

# Análisis de Datos Categóricos

Tarea 01

*Rivera Torres Francisco de Jesús*

*Rodríguez Maya Jorge Daniel*

*Samayoa Donado Víctor Augusto*

*Trujillo Barrios Georgina*

*Marzo 14, 2019*

## Ejercicio 1

Lee el artículo de Tapia, José A. y Nieto, F. Javier. **Razón de posibilidades: una propuesta de traducción de la expresión odds ratio.** *Salud Pública de México.* Julio-Agosto, 1993. Vol. 35 No.4 Pág. 419-424. Elabora un resumen en media cuartilla.

El artículo expone cómo las diversas traducciones hispánicas del término *odds ratio* son imprecisas en distintos sentidos y propone un nuevo término que, según los autores, se adecua mejor tanto lingüística como conceptualmente al significado matemático original. Dentro del repaso histórico por las traducciones e interpretaciones imprecisas, discutidas por los autores, destacan las siguientes. 1. La estipulación de *relative risk* como sinónimo de *odds ratio* en el diccionario de términos estadísticos del *International Statistical Institute* (no obstante este error lingüístico, los autores señalan que el concepto matemático se mantuvo correcto). 2. La confusión producida en textos epidemiológicos en los que se refería a otros términos para hacer alusión al concepto de *odds ratio*: desigualdad relativa, posibilidades relativas, riesgo relativo estimado, por ejemplo. 3. La confusión del concepto matemático con su cálculo y el consecuente surgimiento de términos como *cross product ratio* en los que pierde sentido lingüístico la idea matemática original. 4. La traducción específica como *razón de momios* que sólo tiene sentido en México, por la similitud entre el uso anglosajón del término *odds* (no la traducción literal) y el uso mexicano de *momios*. La confusión entre probabilidad y *odds* reflejada en la traducción como *relación* o *razón de probabilidad*. A partir de las razones mencionadas, los autores mencionan que ciertas propuestas de traducción, como *razón de ventaja*, han resultado salvaguardar el significado matemático de *odds ratio* pero su articulación en frases resulta extraña al español. Por ello proponen traducir *odds* como *posibilidades* que además de ser compatible con el significado y uso del término inglés *odds* permite articular frases que reflejan el sentido original del término.

## Ejercicio 2

¿Qué es la sensibilidad (*sensitivity*) y la especificidad (*specificity*) de una prueba? ¿Qué es la curva ROC? Busca algún ejemplo para ejemplificar estos conceptos.

Para estos dos casos, consideremos la siguiente tabla:

Prueba \ Realidad	Positivo	Negativo
	Verdadero positivo (TP)	Falso Positivo (FP)
Positivo		
Negativo	Falso Negativo (FN)	Verdadero negativo (TN)

La **sensibilidad** de una prueba hace referencia a la capacidad de la prueba para detectar “pacientes” que efectivamente tienen cierta condición (la prueba da positivo). Esto es, es el cociente que existe entre el número de verdaderos positivos dados por la prueba entre el total de pacientes que tienen la condición:

$$\text{sensibilidad} = \frac{TP}{TP + FN}$$

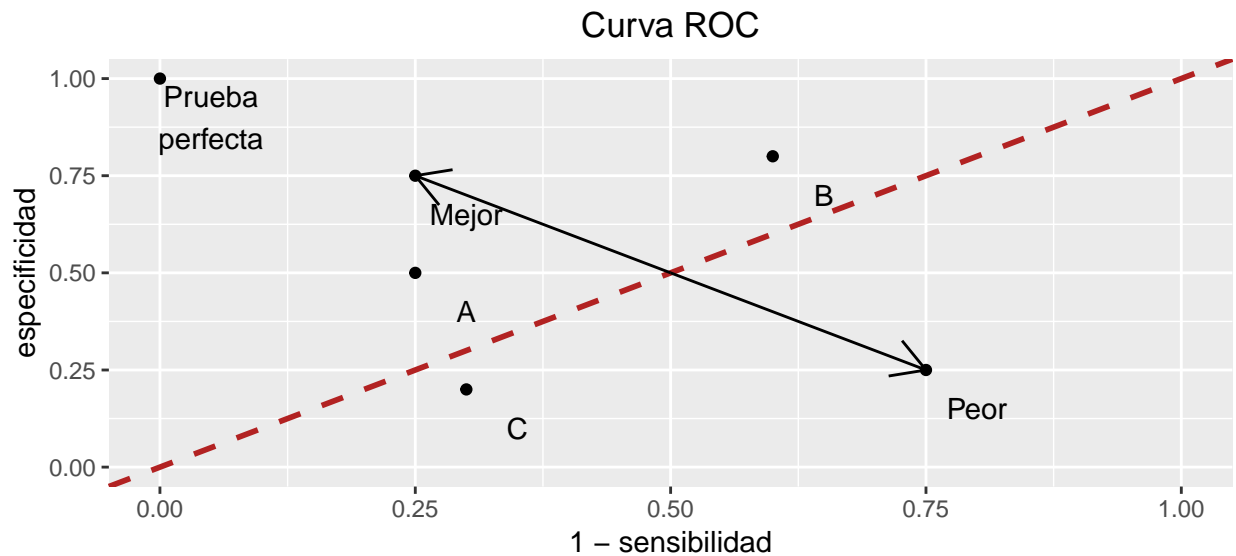
En el caso de una prueba con una alta sensibilidad, se puede afirmar que un resultado negativo es útil para descartar la presencia de la condición (enfermedad). Sin embargo, un resultado positivo, en el caso de una prueba con alta sensibilidad, no se puede usar para confirmar (asegurar) que el paciente padece cierta condición.

La **especificidad** de una prueba hace referencia a la capacidad de la prueba a detectar de forma correcta “pacientes” que carecen de cierta condición (la prueba da negativo). Esto es, es el cociente que existe entre los verdaderos negativos dados por la prueba entre el total de pacientes que efectivamente carecen de la condición en cuestión:

$$\text{especificidad} = \frac{TN}{TN + FP}$$

En el caso de una prueba con alta especificidad, se puede afirmar que un resultado positivo es útil para confirmar la presencia del padecimiento de cierta condición (enfermedad). Mientras que, para un resultado negativo, en el caso de una prueba con alta especificidad, no se puede usar para confirmar que el paciente carece de cierto padecimiento.

La **curva ROC** es una representación gráfica de la sensibilidad respecto a la especificidad cuando se trabaja en pruebas de clasificación binaria (tiene el padecimiento, carece del padecimiento). En el eje  $X$  se coloca el valor de  $1 - \text{sensibilidad}$  y en el eje  $Y$  se coloca el valor de la especificidad. En el caso de la “prueba ideal” se tendría un valor de sensibilidad igual a 1 y un valor especificidad igual a 1, por lo que dicha prueba se colocaría en la coordenada  $(1 - \text{sensibilidad}, \text{especificidad}) = (0, 1)$ .



La gráfica anterior representa la posición de tres pruebas *prueba perfecta*, *prueba A* y *prueba B*. La prueba perfecta es, como se mencionó previamente, es aquella que tiene especificidad y sensibilidad igual a 1. La prueba A tiene una sensibilidad mayor respecto a la prueba B, pero tiene una especificidad menor que la prueba B. La prueba C no tiene un buen desempeño al tener el valor de  $1 - \text{sensibilidad}$  y la especificidad por abajo de la recta puntada (el área “peor”), implica que su desempeño es peor que el de “adivinar” lanzando una moneda para determinar el resultado.

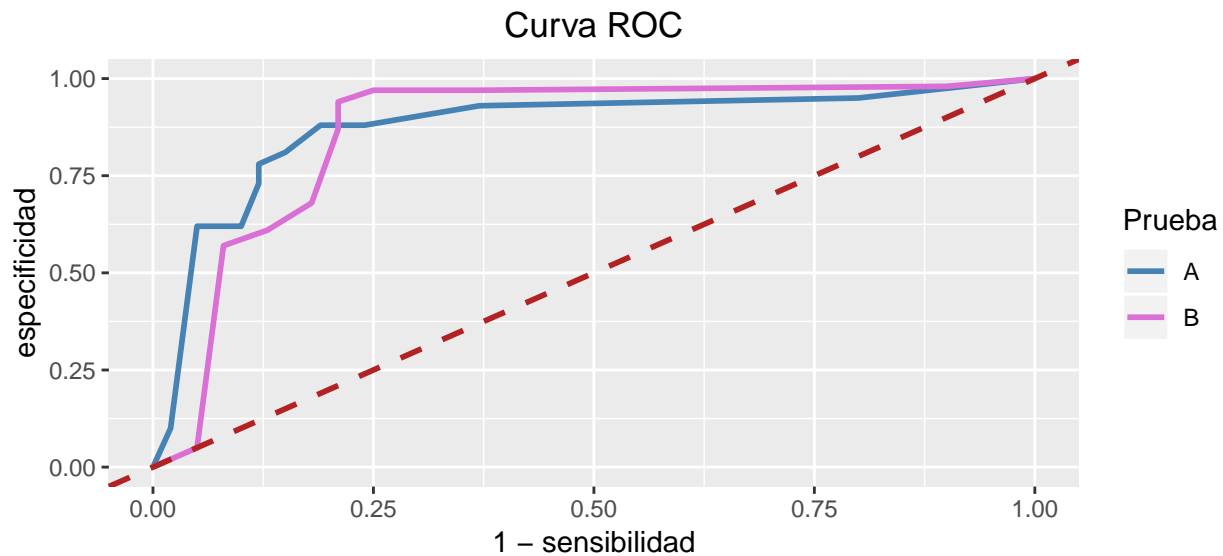
La gráfica anterior permite determinar de una forma visual la comparación entre el desempeño entre distintas pruebas. Por ejemplo, si nos interesa una prueba con alta sensibilidad, se tomaría la prueba A. Mientras que si se requiere una prueba con alta especificidad se elegirá la prueba B.

El nombre de “curva” proviene de la determinación sobre el comportamiento de cada una de las pruebas para distintos “puntos de corte”. Esto es, al determinar una prueba es normal obtener como resultado un valor numérico entre 0 y 1 que indica la “probabilidad de padecer” cierta condición. En este caso, se determina un valor de probabilidad a partir del cual si el resultado de la prueba es mayor a este valor se considerará un resultado positivo y si el resultado de la prueba es menor a este valor se considerará un resultado negativo.

La selección de dicho punto de corte conlleva a que la tabla de contingencia se vea modificada al cambiar los valores de verdaderos positivos (TP), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (TN) y por ende cambia el valor de la sensibilidad y la especificidad.

Lo anterior permite elegir el “punto de corte” de acuerdo con el objetivo de la prueba. Esto es, se requiere crear una prueba que me de una gran certeza que los resultados negativos aseguren una ausencia de una condición, entonces se desea elegir el punto de corte que nos de una alta sensibilidad. Si por el contrario, se busca que los resultados positivos aseguren la presencia de una condición, entonces se desea elegir el punto de corte que nos de una alta especificidad.

Como ejemplo, consideremos las pruebas A y B para determinar si una persona tiene diabetes. Ambas pruebas dan como resultado un “score” (valor entre 0 y 1), y por cada punto de corte definido para determinar si se padece o no diabetes se obtienen distintos valores de sensibilidad y especificidad (ver gráfica inferior).

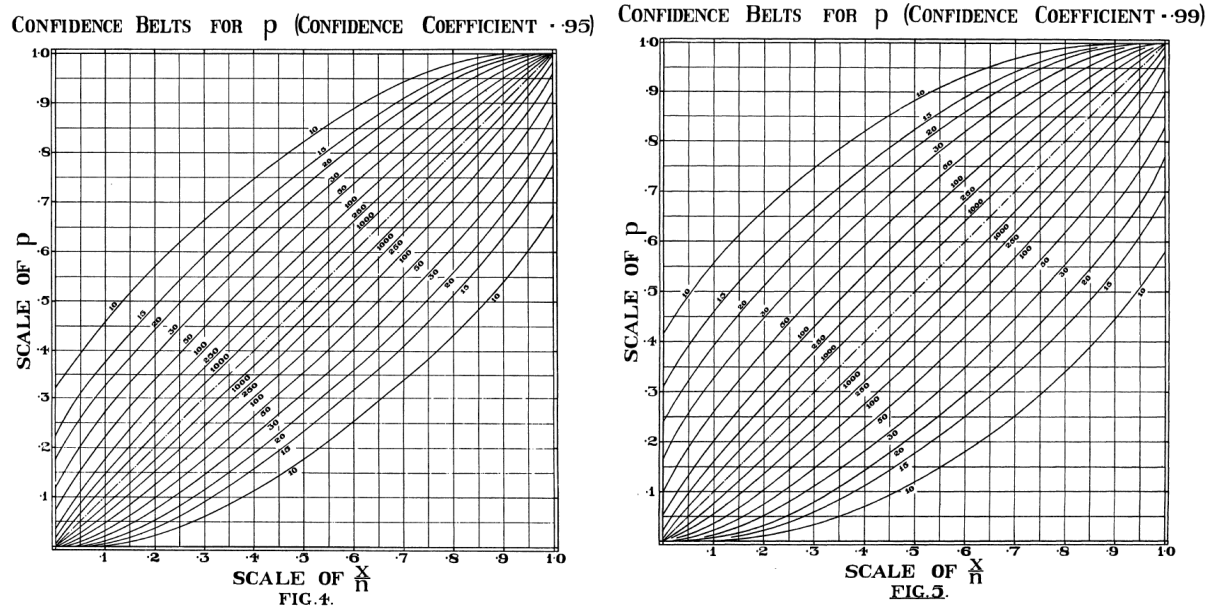


### Ejercicio 3

En el artículo Clopper, C. J. & Pearson, E. S. 1934. **The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial.** *Biometrika*, Vol. 26, No. 4, vienen las tablas para calcular un intervalo EXACTO de confianza para  $p$ . Úsa esas tablas para calcular un intervalo al 95% y otro al 99% para una muestra tamaño  $n = 50$  y  $x = 9$ .

No se encontraron las tablas en el artículo, pero se encontraron las gráficas asociadas a dichas tablas. Se usarán estas gráficas para obtener los intervalos de confianza.

El procedimiento para encontrarlos es fijar el valor de  $x$  en el eje horizontal y seguir una línea recta vertical hasta intersectar las dos curvas que estén asociadas al valor de  $n$ . Ahí se obtendrá la coordenada en  $y$  referente a los límites inferior y superior del intervalo de confianza para  $p$ .



Para el caso de  $n = 50$  y  $x = 9$ , con un intervalo de confianza al 95%, se tiene  $(0.775, 0.960)$ .

Para el caso de  $n = 50$  y  $x = 9$ , con un intervalo de confianza al 99%, se tiene  $(0.740, 0.975)$ .

## Ejercicio 4

Indica el esquema de muestreo y la hipótesis nula y alternativa para cada uno de los siguientes ejemplos:

### Ejemplo a)

En un hospital se decide registrar el número de parejas que lleguen en 1 año con la finalidad de someterse a un tratamiento de fertilidad. Al cabo del año, se genera una tabla en donde se registra el grupo de edad de la mujer al iniciar el tratamiento (20 a 30 y 31 a 45) y el resultado (embarazo o no embarazo).

El esquema de muestreo es **Poisson**

Las hipótesis nula y alternativa son:

$H_0$  = Edad y resultado (del tratamiento) no están relacionados

$H_a$  = Edad y resultado presentan algún grado de asociación

### Ejemplo b)

En un hospital se quiere evaluar la asociación entre edad de la mujer y resultado de un tratamiento de fertilidad. Se establece al inicio del estudio que se registraran los resultados de 722 personas. Al

finalizar el tratamiento se genera una tabla en donde se registra el grupo de edad de la mujer al iniciar el tratamiento (20 a 30 y 31 a 45) y el resultado (embarazo o no embarazo).

El esquema de muestreo es **Multinomial**

Las hipótesis nula y alternativa son:

$H_0$  = Las variables edad y resultado (del tratamiento) son independientes

$H_a$  = Las variables edad y resultado presentan algún grado de dependencia

Expresado en términos de ecuaciones se tiene

$$H_0 = P(\text{edad}, \text{resultado}) = P(\text{edad}) \cdot P(\text{resultado})$$

$$H_a = P(\text{edad}, \text{resultado}) \neq P(\text{edad}) \cdot P(\text{resultado})$$

### Ejemplo c)

En un hospital se quiere evaluar si el resultado de un tratamiento de fertilidad cambia de acuerdo a la edad de la mujer que desea embarazarse. Para ello se deciden registrar los resultados de 350 mujeres cuya edad está entre los 20 y 30 años, y 360 mujeres cuya edad está entre los 31 y 45 años. Al finalizar el tratamiento se genera una tabla en donde se registra el grupo de edad de la mujer al iniciar el tratamiento y el resultado (embarazo o no embarazo)

El esquema de muestreo es **Multinomial-Producto**

Las hipótesis nula y alternativa son:

$H_0$  = Hay homogeneidad en los resultados respecto a los dos grupos de mujeres

$H_a$  = No hay homogeneidad en los resultados respecto a los dos grupos de mujeres

Si denotamos por  $p_1$  la distribución de probabilidad del resultado (del tratamiento) para las mujeres con edad entre 20 a 30 años y por  $p_2$  la distribución de probabilidad del resultado (del tratamiento) para las mujeres con edad entre 31 a 45 años, entonces

$$H_0 = p_1 = p_2 \quad \text{v.s.} \quad H_a = p_1 \neq p_2$$

### Ejercicio 5

Los datos del ejemplo 4a) se han resumido en la siguiente tabla:

Edad de la mujer \ Resultado	Embarazo	No Embarazo
20 a 30 años	26	85
31 a 45 años	146	565

**Inciso 5.1**

Realicen “a mano” la prueba de Ji cuadrada. Concluyan.

**Inciso 5.2**

Corrobores el resultado en R. Muestren el resultado.

**Inciso 5.3**

Calculen la probabilidad de embarazo para mujeres de 20 – 30 años y la probabilidad de embarazo para mujeres de 31 – 45 años

**Inciso 5.4**

Calculen el Odds Ratio a mano y en R. Interpreten el valor OR y los límites superior e inferior del intervalo de confianza.

**Inciso 5.5**

Calculen el Relative Risk a mano y en R. Interpreten el valor RR y los límites superior e inferior del intervalo de confianza.

**Ejercicio 6**

Se realizó un estudio en la India para determinar la asociación entre dos tipos de virus de papiloma humano con la aparición en las lesiones del cérvix. Realicen una prueba exacta de Fisher “a mano” y en R. Interpreten.

Tipo de Virus \ Cervix	Sin lesiones	Con lesiones
HPV 11	47	3
HPV 16	14	42

## Ejercicio 7

Comparación de métodos para generar intervalos de confianza.

### Inciso 7.1

Generen intervalos de confianza al 99% para una  $p$  estimada de  $p = 0.9$  y  $p = 0.5$  con  $n = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$  usando los comandos `binom.exact`, `binom.wilson` y `binom.asymp`.

El código utilizado en R es el siguiente:

### Inciso 7.2

Muestra los límites de los intervalos generados a través de los tres métodos

Tabla 1: Intervalos de confianza para  $p = 0.9$

x	n	media	exacto		wilson		asymp	
			Inferior	Superior	Inferior	Superior	Inferior	Superior
9	10	0.9	0.456	0.999	0.493	0.988	0.656	1.144
18	20	0.9	0.613	0.995	0.621	0.980	0.727	1.073
27	30	0.9	0.680	0.988	0.681	0.974	0.759	1.041
36	40	0.9	0.717	0.983	0.716	0.970	0.778	1.022
45	50	0.9	0.742	0.978	0.740	0.966	0.791	1.009
54	60	0.9	0.759	0.974	0.757	0.963	0.800	1.000
63	70	0.9	0.773	0.970	0.771	0.960	0.808	0.992
72	80	0.9	0.783	0.967	0.781	0.958	0.814	0.986
81	90	0.9	0.791	0.964	0.789	0.956	0.819	0.981
90	100	0.9	0.798	0.962	0.796	0.954	0.823	0.977

Tabla 2: Intervalos de confianza para  $p = 0.5$

x	n	media	exacto		wilson		asymp	
			Inferior	Superior	Inferior	Superior	Inferior	Superior
5	10	0.5	0.128	0.872	0.184	0.816	0.093	0.907
10	20	0.5	0.218	0.782	0.250	0.750	0.212	0.788
15	30	0.5	0.265	0.735	0.287	0.713	0.265	0.735
20	40	0.5	0.295	0.705	0.311	0.689	0.296	0.704
25	50	0.5	0.316	0.684	0.329	0.671	0.318	0.682
30	60	0.5	0.331	0.669	0.342	0.658	0.334	0.666
35	70	0.5	0.343	0.657	0.353	0.647	0.346	0.654
40	80	0.5	0.353	0.647	0.362	0.638	0.356	0.644

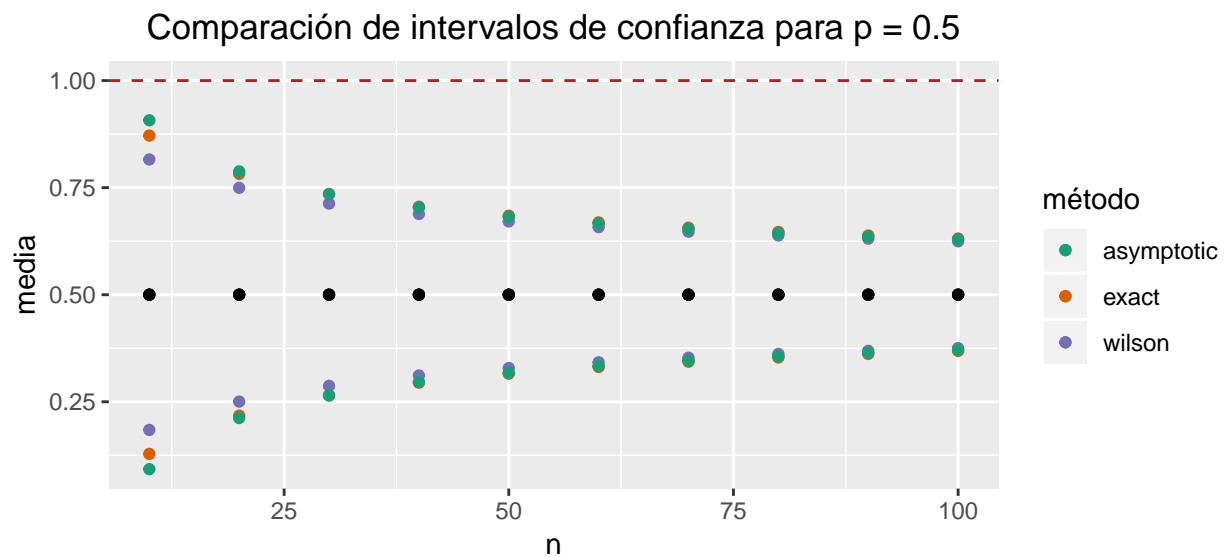
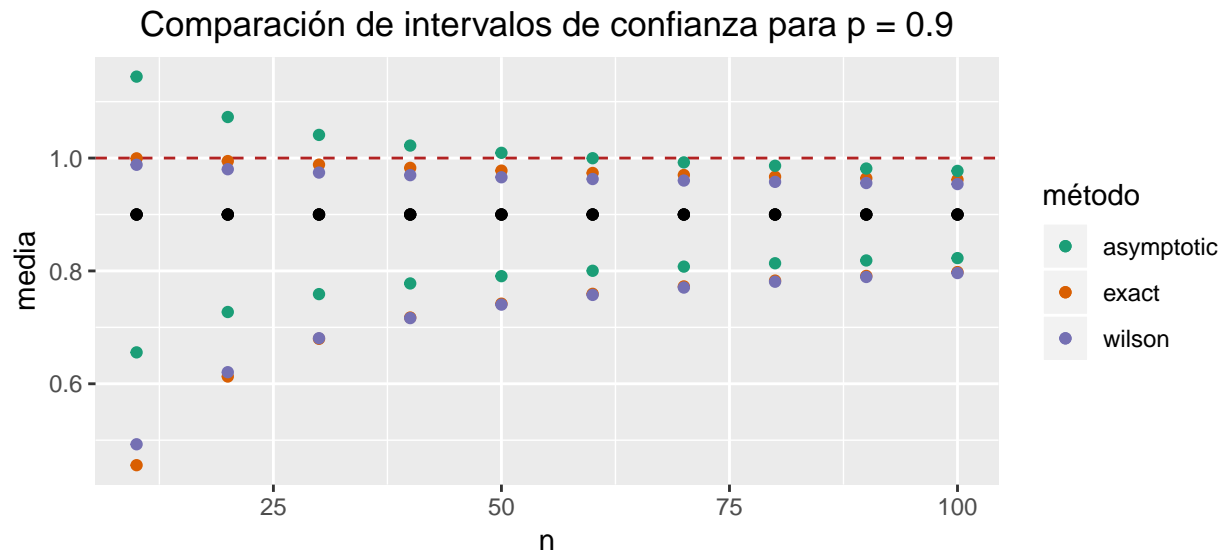


Tabla 2: Intervalos de confianza para  $p = 0.5$  (*continued*)

x	n	media	exacto		wilson		asympt	
			Inferior	Superior	Inferior	Superior	Inferior	Superior
45	90	0.5	0.362	0.638	0.369	0.631	0.364	0.636
50	100	0.5	0.369	0.631	0.375	0.625	0.371	0.629

**Inciso 7.3**

Realicen una gráfica en donde comparen los intervalos generados por los tres métodos y las dos  $p$  estimadas

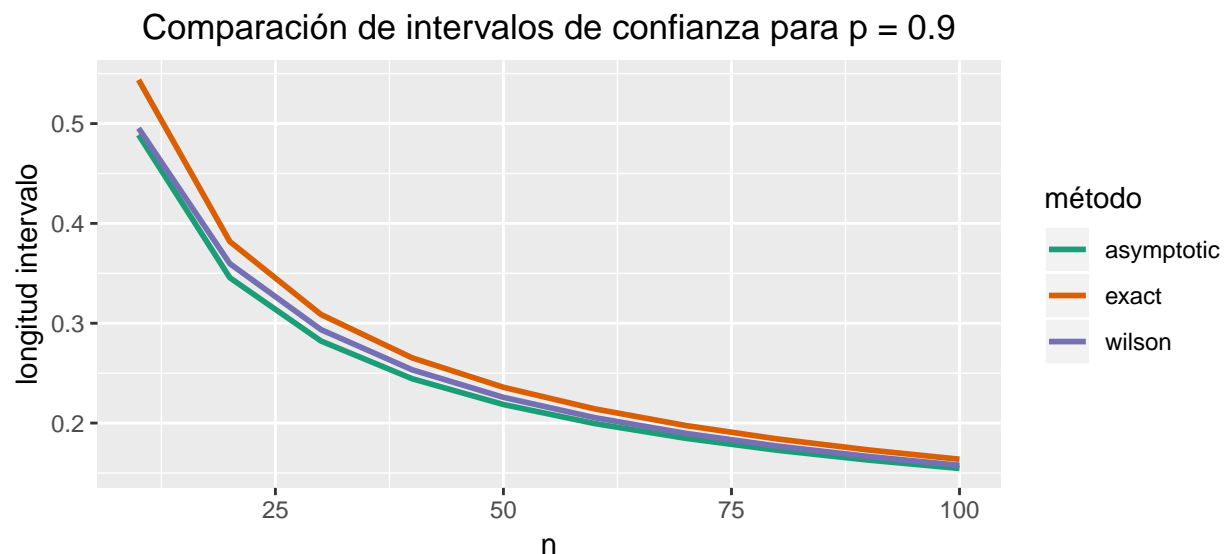


## Inciso 7.4

¿Cuál método elegirían? ¿Cuál método no elegirían? ¿Por qué?

En la gráfica de intervalos de confianza al 99% para  $p = 0.9$ , notamos que el método `binom.asymp` contiene parte del intervalo de confianza mayor a  $p = 1$  para valores con  $n < 60$ . Por tal motivo este método no lo usaríamos cuando  $p$  sea cercano a 0.9.

Graficando la longitud de los intervalos de confianza para  $p = 0.9$ , y considerando que `binom.asymp` no lo usaremos por exceder el valor de  $p = 1$ . Notamos que el método `binom.wilson` es el que genera los intervalos de confianza de menor longitud. Por lo que, para el caso en que  $p \approx 0.9$  es mejor utilizar el método **Wilson** para obtener el intervalos de confianza.



Para el caso cuando  $p = 0.5$  se observa todos los métodos tienen intervalos de confianza dentro del rango  $(0, 1)$ . Por lo tanto, usaremos el criterio de menor longitud para seleccionar el método para el cálculo del intervalos.

En la gráfica inferior, se observa que el intervalo generado con `binom.wilson` es el que tiene la menor longitud para los valores de  $n$  en el rango de  $n = 10, \dots, 100$ . Por lo tanto, para el caso en que  $p \approx 0.5$  se usará el método de **Wilson** para obtener el intervalo de confianza.

