Tarea 1 Econometría Aplicada

Nombre: Rafael **Fecha:** Miércoles 26 de agosto de 2020

Clave única: **Tema:** Estadística Descriptiva

Pregunta 1

(a) Media, varianza e IC de variable *pos*

Dado que la base tiene 49 registros sin información, se considera $n = 100 - 49 = 51$. Sea X_i la tasa de positividad $POS_i = \frac{Confirmed_i}{Tests_i}$ del país i

$$\bar{X} = 0.1120;\ S^2 = 0.0163;\ S = 0.1278$$

Con fundamento en el Teorema del Límite Central, se asume que $\bar{X}$ se comporta en el límite como una distribución normal: $\bar{X} \xrightarrow{d} \mathcal{N}(\mu, \frac{\sigma^2}{n})$ y el intervalo de confianza al 95% para μ es:

$$\bar{X} \pm z_{\frac{\alpha}{2}}(\frac{S}{\sqrt{n}}) = 0.1120 \pm 1.96(\frac{0.1278}{\sqrt{51}})$$

$\therefore \mu_X \in [0.0769, 0.1471]$ a un nivel de confianza de 95%

(b) Prueba de hipótesis tasa de positividad

Se quiere contrastar H_0 : La media de la tasa de positividad mundial es igual a la de México **vs** H_a : La media de la tasa de positividad mundial es menor a la de México. La tasa de positividad de México es igual a 0.5045

Se quiere probar la siguiente hipótesis:

$$H_0 : \mu_X = 0.5045 \quad vs \quad H_a :\ \mu_X < 0.5045$$

El estadístico t queda de la siguiente manera:

$$t = \frac{\sqrt{51}(0.1120 - 0.5045)}{0.1278} = -21.9180$$

El Valor-p es: $\Phi(-21.9180) = (1 - \Phi(21.9180)) \approx 0$, por lo que podemos rechazar H_0 para cualquier nivel de significancia.

Sí hay evidencia al α% para rechazar H_0 (a favor de H_a), es decir, podemos concluir que efectivamente México tiene una tasa de positividad mayor que la media mundial. O dicho de otra forma, podemos concluir que la media de la tasa de positividad mundial es menor a la tasa de positividad de México.

(c) Prueba de hipótesis tasa de fatalidad

Dado que sí tenemos información para todos los países sobre muertes y confirmados, ahora $n = 100$. Sea Y_i la tasa de fatalidad $CFR_i = \frac{Deaths_i}{Confirmed_i}$ del país i. Se quiere contrastar H_0 : La media de la tasa de fatalidad del SARS-CoV-2 es 4 veces menor que la del SARS-CoV **vs** H_a : La media de la tasa de fatalidad del SARS-Cov-2 **no** es 4 veces menor que la del SARS-CoV. Se prueba la siguiente hipótesis bilateral:

$$H_0 : \mu_Y = \frac{0.096}{4} = 0.024 \quad vs \quad H_a : \ \mu_Y \neq \frac{0.096}{4} = 0.024$$

El estadístico t queda de la siguiente manera:

$$t = \frac{\sqrt{100}(0.032 - 0.024)}{0.031} = 2.6022$$

El Valor-p es: $2^*(1\text{-}\Phi(2.6022)) = 0.0092$, por lo que para cualquier nivel de significancia $\alpha > 0.0092$ se puede rechazar H_0

(d) Relación IC - prueba de hipótesis

El intervalo de confianza al 99.08% para μ es:

$$\bar{Y} \pm z_{\frac{\alpha}{2}}(\frac{S}{\sqrt{n}}) = 0.032 \pm 2.604(\frac{0.031}{\sqrt{100}})$$

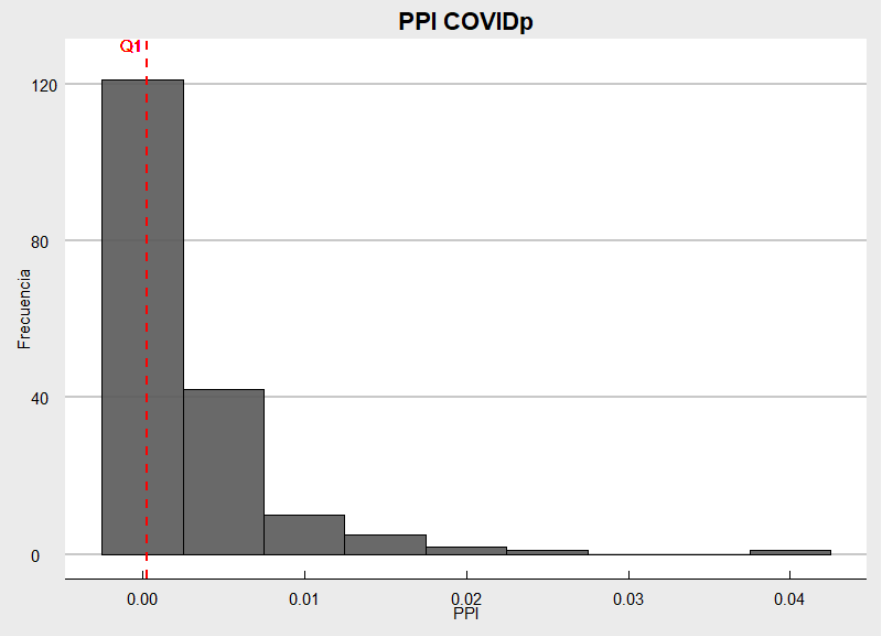
$\therefore \mu \in [0.024, 0.040]$ a un nivel de confianza de 99.08%

Para cualquier $\alpha > 0.0092$, el valor $\frac{0.096}{4} = 0.024$ ya no estará incluido en el intervalo de confianza. El valor-p es el mínimo valor de α para el cual los resultados de la prueba son significativos, es decir, indican que hay que rechazar H_0. Visto de otra forma, el valor-p consiste en calcular la probabilidad de obtener un valor del estadístico t tan extremo como el observado. También, conforme α aumenta, el intervalo de confianza se "hace más pequeño".

Pregunta 2

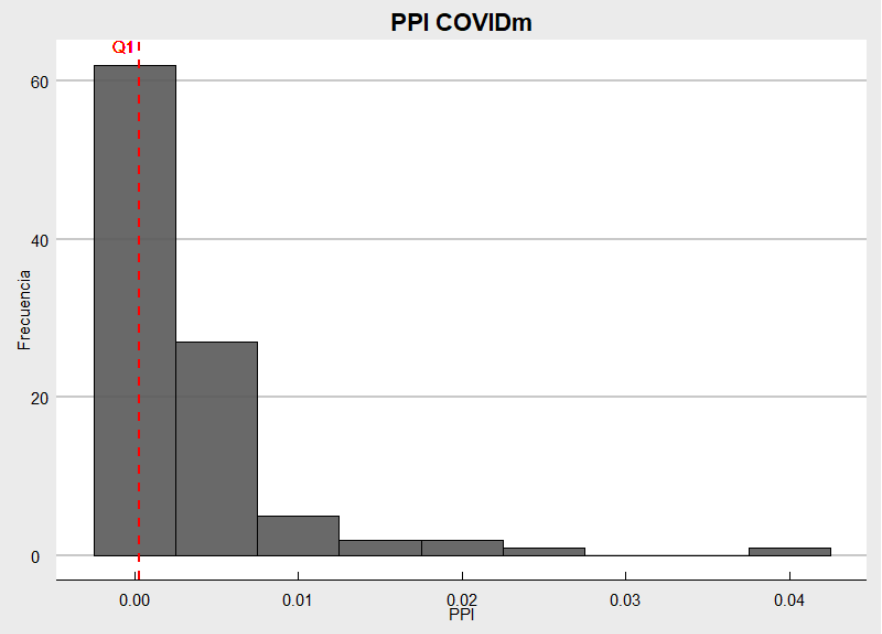
(a) Histograma COVIDp

El primer cuartil de la distribución poblacional de $PPI_i = \frac{Confirmed_i}{Population_i}$ es $Q1 = 0.000265 = 0.03\%$



(b) Histograma COVIDm

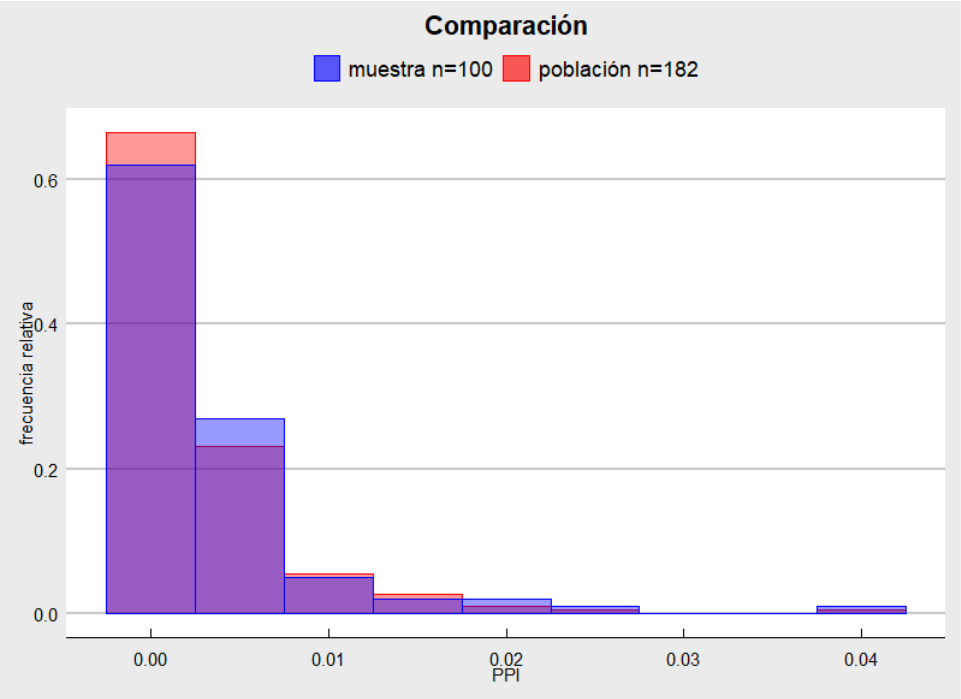
El primer cuartil de la distribución muestral es $Q1 = 0.000355 = 0.04\%$



(c) Relación entre histograma poblacional y muestral

A simple vista, se ven muy similares. Ambos histogramas tienen sesgo a la derecha y en ambos se observan 1 o pocos outliers (alrededor de $PPI = 0.04$)

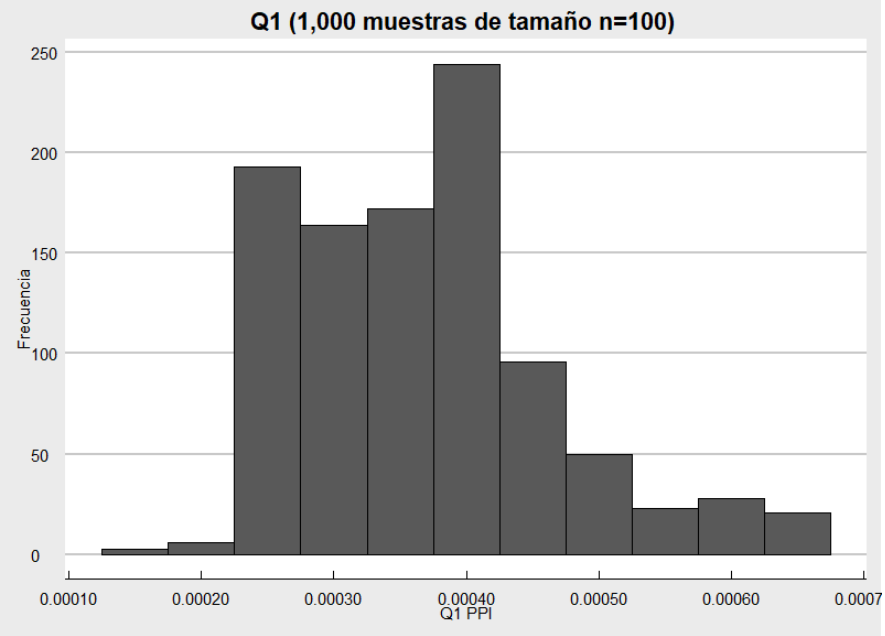
Ex-ante sí esperaba que los histogrmas se parecieran, debido al tamaño de la muestra, es 54% aproximadamente del tamaño de la población. Para comparar de mejor manera ambos histogramas, a continuación se presentan los histogramas con frecuencia relativa, se observa que son muy parecidos. Se podría pensar que la muestra sí es representativa de la población.



(d) Bootstrap

Para el muestreo aleatorio se usó el comando set.seed(5)

A continuación, se muestra el histograma obtenido para el primer cuartil del porcentaje de la población total que ha contraído el virus (PPI), calculado a partir de generar 1,000 submuestras (de tamaño n=100, con reemplazo) de la muestra original BaseCOVIDm.



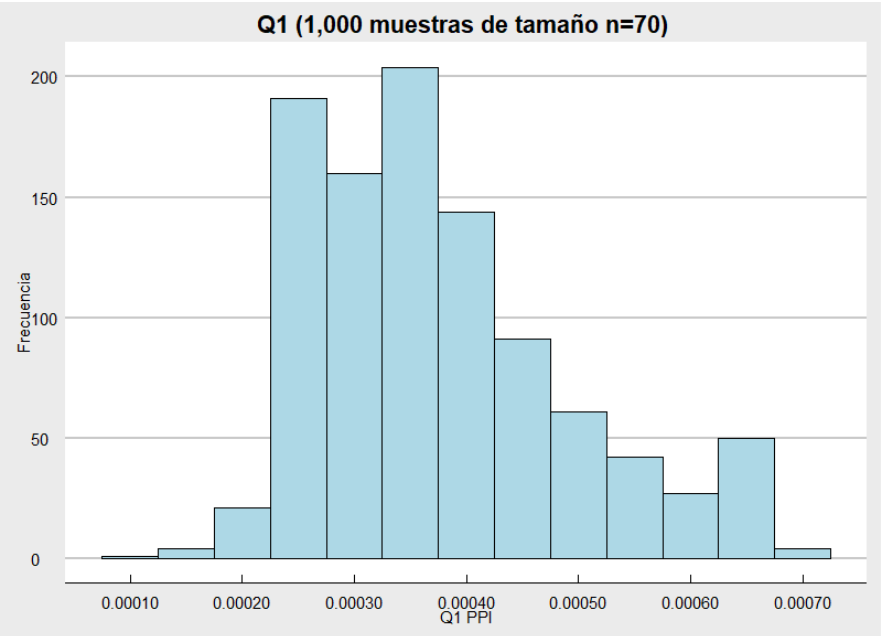
(e) Bootstrap con muestras de tamaño n=70

Primero, se presenta el histograma para el primer cuartil de la población total que ha contraído el virus (PPI), calculado a partir de generar 1,000 submuestras (de tamaño n=70, con reemplazo) de la muestra original BaseCOVIDm.

Dado que ahora $n = 70 < 100$ hay que hacer un ajuste de la varianza

$$Var(cuartil) = \frac{70}{100}\frac{1}{1000}\sum_{k=1}^{1000}(Q_k - E(Q))^2$$

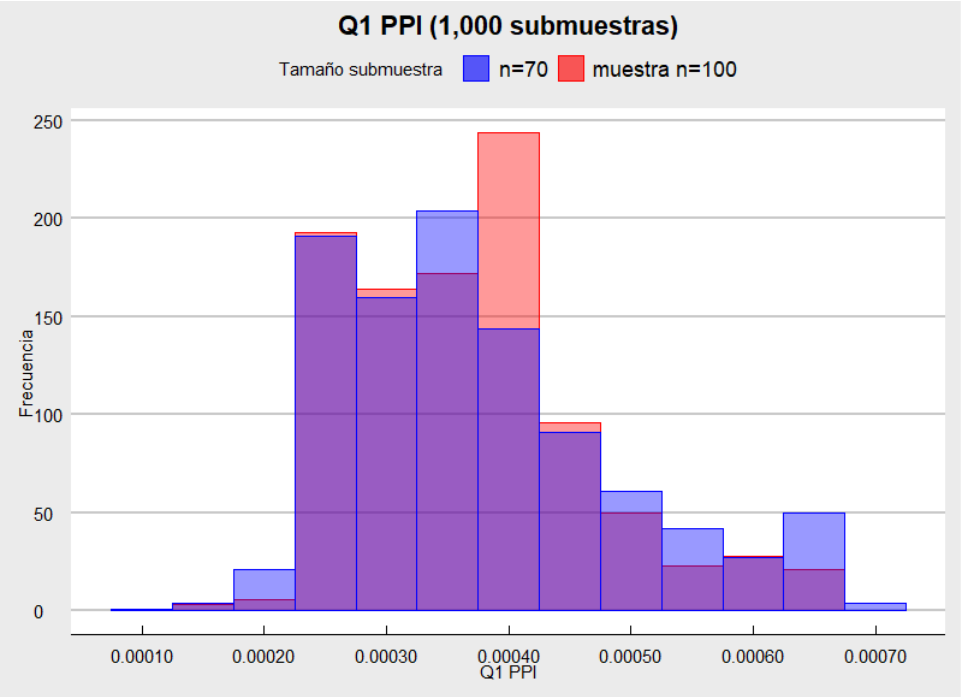
Esto será relevante a la hora de calcular intervalos de confianzas



A continuación, se superponen los histogramas del inciso (c) y (d)

En rojo está el histograma del Q1 para las 1,000 submuestras de tamaño n=100
En azul está el histograma del Q1 para las 1,000 submuestras de tamaño n=70

Llama la atención que el histograma azul da la impresión que tiene mayor varianza (las colas más pesadas)



(f) Intervalo de confianza para el primer cuartil

El primer cuartil poblacional de la variable PPI es $Q_1 = 0.0265\%$
Los pasos para obtener el intervalo de confianza fueron:

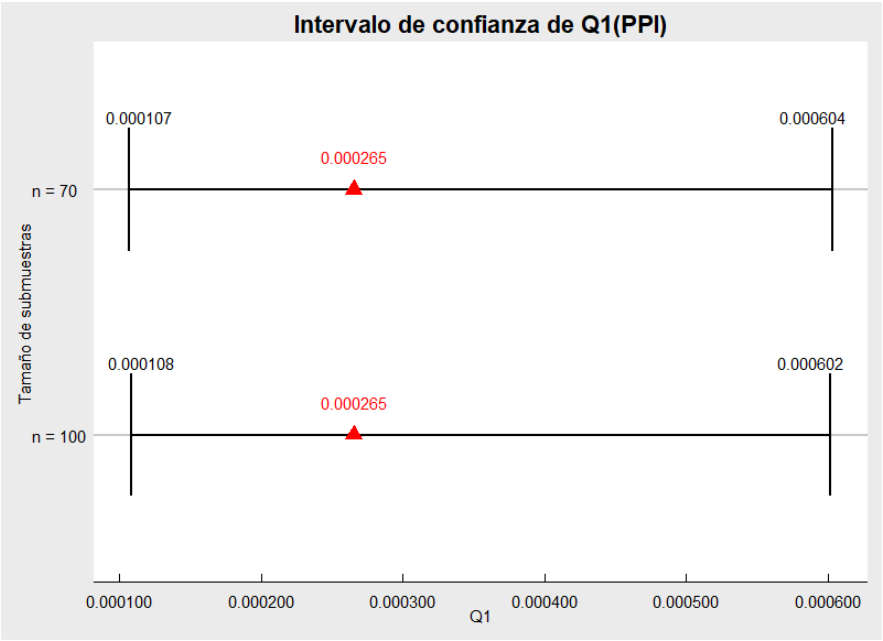
(1) Para cada submuestra $i = 1, \ldots, 1000$ se calculó el primer cuartil $Q1_i^*$ (incisos anteriores)

(2) Se calculó la varianza (y desviación estándar) del vector creado con los valores $Q1_i^*$. Dicha varianza es $Var(Q1)_{bootstrap}$

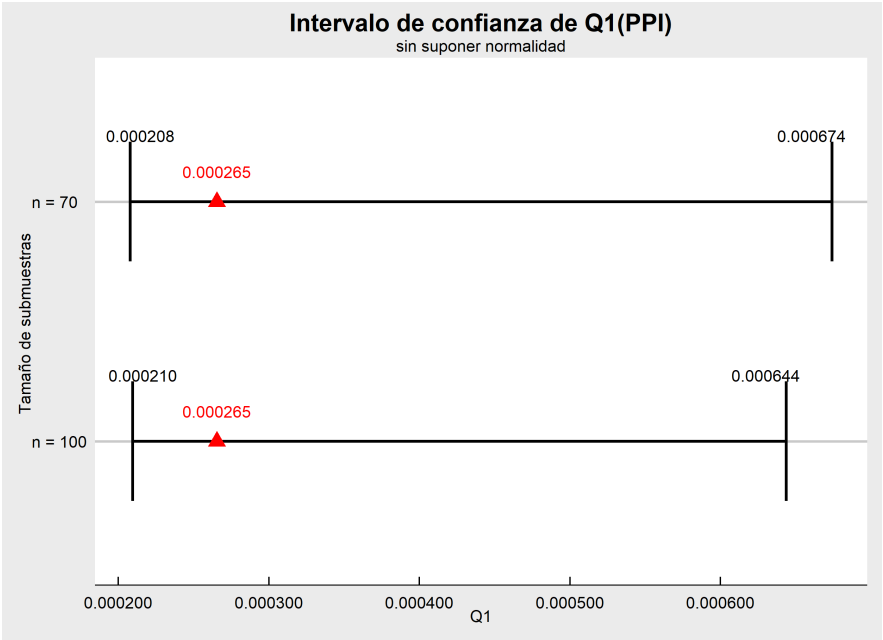
(3) El intervalo de confianza queda de la forma:

$$Q1_{muestra} \pm z_{0.995}\sqrt{Var(Q1)_{bootstrap}}$$

A continuación se muestran gráficamente los intervalos de confianza del primer cuartil con el método bootstrap para 1,000 submuestras de tamaño $n = 100$ y $n = 70$. Llama la atención que los IC son prácticamente los mismos

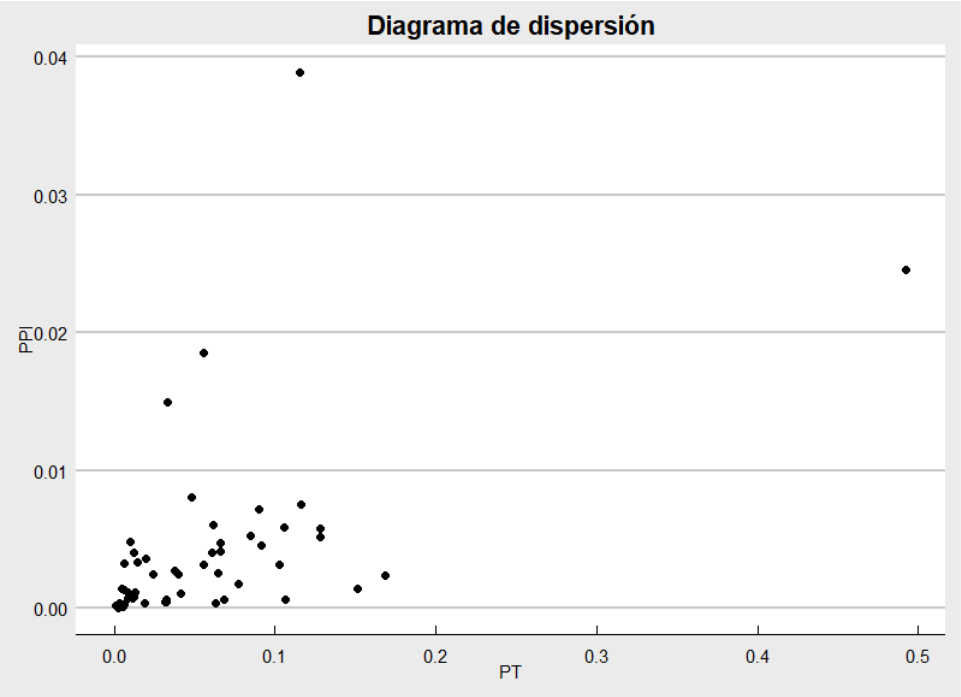


Ahora, sin suponer normalidad (de la distribución de los cuartiles de las 1,000 submuestras) se muestran los mismos intervalos de confianza:

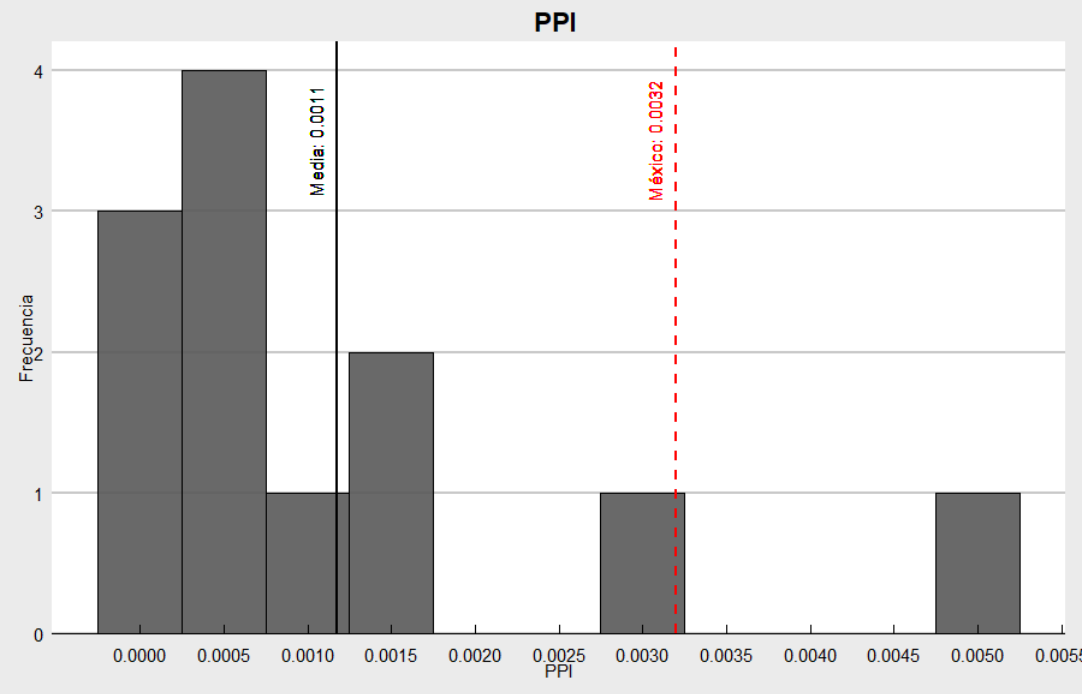


Pregunta 3

(a) Diagrama de dispersión



(b) Histograma

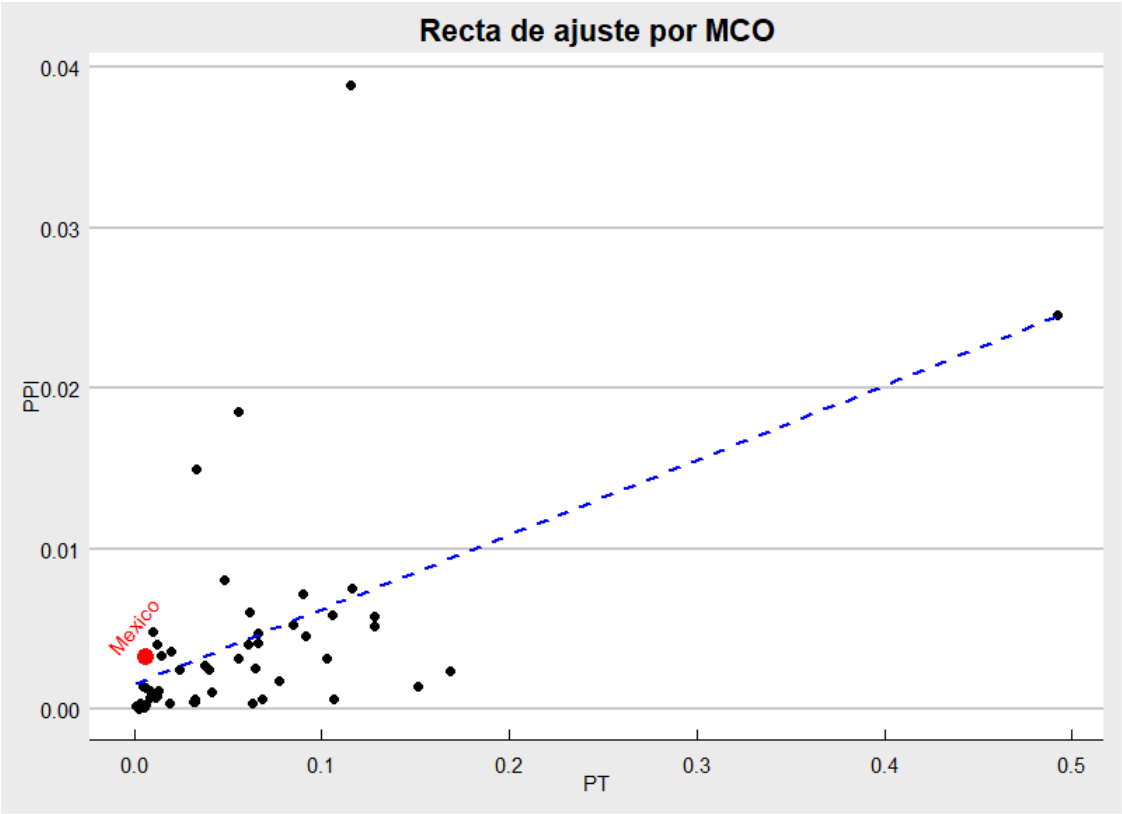


(c) MCO

El histograma nos indica que México tiene un porcentaje de personas infectadas

alto, si lo comparamos con países que aplican un número similar de tests ($PT_{Mex} \pm 0.005$). El PPI de México es mayor a la media de países "similares".
El diagrama de dispersión con la recta de ajuste nos dice algo similar. Para un valor dado del número de pruebas PT, México está por encima del promedio de porcentaje de personas infectadas PPI

La regresión sugiere que a mayor número de tests hay más personas confirmadas con COVID-19.



(d) En el histograma estamos viendo cómo está México en PPI con relación a países similares (PT), mientras que en la regresión vemos explícitamente la relación lineal positiva que existe entre PT y PPI

(e) Habríamos obtenido la función de regresión poblacional, ya no tendríamos que hacer algún proceso de inferencia estadística como normalmente se lleva a cabo con las muestras. La muestra siempre es un subconjunto de la población

Además, los estadísticos son funciones de los datos de la muestra, por los que estos pueden variar dependiendo de la muestra que tomemos, la regresión arrojará resultados distintos si usaramos otra muestra. Se supone que la línea de regresión muestral representa a la regresión poblacional, pero debido a fluctuaciones muestrales, son en el mejor de los casos, sólo una aproximación de la verdadera

regresión poblacional.

Pregunta 4 Los países que no poseen la infraestructura mínima necesaria para poder llevar un registro del transcurso de la epidemia podrían ser los países subrepresentados, yo pensaría que los países más pobres (en América Latina, África o el Sudeste Asiático) o que tienen un gobierno autoritario (que no es transparente y no publica información p.ej. Corea del Norte).

Pienso que lo primero que se debe de hacer antes de hacer cualquier análisis estadístico, es definir bien cuál es la población relevante que queremos estudiar/analizar y después recolectar una muestra que permita hacer inferencias sobre dicha población de interés. En este caso podríamos decir que nuestra *población* son los 182 países. Viéndolo de otra forma, si la población fueran todos los países del mundo (195), entonces los 182 países serían una muestra

Es fundamental definir qué es la población, en la tarea la población fueron los 182 países **que hacen pública su información**, pero igual podríamos definir población como "países africanos" o "países cuyo índice de desarrollo humano es superior a x" o "países con gobiernos populistas" o "países en europa" dependiendo de qué es lo que queramos analizar.

¿Cómo se vería reflejado esto en nuestro estimador y valor estimado muestral?
No sé si entendí bien la pregunta, pero la muestra no cambiaría (seguiría siendo BaseCovidm), pienso que lo que cambiaría son las conclusiones/inferencias que haríamos sobre la población.
