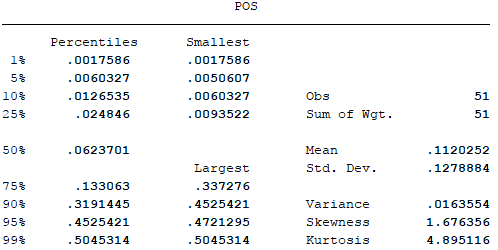
**Tarea 1**

1. Utilizando la base BaseCOVIDm realiza lo siguiente:
2. Calcula la tasa de positividad acumulada para COVID-19 de cada país (POSi). Calcula la media y varianza muestral. Calcula un intervalo de confianza al 95% para la media.

R= La media muestral es 0.1120252 y la varianza muestral es 0.0163554, por lo que un intervalo de confianza al 95% para la media es [0.0769256, 0.1417125], asumiendo una distribución normal.

El valor del intervalo que arroja STATA para una distribución t, es [0.0760561, 0.1479944]. Ambos intervalos muy similares.

1. Dado el número reducido de pruebas que México ha hecho, te planteas la hipótesis de que México debe tener una tasa de positividad mayor que la media mundial. Plantea la prueba de hipótesis relevante para evaluar esta afirmación. Reporta el valor-p e indica que concluyes con respecto a la prueba hipótesis planteada.

R= La tasa de positividad de México (POSMEX)=0.5045314. Esta tasa es observable en la muestra. Por otra parte, la media mundial es desconocida, ya que no tenemos la tasa de positividad para todos los países, por lo tanto, la prueba de hipótesis relevante es:

Ho:

H1:

El estadístico t asociado a esta prueba es t=-21.918005; por lo tanto, el valor-p tiende a 0 y Ho se rechaza en favor de H1.

1. Durante finales del año 2002 comenzó en diversas provincias de China el brote de SARS-CoV, que guarda similitudes con el SARS-CoV-2, virus responsable de la pandemia actual. Nos interesa comparar la tasa de fatalidad (CFR) de ambas pandemias. Un artículo periodístico indica que la tasa de fatalidad media del SARS-CoV es cuatro veces mayor que la del SARS-CoV-2. Plantea la prueba de hipótesis relevante para evaluar dicha afirmación y construye el estadístico t que necesitarás para realizar dicha prueba. Asume que la media de la tasa de fatalidad del SARS-CoV es conocida (de acuerdo a la OMS) e igual a 9.6%

R= Si es cierta la afirmación del artículo periodístico, entonces ; por lo tanto, la prueba de hipótesis relevante para evaluar dicha afirmación es:

Ho:

H1:

Dado que CFRcov2=0.0320818

S2=0.0009645

n= 100 (hay información de esta variable para las 100 observaciones de la muestra)

1. Dada la media de la tasa de fatalidad del SARS-CoV (9.6%), ¿a partir de qué nivel de confianza, el intervalo de confianza relevante para nuestra prueba de hipótesis ya no incluiría el valor de ? Describe la relación entre tu respuesta y el valor-p.

R= El intervalo de confianza relevante es

Donde VC es el valor crítico.

Aquí debe evaluarse el límite inferior del intervalo, para ver si éste incluye al valor señalado.

Con los valores que tenemos:

=0.024

Por lo que VC=2.6023

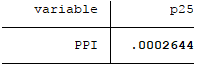
Valor-p= 0.00466\*(2) =0.00932 (para la prueba de dos colas)

Nivel de confianza=1-(0.00932) = 0.99068

En nivel de significancia obtenido es el complemento del valor-p, debido a que el nivel de significancia es y el valor-p es el más pequeña para la cual se rechaza Ho, lo que en términos del intervalo de confianza significa que el intervalo no contiene a de la hipótesis nula.

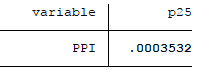
1. Estamos interesados en conocer el porcentaje de la población total de cada país que ha contraído el virus (PPI), por ello creamos la variable PPIi.
2. Utilizando la base BaseCOVIDp construye un histograma de la variable PPI y calcula el primer cuartil de su distribución.

R= El primer cuartil de la distribución es 0.0002644



1. Utilizando la base BaseCOVIDm construye un histograma de dicha variable y calcula el primer cuartil.

 R= El primer cuartil de la distribución es 0.0003532



1. ¿Cuál es la relación que existe entre los histogramas de los incisos anteriores? ¿Ex-ante esperabas que los histogramas se parecieran?

R= Los histogramas que se construyeron utilizando ambas bases de datos (completa de países y muestra de países) son muy parecidas.

Ex ante esperaba que las gráficas se parecieran, ya que la selección de los países que conforman la BaseCOVIDm se llevó a cabo de manera aleatoria, lo que implica que se trata de una muestra independiente e idénticamente distribuida.

1. Utilizando el método Bootstrap, genera 1,000 submuestras del tamaño de la muestra original (n = 100) partiendo de la base BaseCOVIDm. Para cada submuestra, calcula el primer cuartil y grafica un histograma de los 1,000 cuartiles estimados.

R= La gráfica de los cuartiles estimados a partir de las submuestras generadas con el método Bootstrap de tamaño n=100, es la que aquí se muestra.

1. Repite el inciso anterior pero ahora utilizando 1000 submuestras de tamaño 70. Utilizando una gráfica compara el histograma del inciso anterior con el histograma que se produce en este inciso. ¿Qué similitudes y diferencias notas?

R= La gráfica de los cuartiles estimados a partir de las submuestras generadas con el método Bootstrap de tamaño n=70, es la que aquí se muestra.

Ambas gráficas son muy parecidas en cuanto a que sus valores se centran entre .0002 y .0004.

1. Utilizando los resultados de los dos incisos anteriores construye un intervalo de confianza del 99% para el primer cuartil, grafícalos y ubica en dicha gráfica el valor del primer cuartil poblacional.

**Con n=100**

Ybarra=0.0003637

A partir de la fórmula

Se obtiene de forma empírica la Var=0.00000916

El intervalo de confianza del 99% resultante es [0.0001171943, 0.0006102]

**Con L=70**

Ybarra=0.0003776

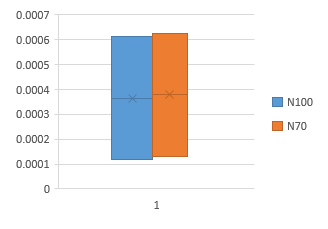
A partir de la fórmula

L=70

Se obtiene de forma empírica la Var=0.0000000092

El intervalo de confianza del 99% resultante es [0.0001310943, 0.0006241]

Dado que primer cuartil poblacional P25=0 .0003532, este se encuentra dentro de ambos intervalos



1. A lo largo de los meses que ha durado la pandemia han surgido diversas preguntas. Una de ellas se enfoca en la relación que existe entre el porcentaje de personas a las que se les han realizado pruebas PT y el porcentaje de personas contagiadas PPI. Utilizando los datos de BaseCOVIDm contestaremos las siguientes preguntas:

a) Realiza un diagrama de dispersión (*scatterplot*) utilizando la variable PT en el eje X y PPI en el eje

R= A partir de la BaseCOVIDm, que contiene 100 observaciones, se obtuvo el siguiente diagrama de dispersión.

b) Explora cuántos países tienen un valor PT 0.005 mayor o menor al valor de PTmex. Utilizando solo estos países, grafica un histograma de la variable PPI. Ubica en dicho gráfico el valor de la media de PPI (solo utilizando este conjunto de países) y ubica el valor de la variable PPI correspondiente a México. ¿Está México por encima o por debajo de la media de estos países?

R= 12 países se encuentran en ese rango de valores para la variable PT y México (línea verde) está por encima de la media para estos países (línea roja).

c) ¿Está México por encima o por debajo de la recta? ¿Cómo se relaciona esta respuesta con la

pregunta del inciso anterior?

R: México se encuentra por encima de la recta, lo cual es consistente con la respuesta del inciso anterior, en la que se verificó que México se encontraba por encima de la media condicional de los países similares a México en términos del porcentaje de personas a las que se les han realizado pruebas.

d) Describe brevemente (menos de 200 palabras) qué similitudes y diferencias tienen ambas estrategias.

La estrategia b nos proporciona información de lo que ocurre con la variable PPI en un rango acotado de valores de la variable PT; los datos dicen que, en promedio, países con similar proporción de pruebas realizadas han tenido menor porcentaje de personas contagiadas que México, ¿debería México realizar más pruebas si quiere prevenir contagios?

La estrategia anterior no nos permite identificar si al menos existe una relación entre el número de pruebas y el porcentaje de personas contagiadas, mientras que la estrategia c sí nos permite identificar dicha relación. Sin embargo, si continuáramos con la estrategia b para otros rangos de la variable PT, y graficáramos la media de PPI para dichos rangos, obtendríamos información similar a la que obtenemos con la estrategia c.

Finalmente, ambas estrategias difieren en que la estrategia c busca modelar la relación que existe entre ambas variables, mientras que la estrategia b es más útil para un análisis descriptivo-comparativo.

e) Describe brevemente (menos de 200 palabras) qué diferencia teórica hubiera hecho utilizar la base BaseCOVIDp para contestar las preguntas.

Dado que la BaseCOVIDp contiene a la población (total de países para los cuales hay información), el modelo b no cambiaría significativamente, puesto que su utilidad es descriptiva y nos ayuda a establecer comparaciones entre países, más que ayudarnos a modelar la relación entre PPI y PT.

En cuanto al modelo c, en lugar de estimar la como un parámetro, encontraríamos la verdadera relación que existe entre PPI y PT; además de que el modelo no tendría utilidad predictiva porque ya conoceríamos lo que ocurre con toda la población de países.

1. Considerando que en el mundo hay un total de 195 países y que la base poblacional que se proporcionó (BaseCOVIDp) tiene información de únicamente 182, ¿qué países pueden estar subrepresentados en la base muestral?, ¿cómo afecta esto a la definición de nuestros parámetros poblacionales?,¿cómo se verá reflejado esto en nuestro estimador y en el valor estimado muestral?

El hecho de que solo exista información para 182 de los 195 países probablemente implica que la causa de ello no es aleatoria, por lo que podrían estar subrepresentados países cuyos gobiernos tienen algún interés particular en evitar que las cifras de la pandemia se conozcan, debido a que reflejarían un mal manejo de ella. También podrían estar subrepresentados países pequeños, en los que posiblemente la pandemia no ha tenido una alta incidencia y, por lo tanto, no existe un gran interés por reportar cifras.

Si este es el caso, probablemente la muestra solo sea representativa de la población que incluye los 182 países y no puedan extrapolarse los resultados de las estimaciones al resto de países.