Grau d'Estadística: Estadística Mèdica

Problemas del Bloque 3

Fecha de entrega: domingo 2 de diciembre 2018

Alumna: Laura Julià Melis

EJERCICIO 1

Presentad un ejemplo (real) de un estudio de cohorte y otro de un estudio caso-control y contestad las siguientes preguntas para cada estudio:

i) Estudio de cohorte:

Fuente: *Pages 1373–1378, https://doi.org/10.1093/jnci/89.18.1373*

Abstract

Background: Breast cancer mortality and incidence rates vary by geographic region in the United States. Previous analytic studies have measured mortality, not incidence, and have used regional prevalences to control for geographic variation in risk factors rather than adjusting for risk factors measured at the level of the individual. We prospectively evaluated regional variation in breast cancer incidence rates in the Nurses' Health Study and assessed the influence of breast cancer risk factors measured at the individual level. *Methods:* The Nurses' Health Study cohort was established in 1976 when 121 700 female nurses aged 30–55 years living in 11 U.S. states were enrolled. These states represent all four regions of the continental United States. We identified 3603 incident cases of invasive breast cancer through 1992 (1 794 565 person-years of follow-up). We calculated relative risks (RRs) adjusted for age and for age and established risk factors (i.e., multivariateadjusted analysis), comparing California, the Northeast, and the Midwest with the South. *Results:* For premenopausal women, there was little evidence of regional variation in breast cancer incidence rates, either in age-adjusted or in multivariate-adjusted analyses. For postmenopausal women in California, age-adjusted risk was modestly elevated (RR = 1.24; 95% confidence interval [CI] = 1.05–1.47); after adjusting for age and for established risk factors, the excess rate in California was attenuated by 25% (RR = 1.18; 95% CI = 1.00–1.40). No excess of breast cancer incidence was observed for postmenopausal women in either the Northeast or the Midwest. *Conclusions:* Little regional variation in age-adjusted breast cancer incidence rates was observed, with the exception of a modest excess for postmenopausal women in California. Adjustment for differences in the distribution of established risk factors explained some of the excess risk in California.

a) ¿Cuál es el objetivo principal del estudio?

Evaluar prospectivamente la variación regional de las tasas de incidencia de cáncer de mama del estudio *The Nurses' Health Study* y valorar la influencia de los factores de riesgo de cáncer de mama (medidos a nivel individual).

b) ¿Cuál es la población de interés?

Las mujeres estadounidenses.

c) ¿Cuáles son las muestras y qué tamaños tienen?

La del estudio *The Nurses' Health Study*: 121 700 enfermeras de entre 30 y 55 años que vivían en 11 estados de EEUU (California, Connecticut, Florida, Maryland, Massachusetts, Michigan, Nueva Jersey, Nueva York, Ohio, Pensilvania y Texas) en el año 1976.

d) ¿Cuáles son la(s) enfermedad(es) y la(s) exposición (exposiciones) de interés?

La enfermedad de interés es el cáncer de mama y la exposición es el estado en el que reside la persona estudiada.

e) ¿Cuál será la razón por el diseño escogido en cada caso?

En un estudio de cohorte prospectivo, los individuos no tienen la enfermedad de interés (en este caso, cáncer de mama) y son seguidos durante un tiempo para observar la frecuencia con que la enfermedad aparece (incidencia) en cada uno de los grupos. Este es el objetivo del presente estudio y por ello este diseño es el adecuado.

ii) Estudio de caso-control:

Fuente: 1990 Sep;108(9):1274-80, https://www.ncbi.nlm.nih.gov/pubmed/2400347

Abstract
Uveal melanoma threatens life, as well as sight. To evaluate the effect of constitutional factors and UV radiation on the risk of uveal melanoma, 197 cases in New England were compared with 385 matched population controls, identified by random-digit dialing, and 337 cases residing within the United States were compared with 800 sibling controls. In the population-based comparison, estimated relative risks (RRs) of uveal melanoma, after adjustment for other factors, were elevated for the following: ancestry from more northern latitudes with a substantially elevated risk for Northern European ancestry (RR, 6.5; 95% confidence interval [CI], 1.9 to 22.4) and more than a twofold risk for British ancestry (RR, 2.4; 95% CI, 1.1 to 5.1), as compared with Southern European or other Mediterranean heritage; light skin color as compared with dark (RR, 3.8; 95% CI, 1.1 to 12.6); and 10 or more cutaneous nevi as compared with none (RR, 2.7; 95% CI, 1.5 to 4.9). There was a statistically significant trend for increasing risk with more northern heritage and more moles. Southern residence (below latitude 40 degrees N) for more than 5 years also increased risk (RR, 2.8; 95% CI, 1.1 to 6.9), as compared with none. In both comparisons, use of sunlamps was a risk determinant (RR, 3.4; 95% CI, 1.1 to 10.3 with random-digit dialed controls and RR, 2.3; 95% CI, 1.2 to 4.3 with sibling controls, comparing occasional or frequent use to never use), as was intense sun exposure (RR, 1.7; 95% CI, 0.9 to 3.0 and RR, 2.1; 95% CI, 1.4 to 3.2, respectively). However, birthplace below latitude 40 degrees N and outdoor work were associated with a lower risk.
(ABSTRACT TRUNCATED AT 250 WORDS).

a) ¿Cuál es el objetivo principal del estudio?

Evaluar el efecto de los factores constitucionales y la radiación UV sobre el riesgo de melanoma uveal.

b) ¿Cuál es la población de interés?

La población de Estados Unidos.

c) ¿Cuáles son las muestras y qué tamaños tienen?

En Nueva Inglaterra se compararon 197 casos con 385 controles y en los Estados Unidos se compararon 337 casos con 800 controles, donde casos hace referencia a sujetos con la enfermedad y controles a sujetos sin la enfermedad.

d) ¿Cuáles son la(s) enfermedad(es) y la(s) exposición (exposiciones) de interés?

La enfermedad de interés es el melanoma uveal (tumor maligno en el ojo) y las exposiciones, las siguientes:

- Nivel de latitud de la ascendencia del sujeto (ascendencia en el norte de Europa y en el Sur o Mediterráneo).
- Color de piel (claro y oscuro).
- Presencia de nevus cutáneos. (ninguno y 10 o más)
- Años resididos en el sur (ninguno y más de 5)
- Uso de lámparas solares.
- Exposición intensa al sol.
- Latitud del lugar de nacimiento (por debajo de 40 grados N y por encima)
- Trabajo al aire libre.

e) ¿Cuál será la razón por el diseño escogido en cada caso?

En este caso el estudio de casos y controles es útil para conseguir el objetivo deseado ya que al investigar si los sujetos estuvieron expuestos o no a las características de interés (mencionadas en el apartado d) permite comparar la proporción de expuestos en ambos grupos (casos y controles) y poder concluir qué factores son de riesgo para padecer la enfermedad (melanoma uveal).

EJERCICIO 2

¿A qué tipo de medida epidemiológica nos referimos a continuación?

Medidas vistas en clase *(para tenerlo presente):*

- **Prevalencia de una enfermedad (D).** Proporción de individuos afectados por la enfermedad (X) entre la población de interés(N) en un momento dado t. [P=X/N]
- **Incidencia acumulada o riesgo.** Proporción de nuevos casos durante un periodo de tiempo dividido entre la población inicial sin enfermedad.
- **Tasa de incidencia o incidencia.** Número de nuevos casos (I) por unidad de persona-tiempo en riesgo. Es la tasa a la que ocurren nuevos eventos en una población. Es necesario conocer el tiempo que está en riesgo el sujeto de estudio. Estimada en estudios de cohorte.
- **Riesgo relativo.** Tasa del riesgo de padecer una enfermedad de la gente expuesta entre el riesgo de padecerla que tiene la gente no expuesta.
- **Diferencia de riesgos.** Es la diferencia entre el riesgo de padecer enfermedad entre los sujetos expuestos y los no expuestos.
- **Odds ratio.** Probabilidad de enfermar entre las personas expuestas entre la probabilidad de enfermar entre las personas no expuestas
- **Riesgo atribuible.** Proporción de casos de enfermedad en la población que son atribuibles a la exposición.

a) **El número de casos de depresión postparto entre las mujeres de un estudio de cohorte dividido por el tiempo total bajo riesgo de estas mujeres durante el año que duró el estudio.**

Tasa de incidencia.

b) **El número de personas residentes en Barcelona que enfermaron de gripe entre el 1 de noviembre de 2014 y el 31 de marzo de 2015 dividido por el número de personas residentes en Barcelona el 1 de noviembre de 2014 libres de la gripe.**

Incidencia acumulada (proporción de incidencia o riesgo)

c) **El número de casos de malaria en África occidental en 2015 dividido por el número habitantes de África occidental en 2015.**

Prevalencia

d) **El número de nuevos casos de cáncer de pulmón en hombres de 55 a 59 años en la provincia de Barcelona durante el año 2014 dividido por el número de hombres de 55 a 59 años en la provincia de Barcelona el 1 de julio de 2014.**

Tasa de incidencia anual (es el número de nuevos casos durante el año dividido entre la población a mitad del año.)

e) **El número de neonatos con microcefalia en Centroamérica entre diciembre de 2015 y febrero de 2016 dividido por el número de todos los neonatos en Centroamérica en este periodo.**

Prevalencia

EJERCICIO 3

En el artículo *Prediction of Psychosis in Adolescents and Young Adults at High Risk* de Ruhrmann et al. (2010) se presentan los resultados de un estudio prospectivo sobre la psicosis en gente joven de alto riesgo. La medida de interés es la incidencia de la psicosis y en el resumen se presenta como resultado:

"At 18-month follow-up, the incidence rate for transition to psychosis was 19 %."

En el apartado *Statistical analysis* se explica cómo se ha hecho este cálculo.

¿Es correcto el uso del término *incidence rate* en este caso? Razonad la respuesta.

El estudio realiza el ratio de Hazard acumulado, como se muestra a continuación:

"The iIR of transition to psychosis after 6, 9, 12, and 18 months was 7%, 11%, 14%, and 19%, respectively (Figure 1)."

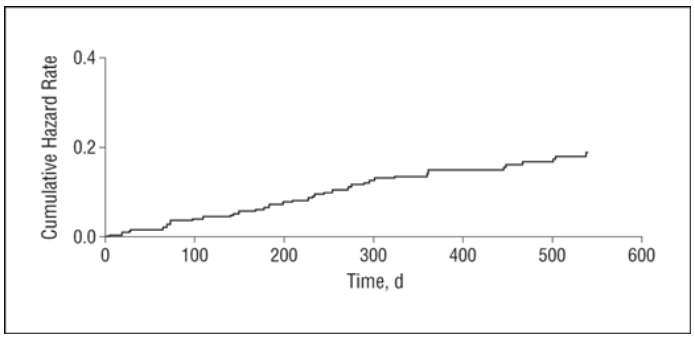

Figure 1: Kaplan-Meier survival analysis for 18-month follow-up (n = 245).

Esta ratio acumulado de Hazard se interpreta como la probabilidad de error o fallo en el tiempo t dada la supervivencia hasta ese tiempo t. En este estudio, corresponde a la probabilidad de que ocurra la transición a psicosis en el mes t después de t meses de seguimiento (t= 6, 9, 12 y 18). Y al multiplicar la probabilidad obtenida por 100, es como se puede obtener esta medida en tanto por ciento

Pero, por otro lado, la tasa de incidencia es una medida que se interpreta en unidades de personas-años y se interpreta como el número de casos nuevos que se dan al observar x personas durante t años, y que se calcula dividiendo en número de casos nuevos (que correspondería al número de sujetos que padecen la transición a psicosis) entre el tiempo bajo riesgo (que sería el número de personas estudiadas en t=1 año, por ejemplo).

Así pues, no es correcto usar el término "ratio de incidencia" en este caso.

EJERCICIO 4

Se quiere estudiar la incidencia de una cierta enfermedad en función de una exposición de interés. Para ello se dispone de los datos –casos de enfermedad y tiempo total de seguimiento en distintos grupos de edad– de la Tabla 1.

Tabla 1: Datos de incidencia de una enfermedad de interés

	Expuestos a E		No-Expuestos a E	
Edad	Casos	*Person-years*	Casos	*Person-years*
35–44	32	52407	2	18790
45–54	104	43248	12	10673
55–64	206	28612	28	5710

a) Estimad las tasas de incidencia (en casos por 10 000 personas-años) para ambos grupos de exposición y cada grupo de edad. Para ello podéis usar la función pois.exact del paquete epitools.

La metodología que se utiliza es la siguiente:

1. Obtenemos la tasa de incidencia (por persona-año) con:

   `pois.exact(casos, pt = personas-años)$rate`

2. Multiplicamos la tasa por 10000 para obtener la tasa en casos por 10 000 personas-años.

Código utilizado:

```
library(epitools)

# Expuestos: Grupo edad 35-44
pois.exact(32, pt = 52407)$rate*10000

# Expuestos: Grupo edad 45-54
pois.exact(104, pt = 43248)$rate*10000

# Expuestos: Grupo edad 55-64
pois.exact(206, pt = 43248)$rate*10000


# No expuestos: Grupo edad 35-44
pois.exact(2, pt = 12790)$rate*10000

# No expuestos: Grupo edad 45-54
pois.exact(12, pt = 10673)$rate*10000

# No expuestos: Grupo edad 55-64
pois.exact(28, pt = 5710)$rate*10000
```

Tabla de resultados:

	GRUPO	Tasa de incidencia estimada (en casos por 10 000 personas-años)
EXPUESTOS	Edad 35-44	6.106055
	Edad 45-54	24.04735
	Edad 55-64	47.63226
NO EXPUESTOS	Edad 35-44	1.563722
	Edad 45-54	11.24332
	Edad 55-64	49.03678

La idea (poniendo como ejemplo el grupo de expuestos de 35-44 años) es que, si se dan 32 casos por 52407 personas-años, se esperan 32 casos por 52407 personas observadas durante un año. Al desear conocer la tasa de incidencia en casos por 10000 personas-años, lo que se nos está preguntando es cuál será el número de casos que se esperan si se observan 10000 personas durante un año. Y entonces, analíticamente podemos obtener la tasa de incidencia con una regla de tres:

$$x = \frac{10000 \cdot 32}{52407} = 6.10605453$$

Que es equivalente a la tasa obtenida con la metodología anterior utilizando la función de r "pois.exact".

b) Calculad los intervalos de confianza del 95% aproximados (en casos por 10 000 personas-años) en el grupo de edad de 35 a 44 años usando la fórmula de la Transparencia 21/27 del Tema 2.

$$CI(I_r; 1-\alpha) = \hat{I}_r \pm z_{1-\alpha/2} \cdot \sqrt{\hat{I}_r/\Delta T}$$

Donde $\hat{I}_r = \frac{I}{\Delta T}$ es el nombre de incidentes (nuevos casos) que se dan durante todo el tiempo bajo riego (ΔT).

En el apartado anterior hemos obtenido que el grupo expuesto de edad entre 35 y 44 años tiene una tasa de incidencia estimada $\hat{I}_r = 6.106055$ en casos por 10000 personas-años (tasa por año de 0.0006106055) y la tabla del enunciado nos dice que $\Delta T = 52407$. Además, como se pide un nivel de confianza del 95%, ($1 - \alpha = 0.95$), sabemos que $\alpha = 0.05$, por lo que mirando en la tabla de la distribución normal obtenemos que $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1{,}96$.

Así pues, sólo nos queda substituir estos valores en la fórmula anterior:

$$CI_{0,95} = 0.0006106055 \pm 1{,}96 \cdot \sqrt{\frac{0.0006106055}{52407}} = [0.0003990415, 0.0008221695]$$

Y para obtenerlo en casos por 10000 personas-años, multiplicamos el resultado por 10000 y obtenemos:

$$\boxed{CI_{0,95} = [3.990415, 8.221695]}$$

También lo podemos obtener con la función de R **`pois.exact`** haciendo lo siguiente:

```
pois.exact(32, pt = 52407)$lower*10000 # para el limite inferior
pois.exact(32, pt = 52407)$upper*10000 # para el límite superior
```

Con lo que se obtienen unos valores diferentes pero similares:

Tasa por año	Tasa en casos por 10 000 personas-años	Método	Intervalo de confianza al 95%
0.0006106055	6.106055	Fórmula	[3.990415, 8.221695]
		R	[4.176537, 8.619932]

c) Estimad la incidencia acumulada en un periodo de 7 años para una persona expuesta que acaba de cumplir 55 años.

La incidencia acumulada es la proporción de nuevos casos durante un periodo de tiempo dividido entre la población inicial sin enfermedad.

$$CI(\Delta) = \frac{I}{N_0}$$

Donde I es el número de casos nuevos (incidentes) y N_0, el tamaño de la población inicial libre de enfermedad.

Al acabar de cumplir 55 años, esta persona pertenece al grupo de edad 55-64, que para los sujetos expuestos la tasa de incidencia es de 206 casos por 28612 personas-años.

Como ya se ha comentado antes, esto se interpreta como que en el grupo con esas características se esperan 206 casos por 28612 personas observadas en un año. O lo que es lo mismo, pero dicho de otra manera más conveniente para esta pregunta, se esperan 206 casos por $\frac{28612}{7} \cong 4087$ personas observadas en 7 años.

Así, ya tenemos que el número de casos nuevos en un periodo de 7 años es $I = 206$ y que el tamaño de la población observada inicial es $N_0 = 4087$. Substituimos en la fórmula y obtenemos:

$$CI(7) = \frac{206}{4087} = 0.05040372$$

Este resultado se puede interpretar diciendo que la probabilidad de que una persona no enferma, expuesta y de un rango de edad entre 55 y 64 años pueda enfermar es de 0,0504.

EJERCICIO 5

El paquete epitools contiene el *data frame* wcgs con datos de un estudio de cohorte cuyo fin era estudiar la relación entre enfermedades cardiovasculares y una serie de posibles factores de riesgo. Para obtener información sobre estos datos y para visualizarlos podéis ejecutar las siguientes instrucciones en R:

```
> library(epitools)
> ?wcgs
> data(wcgs)
> View(wcgs)
> summary(wcgs)
```

La variable chd69 indica la aparición (chd69 = 1) o ausencia (chd69 = 0) de una enfermedad cardiovascular mientras la variable time169 contiene los tiempos bajo riesgo (en días) para tal enfermedad.

a) Estimad, si es posible, las siguientes medidas de asociación entre una enfermedad cardiovascular y el tipo de comportamiento (variable dibpat0) e interpretad los valores obtenidos:

Tabla de valores en la base de datos:

	ENFERMEDAD		
COMPORTAMIENTO	SÍ (chd69 = 1)	NO (chd69 = 0)	TOTAL
SÍ (dibpat0 = 1)	178	1411	1589
NO (dibpat0 = 0)	79	1486	1565
TOTAL	257	2897	3154

*Datos sacados con el siguiente código: "`table(df$chd69, df$dibpat0)`".

- El riesgo relativo.

  Es la tasa del riesgo de padecer una enfermedad de la gente expuesta entre el riesgo de padecerla que tiene la gente no expuesta.

$$RR = \frac{P(D|E)}{P(D|\bar{E})}$$

  Y si substituimos en la fórmula los valores correspondientes, obtenemos:

$$RR = \frac{178/1589}{79/1565} = \boxed{2.219133}$$

  Esto es, el riesgo de padecer una enfermedad cardiovascular es 2,219 veces más alta entre la gente con un comportamiento XXX.

- La diferencia de riesgos.

  Es la diferencia entre el riesgo de padecer enfermedad entre los sujetos expuestos y los no expuestos.

$$RD = P(D|E) - P(D|\bar{E})$$

  En el presente caso:

$$RD = \frac{178}{1589} - \frac{79}{1565} = \boxed{0.06154087}$$

- El odds ratio.

  Es la probabilidad de enfermar entre las personas expuestas entre la probabilidad de enfermar entre las personas no expuestas.

$$OR = \frac{odds(D|E)}{odds(D|\bar{E})} \text{ donde } odds(D) = \frac{P(D)}{1-P(D)}.$$

Por lo que:

$$OR = \frac{P(D|E)/(1-P(D|E))}{P(D|\bar{E})/(1-P(D|\bar{E}))} = \frac{\frac{178}{1589}/\frac{1411}{1589}}{\frac{79}{1565}/\frac{1486}{1565}} = \frac{178 \cdot 1486}{1411 \cdot 79} = \boxed{2.372929}$$

Esto es, la probabilidad de tener una enfermedad cardiovascular es 2,373 veces mayor entre los sujetos con el comportamiento SI que entre los sujetos con el comportamiento NO.

Y, dado que el odds ratio también se puede escribir como:

$$OR = RR \cdot \frac{1-P(D|\bar{E})}{1-P(D|E)}$$

El riesgo relativo siempre es más cercano a 1 que el OR. En nuestro caso, al ser RR mayor a 1, ocurre que OR>RR>1 (2.373>2.219>1).

- El riesgo atribuible en la población.

Es la proporción de casos de enfermedad en la población que son atribuibles a la exposición.

$$PAR = P(E|D)\left(1-\frac{1}{RR}\right) = \frac{178}{257}\left(1-\frac{1}{2.219133}\right) = \boxed{0.3805}$$

- El riesgo atribuible entre los expuestos.

Es la proporción de casos con enfermedad entre los sujetos expuestos que es atribuible a la exposición.

$$EAR = \frac{P(D|E)-P(D|\bar{E})}{P(D|E)} = \frac{RR-1}{RR} = 1-\frac{1}{RR} = \boxed{0.5493736}.$$

b) Estimad las tasas de incidencia (en casos por 10 000 personas-años) y calculad los intervalos de confianza del 98% para ambos tipos de comportamiento.

Código utilizado:

```
casosexposats <- 0
tempsexposats <- 0
casosNoexposats <- 0
tempsNoexposats <- 0

for(i in 1:nrow(df)){
  if(df$chd69[i]  == 1){
    if(df$dibpat0[i]  == 1){
        casosexposats <- casosexposats+1
        tempsexposats<-tempsexposats + df$time169[i]
    }else{
      casosNoexposats <- casosNoexposats+1
      tempsNoexposats<-tempsNoexposats + df$time169[i]
    }
  }
}
taula <- matrix(c(casosexposats, casosNoexposats, tempsexposats, tempsNoexposats),2,2)
```

Tabla de datos:

Comportamiento	Nuevos casos (chd69 = 1)	Tiempo bajo riesgo
dibpat0=1	178	282075
dibpat0=0	79	143183

La tasa de incidencia se calcula como:

$$\hat{I}_r = \frac{I}{\Delta T}$$

Donde I es el número de casos nuevos i ΔT, el tiempo bajo riesgo. Substituimos para cada grupo según el tipo de comportamiento:

$$\hat{I}_{\text{dibpat0=1}} = \frac{I}{\Delta T} = \frac{178}{282075} = \boxed{0.0006310378}$$

$$\hat{I}_{\text{dibpat0=0}} = \frac{I}{\Delta T} = \frac{79}{173183} = \boxed{0.0005517415}$$

En casos por 10000 personas años serian: 6.31 casos para dibpat0=1 y 5.517 casos para dibpat0=0.

Ahora ya podemos calcular los intervalos de confianza:

$$CI\left(\hat{I}_{\text{dibpat0=1}}; 0{,}98\right) = 0.0006310378 \pm 2{,}325 \cdot \sqrt{\frac{0.0006310378}{282075}} = [0.0005210693, 0.0007410063]$$

$$CI(\hat{I}_{\text{dibpat0=0}}; 0{,}98) = 0.0005517415 \pm 2{,}325 \cdot \sqrt{\frac{0.0005517415}{173183}} = [0.00042051, 0.000682973]$$

Y para obtenerlo en casos por 10000 personas-años, multiplicamos el resultado por 10000 y obtenemos:

$$\boxed{CI\left(\hat{I}_{\text{dibpat0=1}}; 0{,}98\right) = [5.210693, 7.410063]}$$

$$\boxed{CI\left(\hat{I}_{\text{dibpat0=0}}; 0{,}98\right) = [4.2051, 6.82973]}$$

c) Usad la función rateratio del paquete epitools para calcular la razón entre ambas tasas de incidencia (*incidence rate ratio*) y el intervalo del 99% correspondiente. Interpretad los valores obtenidos.

Código utilizado:

```
library(epitools)
casos <- c(exposed = 178, unexposed = 79)
pyears <- c(exposed = 282075, unxposed = 173183)
rateratio(casos,pyears, conf.level=0.99)
```

Salida de resultados:

```
rate ratio with 99% C.I.
                 estimate     lower    upper
exposed   1.0000000        NA       NA
unexposed 0.8753873 0.6122409 1.231886
```

La razón entre ambas tasas de incidencia es

$$IRR = \frac{1.0000000}{0.8753873} = \boxed{1.142352}$$

Y el intervalo del 99% de confianza,

$$IC(IRR)_{0.99} = \left[\frac{1}{1.231886}, \frac{1}{0.6122409}\right] = \boxed{[0.8117634, 1.633344]}$$