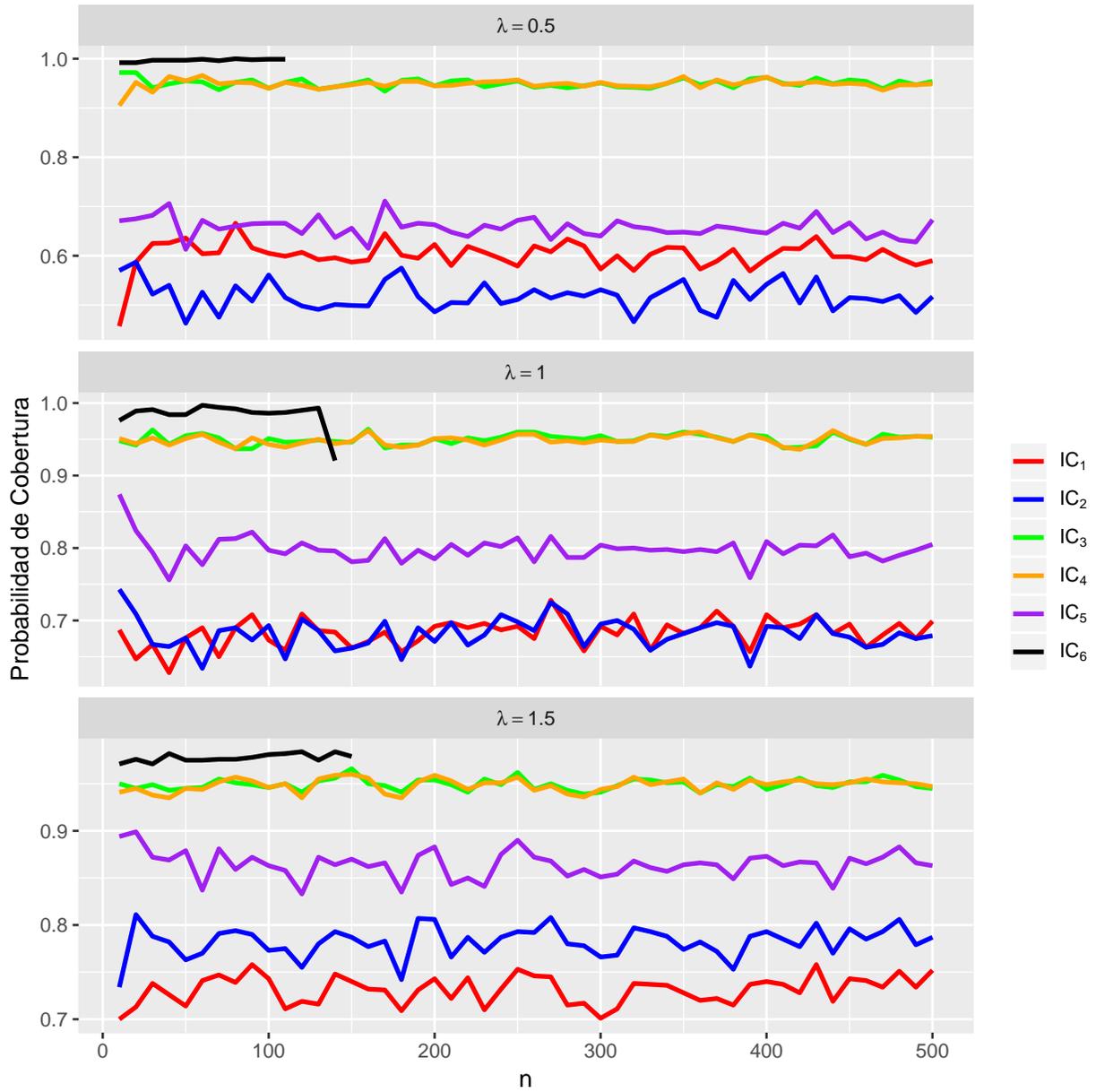


Figura 5-29: Probabilidad de cobertura de los intervalos de confianza para combinaciones de λ y n

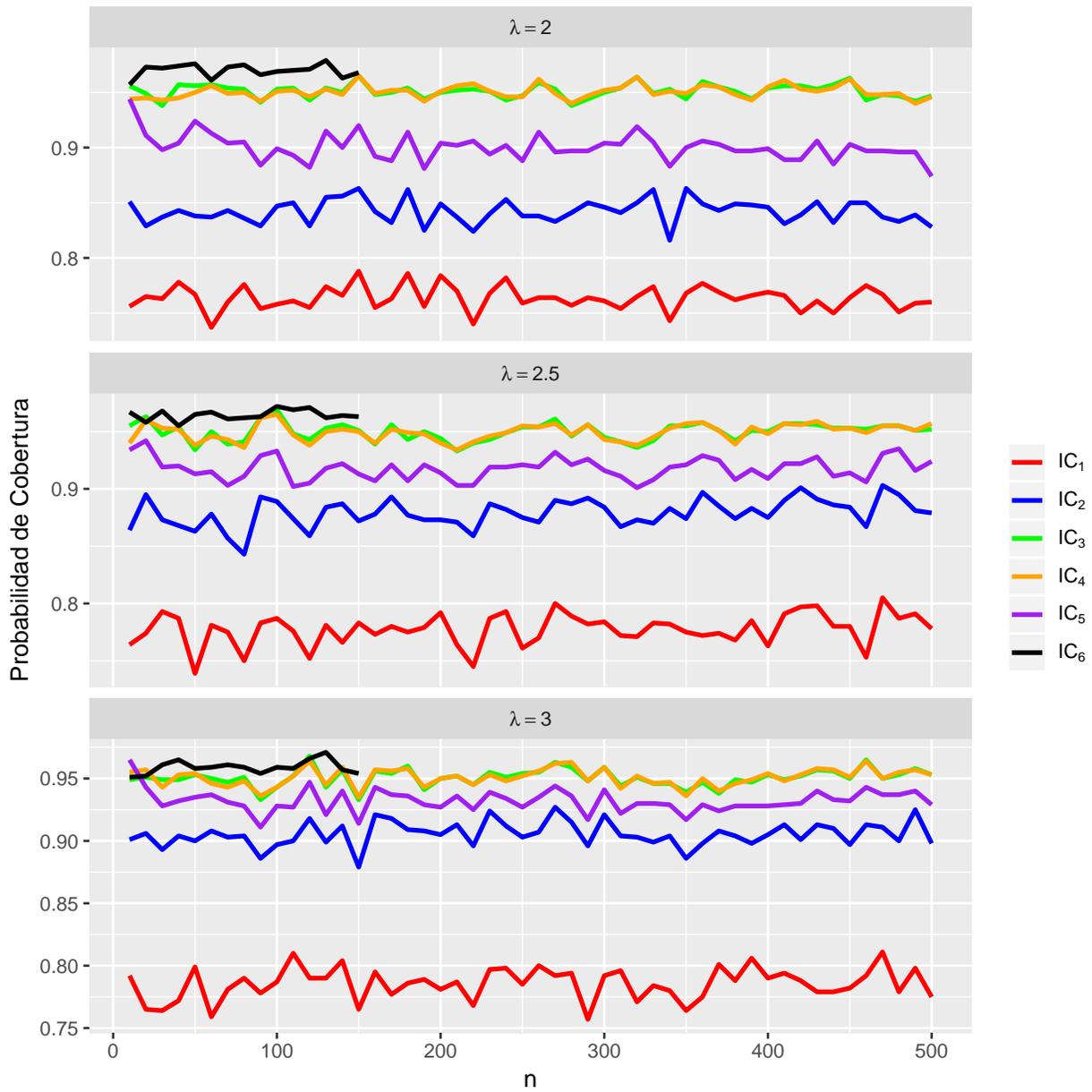


Figura 5-30: Probabilidad de cobertura de los intervalos de confianza para combinaciones de λ y n

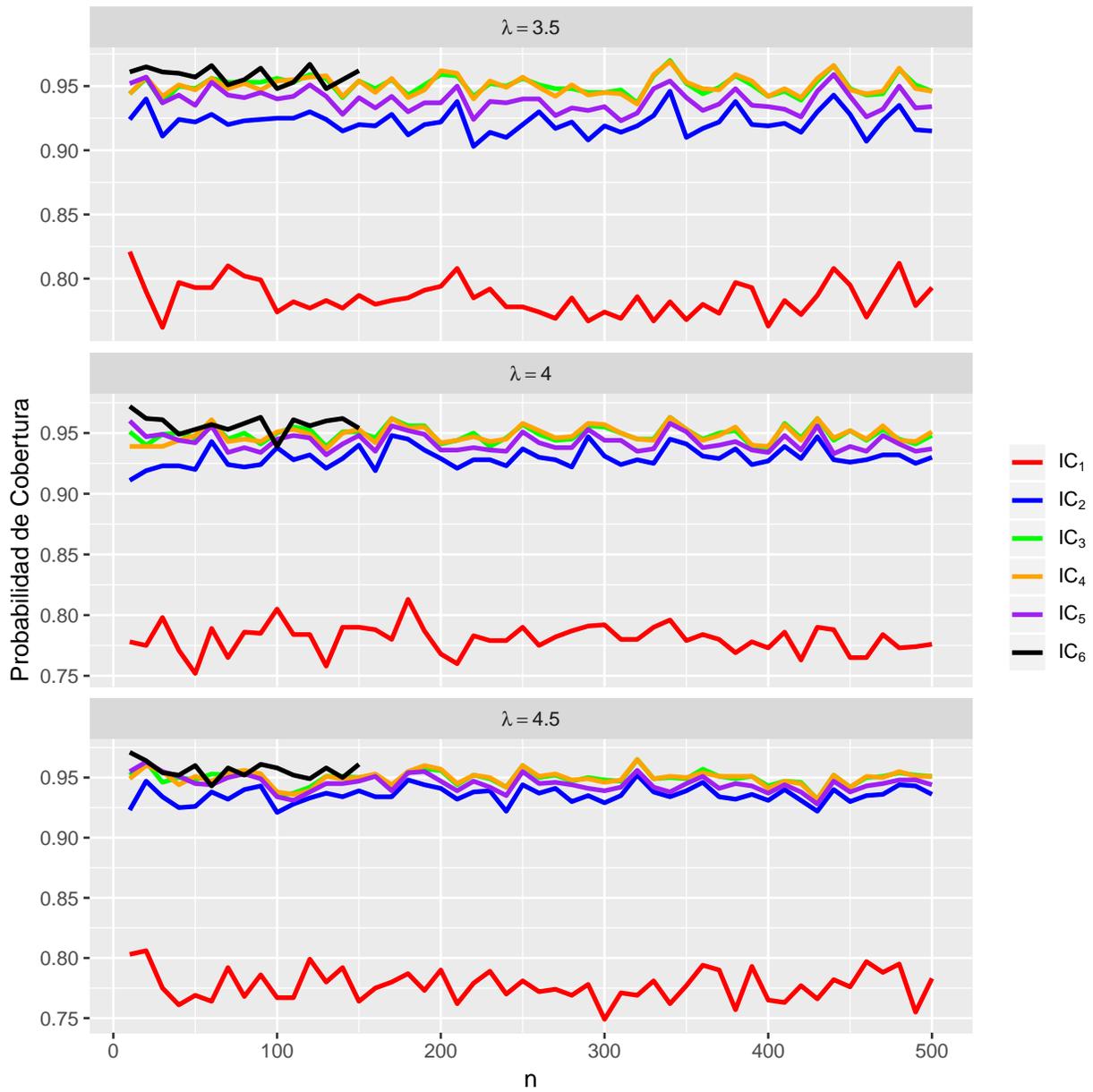


Figura 5-31: Probabilidad de cobertura de los intervalos de confianza para combinaciones de λ y n

6 Aplicaciones

Como se mencionó en la introducción la DPTC tiene múltiples aplicaciones. En esta sección se consideran cuatro de estas aplicaciones y se obtienen los mejores estimadores opresentados en el texto. La primera aplicación consiste en estimar el número de unidades de proteína presentes en las células, en la segunda el número de adictos de marihuana en 2006 en Tailandía, en la tercera el número de trabajadores en una empresa, y en la cuarta el número total de hogares. En la mayoría de los casos no se pudo calcular el estimador propuesto por Vélez and Correa (2013) debido al tamaño de muestra.

6.1. Estimación del número de unidades de proteína presentes en las células

En un experimento se desea contar el número de unidades (epítopes) de proteína en las células. Para lograr este objetivo es necesario etiquetar los epítopes, pues solo se identifican aquellas unidades con una o más etiquetas. Para poder contabilizar el número total de epítopes se debe estimar el número de unidades que no tienen etiquetas. En esta estimación se asume que los datos siguen una distribución Poisson, tal y como lo hicieron Mattews and Appleton (1993). La tabla **6-1** muestra la frecuencia con que se encuentran i etiquetas en un epítope de proteína(Cullen et al., 1990).

i	1	2	3	4	5
n_i	122	50	18	4	4

Tabla 6-1: Frecuencia con que se encuentran i etiquetas en un epítope de proteína(Cullen et al., 1990)

La tabla **6-2** presenta las estimaciones del promedio de etiquetas en un epítope (λ) y el número de epítopes sin etiquetas (n_0). Para encontrar los estimadores bayesianos se asumió que el verdadero valor de λ es 1. Se encontró que el estimador más alto para λ fue $\hat{\lambda}_1 = 0.9929029$ y el más bajo $\hat{\lambda}_0 = 0.9905853$. Con respecto a los estimadores de n_0 se encontró que el más alto fue $\hat{n}_{08} = \hat{n}_{010} = 117$ y el más bajo fue $n_{011} = 116$.

j	$\hat{\lambda}_j$	\hat{n}_{0j}
8	0.9907275	117
10	0.9905853	117
11	0.9929029	116

Tabla 6-2: Estimaciones del promedio de etiquetas en un epítotope (λ) y el número de epítotope sin etiquetas (n_0)

La tabla **6-3** presenta estimaciones por intervalo para el promedio de etiquetas en un epítotope. En la tabla LI_j y LS_j denotan el límite inferior y límite superior del intervalo IC_j respectivamente. Se observa que el intervalo de menor longitud fue IC_3 .

j	LI_j	LS_j	Longitud
3	0.834030	1.176532	0.342502
4	0.828614	1.169917	0.341303

Tabla 6-3: Estimaciones por intervalo para el promedio de etiquetas en un epítotope

6.2. Estimación del número de adictos de marihuana en 2006 en Tailandia

Saber el número de usuarios que consumen cierta sustancia alucinógena es un indicador de la efectividad con que los gobiernos están afrontando los problemas del micro tráfico, salud pública, entre otros. Sin embargo, solo es observable el número de individuos que han entrado en un programa de rehabilitación para tratar una adicción, desconociendo el número de adictos que no han sido tratados. Realizar una predicción de esta cifra requiere conocer el número de adictos que no han sido tratados. La tabla **6-4** muestra la frecuencia con la cual un usuario de marihuana ha ingresado a tratamiento de adicción i veces (Tailandia-2006) (Vivatwongkasem, 2016):

Ingreso a tratamiento (i)	0	1	2	3	4	5
Número de casos (Frecuencias n_i)	-	5,445	1,025	158	21	1

Tabla 6-4: Frecuencia con la cual un usuario de marihuana ha ingresado a tratamiento de adicción i veces (Tailandia-2006)

La tabla **6-5** presenta las estimaciones del promedio de ingresos a tratamiento (λ) y número de usuarios de marihuana que no han ingresado a tratamiento (n_0) con los diferentes métodos presentados. Para encontrar los estimadores bayesianos se asumió que el verdadero valor de

λ es 0.4. La mayor estimación del promedio de ingresos a tratamiento fue $\hat{\lambda}_7 = 0,3972735$ y la menor $\hat{\lambda}_1 = 0,3972290$. A su vez la máxima cantidad estimada para el número de adictos que no han recibido tratamiento fue $\hat{n}_{010} = 13,636$ y la mínima $\hat{n}_{011} = 13,633$.

i	λ_i	n_i
8	0.3972305	13,635
10	0.3972290	13,636
11	0.3972735	13,633

Tabla 6-5: Estimaciones del promedio de ingresos a tratamiento (λ) y número de usuarios de marihuana que no han ingresado a tratamiento (n_0)

La tabla **6-6** muestra los intervalos de confianza para el promedio de ingresos a tratamiento (λ) con los diferentes métodos presentados. Se observa que el intervalo con menor longitud fue el IC_4 . Debido al tamaño de muestra IC_6 no pudo ser calculado.

j	LI_j	LS_j	
3	0.3775833	0.4178935	0.0403102
4	0.3774407	0.4176587	0.040218

Tabla 6-6: Intervalos de confianza para el promedio de ingresos a tratamiento (λ)

6.3. Estimación del número de trabajadores en una empresa

Conocer el número total de empleados de una empresa y el número de accidentes sufridos sirve para evaluar la política de seguridad y salud en el trabajo que tiene la empresa. En la aplicación que se va a presentar se asume que la única información que se cuenta es la muestra la frecuencia con que se encuentra un empleado que ha sufrido i accidentes laborales (tabla **6-7**) (Plackett, 1953). Por ejemplo, se observa que 2039 empleados han sufrido un accidente. En este caso el número total de empleados de la empresa es desconocido. Por lo cual es necesario conocer el número de empleados que no han padecido accidentes (n_0)

i	1	2	3	4	6
n_i	2039	312	35	3	1

Tabla 6-7: Frecuencia con que se encuentra un empleado que ha sufrido i accidentes laborales (Plackett, 1953)

La tabla **6-8** presenta las estimaciones del promedio de accidentes padecidos por un empleado (λ) y el número de empleados que no han padecido accidentes (n_0). Para encontrar los estimadores bayesianos se asumió que el verdadero valor de λ es 0.31. En este caso los tres estimadores del parámetro λ considerados coincidieron. Con respecto al número de empleados que no han padecido accidentes se encontró que el estimador más alto fue $\hat{n}_{08} = \hat{n}_{010} = 6,458$ y el más bajo $\hat{n}_{011} = 6,455$.

j	λ_i	n_{0i}
8	0.3150000	6,458
10	0.3150000	6,458
11	0.3150000	6,455

Tabla 6-8: Estimaciones del promedio de accidentes padecidos por un empleado (λ) y el número de empleados que no han padecido accidentes (n_0)

La tabla **6-9** presenta intervalos de confianza del promedio de accidentes padecidos por un empleado (λ). Se observa que el intervalo de menor longitud fué IC_3 , seguido de IC_5 e IC_1 :

j	LI_j	LS_j	Longitud
3	0.2860261	0.3466466	0.0606205
4	0.2855192	0.3460884	0.0605692

Tabla 6-9: Intervalo de confianza del promedio de accidentes padecidos por un empleado (λ).

6.4. Estimación del número total de hogares

Con información de la frecuencia con la cual se presentan i casos de cólera en un hogar (tabla **6-10**) en una villa de la India, Dahiya and Gross (1973) estimó el número total de hogares que habitaban dicha villa.

i	1	2	3	4
n_i	32	16	6	1

Tabla 6-10: Frecuencia con la cual se presentan i casos de cólera en un hogar en una villa de la India.

La tabla **6-11** presenta las estimaciones del promedio de personas que presentan casos de cólera por hogar (λ) y el número de hogares sin casos (n_0) con los diferentes métodos presentados. Para encontrar los estimadores bayesianos se asumieron que el verdadero valor de λ es 1.5.

Se encontró que el estimador más alto del promedio de personas que presentan casos de cólera por hogar fue $\hat{\lambda}_{11} = 1.4484439$ y el más bajo $\hat{\lambda}_{10} = 1.4402972$. Con respecto al número de hogares sin casos se encontró el mismo valor para todos los estimadores.

j	λ_i	n_{0i}
8	1.4425312	22
10	1.4402972	22
11	1.4484439	22

Tabla 6-11: Estimación del promedio de personas que presentan casos de cólera por hogar (λ) y el número de hogares sin casos (n_0)

La tabla **6-12** presenta intervalos de confianza para el promedio de personas que presentan casos de cólera por hogar. Se observa que el intervalo de menor longitud fue IC_4 .

j	LI_j	LS_j	Longitud
3	0.834030	1.176532	0.342502
4	0.828614	1.169917	0.341303

Tabla 6-12: Intervalos de confianza para el promedio de personas que presentan casos de cólera por hogar

7 Conclusiones

En este trabajo de grado se presentaron diferentes estimadores puntuales y por intervalo para los parámetros λ y n_0 de la distribución Poisson Truncada en Cero. Esta distribución ha mostrado ser útil en diferentes problemas de conteo. Algunas de las aplicaciones consideradas fueron la estimación del número de unidades de proteína presentes en las células, estimación del número de adictos de marihuana en 2006 en Tailandia, estimación del número de trabajadores en una empresa y la estimación del número de hogares.

En la comparación de los estimadores puntuales se utilizaron el ECM y el sesgo promedio. Se observó que el estimador $\hat{\lambda}_{12}$ es el que tiene menor ECM, no obstante su sesgo está entre los más altos. Adicionalmente el algoritmo propuesto para encontrar este estimador falla a medida que incrementa el tamaño de muestra. De este modo λ_8 o λ_{11} son una mejor elección pues no presenta problemas en su cálculo al aumentar el tamaño de muestra, además de tener un sesgo promedio cercano a cero y poseen el segundo menor *ECM*. En cuanto a los estimadores de n_0 se encontró que aunque el *ECM* de n_{012} es pequeño, su sesgo promedio está entre los más grandes. Por tanto, n_{011} presenta mejores propiedades en cuanto a *ECM* y *SP*.

En la comparación de los intervalos de confianza se utilizó la probabilidad de cobertura y la longitud del intervalo. Por un lado, para obtener la máxima cobertura en muestras pequeñas se prefiere el intervalo IC_6 y en muestras grandes se prefieren IC_3 e IC_4 . Por otro lado, se observó que IC_1 es el intervalo con menor longitud, no obstante este estimador tiene una cobertura real muy pequeña.

En resumen, para estimar λ se recomienda utilizar λ_{11} y λ_8 , en la estimación de n_0 se debe utilizar n_{011} , y en la estimación por intervalo es preferible usar IC_6 en muestras pequeñas e IC_3 o IC_4 en muestras grandes.

Bibliografía

- Berkson, J. (1980). Minimum chi-square, not maximum likelihood 8(3):457–487.
- Blumental, S., Dahiya, R. C., and Gross, J. A. (1978). Estimating the Complete Size from an Incomplete Sample. *Journal of the American Statistical Association*, 73:182–187.
- Cohen, A. C. (1960). Estimating the Parameter in a Condicional Poisson Distribution. *Biometrics*, 16:203–211.
- Cullen, M. J., Walsh, J., Nicholson, L. V. B., and Harris, J. B. (1990). Ultrastructural localization of dystrophin in human muscle by using gold immunolabelling. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 240(1297):197–210.
- Dahiya, R. C. and Gross, J. A. (1973). Estimating the Zero Class from a Truncated Poisson Sample. *Journal of the American Statistical Association*, 68:731–733.
- Daidoji, K. and Iwasaki, M. (2012). On interval estimation of the poisson parameter in a zero–truncated poisson distribution. *J. Jpn. Soc. Comp. Statist.*, 25:1–12.
- David, F. N. and Johnson, N. L. (1952). The Truncated Poisson. *Biometrics*, 8:275–285.
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 39:1–38.
- Finney, D. J. and Varley, G. C. (1955). An Example of the Truncated Poisson Distribution. *Biometrics*, 11:387–394.
- Harris, R. R. and Kanji, G. K. (1983). On the use of minimum chi-square estimation. 32(4):379–394.
- Hassan, A., Ahmad, P., and Bhatti, M. (2008). On the bayes estimator of parameter and reliability function of the zero-truncated poisson distribution. *Journal of the Korea Society for Industrial and Applied Mathematics*, 12:97–108.
- Irwin, J. O. (1959). Note: On the Estimation of the Mean of a Poisson Distribution from a Sample with the Zero Class Missing. *Biometrics*, 15:324–326.

- Kalbfleisch, J. G. (1985). *Probability and Statistical Inference Volumen 2.*, volume 2 of *Springer Texts in Statistics*. Springer, second edition edition.
- Mattews, J. N. S. and Appleton, D. R. (1993). An Application of the Truncated Poisson Distribution to Immunogold Assay. *Biometrics*, 49:617–621.
- Mir, K. A. (2011). Estimation in Truncated Poisson Distribution. *Mathematica Slovaca*, 61:289–296.
- Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the theory of statistics*. McGraw-Hill, Tokio, third edition.
- Moore, P. G. (1954). A Note on Truncated Poisson Distribution. *Biometrics*, 10:402–406.
- Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, USA, 1 edition.
- Plackett, R. L. (1953). The Truncated Poisson Distribution. *Biometrics*, 9:485–488.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rangan, S., Fletcher, A. K., and Goyal, V. K. (2012). Asymptotic analysis of map estimation via the replica method and applications to compressed sensing. *IEEE Transactions on Information Theory*, 58(3):1902–1923.
- Rider, P. R. (1953). Truncated Poisson Distributions. *Journal of the American Statistical Association*, 48:826–830.
- Rizzo, M. L. (2007). *Statistical Computing with R*. Chapman & Hall / CRC, Bowling Green, Ohio.
- Sanathanan, L. (1977). Estimating the Size of a Truncated Sample. *Journal of the American Statistical Association*, 72:669–672.
- Sarabia, J., Déniz, E., and Vázquez-Polo, F. (2006). *Estadística Actuarial: Teoría y aplicaciones*.
- Tate, R. F. and Goen, R. L. (1958). Minimum Variance Unbiased Estimation for the Truncated Poisson Distribution. *The Annals of Mathematical Statistics*, 29:755–765.
- Tsiatis, A. A., Rosner, G. L., and Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics*, 40(3):797–803.
- Vélez, J. I. and Correa, J. C. (2013). Estimación bayesiana del parámetro de la distribución poisson censurada sin ceros. *Cuadernos de Estadística Aplicada*, 1:1–9.

- Viwatwongkasem, C. (2016). Em algorithm for truncated and censored poisson likelihoods. *Procedia Computer Science*, 86:240–243.