**Trabajo Práctico en R**

**Segunda Parte**

**Alumno: ……………………………**

**Registro: …………………………..**

**Ejercicio 12**

**La distribución Normal**

Este ejercicio es una adaptación del incluido en el libro OpenIntro Statistics. La versión original y en inglés del ejercicio se puede consultar en el siguiente vínculo:

<http://htmlpreview.github.io/?https://github.com/andrewpbray/oiLabs-base-R/blob/master/normal_distribution/normal_distribution.html>

En este ejercicio se analiza la distribución de probabilidad más relevante en Estadística: la **distribución Normal**. Utilizaremos datos de dimensiones del cuerpo humano. En particular, se utilizará un set de datos que contiene información de 247 hombres y 260 mujeres, todos considerados adultos saludables.

La base tiene 25 variables, sobre las que se puede encontrar una descripción en el siguiente link: <https://www.openintro.org/stat/data/bdims.php>

Aquí sólo nos centraremos en la altura en centímetros y el sexo.

Comencemos construyendo un histograma de la altura de hombres y otro para las mujeres:

***Insertar gráfico aquí***

¿Qué diferencias observa entre hombres y mujeres?

***Insertar respuesta aquí***

Notar que ambas distribuciones tienen forma de campana y son aproximadamente simétricas. ¿Podemos decir que siguen distribuciones aproximadamente normales? Para una primera aproximación, graficaremos encima de cada histograma una distribución Normal. Esta distribución Normal deberá tener la misma media y desvío standard que los datos originales.

Empecemos con el caso de las mujeres:

***Insertar gráfico aquí***

Basado en este gráfico, ¿considera que la altura de las mujeres sigue una distribución aproximadamente Normal?

***Insertar respuesta aquí***

Veamos el caso de los hombres

***Insertar gráfico aquí***

Basado en este gráfico, ¿considera que la altura de las mujeres sigue una distribución aproximadamente Normal?

***Insertar respuesta aquí***

Evaluar a ojo la forma del histograma para ver si los datos se distribuyen de manera Normal puede ser útil, pero no siempre es posible determinar qué tan buena es la aproximación. Un camino alternativo consiste en construir un gráfico "QQ", nombre que proviene de Quantile-Quantile, (Quantile significa cuantil o percentil). El gráfico consiste en comparar si los percentiles de la distribución que estamos analizando se corresponden con los de una distribución Normal.

A modo de ejemplo, el primer cuartil (percentil 25) de las alturas de las mujeres de la base de datos es:

***Insertar respuesta aquí***

Por su parte, si la población fuera exactamente normal, con media y desvíos iguales a los de nuestra base de datos, el primer cuartil (percentil 25) debería ser:

***Insertar respuesta aquí***

Notará que son valores bastante parecidos. El gráfico QQ consiste en realizar esta comparación para todos los percentiles.

Veamos el gráfico para la altura de las mujeres

***Insertar gráfico aquí***

Si la distribución fuera exactamente Normal, cada circulito aparecería exactamente sobre la línea recta. Si bien en este caso la correspondencia no es exacta, podemos ver que es una muy buena aproximación, salvo en las colas de la distribución.

Veamos el gráfico QQ para la altura de los hombres:

***Insertar gráfico aquí***

Al igual que para el caso de las mujeres, podemos concluir que la distribución Normal es una buena aproximación de la distribución de la altura de los hombres. De esta forma, podemos utilizar todo lo que sabemos de la distribución Normal para responder preguntas sobre la altura de las personas.

A partir de ahora nos enfocaremos en valores específicos de nuestro conjunto de datos y veremos si la distribución normal produce estimaciones razonables de las probabilidades.

Suponiendo una distribución Normal, ¿Cuál es la probabilidad de que una mujer adulta seleccionada al azar mida más de 170 cm.?

***REALICE EL CÁLCULO ANALÍTICAMENTE, UTILICE LUEGO R PARA COMPROBAR LA RESPUESTA. Suponga una media de 164,87 cm. y un desvío de 6,5446 cm.***

***Insertar respuesta aquí***

En la base de datos utilizada, ¿cuál es la proporción de mujeres que mide más de 170 cm.?

***Insertar respuesta aquí***

¿La proporción es cercana a la estimada mediante la distribución Normal? Interprete.

***Insertar respuesta aquí***

Suponiendo una distribución Normal, ¿Cuál es la probabilidad de que un hombre adulto seleccionado al azar mida más de 170 cm.?

***REALICE EL CÁLCULO ANALÍTICAMENTE, UTILICE LUEGO R PARA COMPROBAR LA RESPUESTA. Suponga una media de 177,75 cm. y un desvío de 7,1836 cm.***

***Insertar respuesta aquí***

En la base de datos utilizada, ¿cuál es la proporción de hombres que mide más de 170 cm.?

***Insertar respuesta aquí***

¿La proporción es cercana a la estimada mediante la distribución Normal? Interprete.

***Insertar respuesta aquí***

**Ejercicio 13**

**Aproximación de la Binomial por la Normal**

En la primera parte del TP se lanzó una moneda equilibrada 1 millón de veces, obteniéndose proporciones de caras y cecas cercanas al 50%-50%. En este ejercicio analizaremos qué tan probable es obtener resultados distintos a estos.

En principio, la cantidad de resultados "cara" en 1 millón de lanzamientos puede ser cualquier número entre 0 y 1 millón. Dado que la probabilidad de que se obtenga "cara" en un lanzamiento en particular es la misma en todos los lanzamientos, la distribución de probabilidades de la cantidad de caras obtenidas se puede representar como una variable Binomial con parámetros n=1000000 y p=0.5, ya que la moneda es equilibrada. Sin embargo, dado el alto valor de 'n', puede ser dificultoso calcular la distribución binomial. Como se vio en clase, la Distribución Normal puede dar una buena aproximación de la distribución binomial cuando se cumple que np>10 y nq>10.

¿Se cumplen los supuestos para el caso de moneda equilibrada?

***REALICE EL CÁLCULO ANALÍTICAMENTE, UTILICE LUEGO R PARA COMPROBAR LA RESPUESTA.***

***Insertar respuesta aquí***

¿Cuál es la media y el desvío de la distribución normal utilizada para aproximar la distribución de la cantidad de caras obtenidas con la moneda equilibrada?

***REALICE EL CÁLCULO ANALÍTICAMENTE, UTILICE LUEGO R PARA COMPROBAR LA RESPUESTA.***

***Insertar respuesta aquí***

Realizar el gráfico de esta distribución

***Insertar gráfico aquí***

¿Cuál es la probabilidad de que se obtenga más de un 50,3% de caras? Es decir, más de 503000 caras del millón de lanzamientos.

Utilice la distribución Normal para aproximar la binomial e interprete los resultados.

***Insertar respuesta aquí***

Si hubiese podido calcular la distribución Binomial, por ejemplo, usando R, ¿Cuál sería la probabilidad exacta de que salgan más de 503000 caras del millón de lanzamientos

***Insertar respuesta aquí***

¿Cuál es la diferencia entre la probabilidad aproximada por la normal y la exacta?

***Insertar respuesta aquí***

¿Considera que la aproximación mediante la Normal es razonable?

***Insertar respuesta aquí***

Sin embargo, recuerde que cuando se aproxima una variable discreta (binomial) mediante una variable continua (normal) es recomendable aplicar un "factor de corrección". En este caso, al hacer la aproximación con la normal tendríamos que haber calculado la probabilidad de que la variable sea mayor a 503000 + 0.5.

¿Cuál es esta probabilidad?

***Insertar respuesta aquí***

Una vez aplicado el factor de corrección ¿Cuál es la diferencia entre la probabilidad aproximada por la normal y la exacta?

***Insertar respuesta aquí***

La aplicación del factor de corrección, ¿mejoró o empeoró la aproximación?

***Insertar respuesta aquí***

¿Cuál es la probabilidad de que se obtenga menos de un 50,1% de caras? Es decir, menos de 501000 caras del millón de lanzamientos

Utilice la distribución Normal para aproximar la binomial, probando tanto con o sin el factor de corrección para comparar los resultados. Interprete los resultados.

***Insertar respuesta aquí***

Si hubiese podido calcular la distribución Binomial, por ejemplo, usando R, ¿Cuál sería la probabilidad exacta de que salgan menos de 501000 caras del millón de lanzamientos

***Insertar respuesta aquí***

¿Cuál es la diferencia entre la probabilidad exacta y ambas aproximaciones?

***Insertar respuesta aquí***

La aplicación del factor de corrección, ¿mejoró o empeoró la aproximación?

***Insertar respuesta aquí***

¿Cuál es la probabilidad de que se obtenga entre 49,99% y 50,01% de caras? Es decir, entre 499900 y 500100 caras en el millón de lanzamientos.

Utilice la distribución Normal para aproximar la binomial, probando tanto con o sin el factor de corrección para comparar los resultados. Interprete los resultados.

***Insertar respuesta aquí***

Si hubiese podido calcular la distribución Binomial, por ejemplo, usando R, ¿Cuál sería la probabilidad exacta de que salgan entre 499900 y 500100 caras del millón de lanzamientos

***Insertar respuesta aquí***

¿Cuál es la diferencia entre la probabilidad exacta y ambas aproximaciones?

***Insertar respuesta aquí***

La aplicación del factor de corrección, ¿mejoró o empeoró la aproximación?

***Insertar respuesta aquí***

**Ejercicio 14**

**Fundamentos de la inferencia estadística: Distribuciones Muestrales**

Este ejercicio es una adaptación del incluido en el libro OpenIntro Statistics. La versión original y en inglés del ejercicio se puede consultar en el siguiente vínculo:

<http://htmlpreview.github.io/?https://github.com/andrewpbray/oiLabs-base-R/blob/master/sampling_distributions/sampling_distributions.html>

El objetivo es analizar cómo los estadísticos de una muestra aleatoria pueden servir como estimaciones puntuales de parámetros poblacionales. En particular, interesa formular una **distribución muestral** de la estimación para aprender más acerca de las propiedades de la misma.

Se utilizarán datos del mercado inmobiliario de la ciudad de Ames, Iowa, EEUU. En particular, nos enfocaremos en todas las ventas de hogares ocurridas en esa ciudad entre 2006 y 2010. Este conjunto de datos representa nuestra **población** de interés. Mediante pequeñas muestras de esta población, intentaremos aprender más sobre estas transacciones inmobiliarias.

La base de datos contiene una gran cantidad de variables. Para empezar, sólo nos centraremos en una de ellas: el área del hogar en pies cuadrados, que la convertiremos a metros cuadrados para facilitar su interpretación.

Daremos una mirada rápida a la distribución del área de los hogares calculando los principales indicadores y construyendo el histograma:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mínimo** | **Q1** | **Mediana** | **Q3** | **Máximo** | **Media** |
| COMPLETAR | COMPLETAR | COMPLETAR | COMPLETAR | COMPLETAR | COMPLETAR |

***Insertar gráfico aquí***

En este ejercicio tenemos acceso a la población completa, pero esto rara vez sucede en la vida real. Recolectar información sobre el total de la población suele ser extremadamente costoso o incluso imposible. Por este motivo, se suele tomar una muestra de la población y utilizarla para entender las propiedades de la misma.

Si estamos interesados en estimar el área promedio de los hogares en Ames, podemos empezar tomando una muestra aleatoria de 50 hogares.

En base a esta muestra, nuestra mejor estimación puntual sobre el área promedio de los hogares es el promedio de la muestra de 50 hogares.

Tome una muestra de 50 hogares y calcule el área promedio de esa muestra.

***Insertar respuesta aquí***

Dependiendo de cuáles fueron los 50 hogares incluidos en la muestra, el promedio estimado estará un poco por encima o por debajo del verdadero promedio de la población. Sin embargo, la media muestral suele resultar un buen estimador del tamaño promedio de los hogares, dando un resultado no muy diferente al real y requiriendo para su cálculo menos del 3% de los hogares que conforman la población.

Tomemos otra muestra aleatoria, también de 50 hogares. ¿El promedio da muy distinto al de la primera muestra?

***Insertar respuesta aquí***

Supongamos que tomamos dos muestras más, una de tamaño 100 y otra de 1000. ¿Cuál cree que dará una estimación más precisa de la media poblacional? Justifique.

***Insertar respuesta aquí***

No debería sorprender que, cada vez que se toma una muestra aleatoria, se obtiene un área promedio diferente. Es útil tomar dimensión de cuánta variabilidad debemos esperar cuando estimamos la media poblacional mediante medias muestrales. La distribución de los promedios muestrales, denominada **distribución muestral**, nos puede ayudar a entender esta variabilidad.

En este ejemplo, como tenemos acceso al total de la población, podemos construir la distribución muestral repitiendo los pasos anteriores muchas veces. En particular, vamos a generar 5000 muestras de 50 hogares y calcular la media para cada una de ellas.

Veamos el histograma de las 5000 medias muestrales:

***Insertar gráfico aquí***

Aproximadamente, ¿cuál es el centro de esta distribución?

***Insertar gráfico aquí***

¿Qué esperaría que ocurra si, en lugar de tomar 5000 muestras, toma 50000? Para comprobarlo, tome 50000 muestras, calcule la media muestral de cada una de ellas y presente el histograma de esta las mismas.

***Insertar gráfico aquí***

¿Se modificó algo en el gráfico?

***Insertar respuesta aquí***

La distribución muestral que acabamos de computar nos dice mucho acerca de la estimación del área promedio de los hogares en Ames. Como **la media muestral es un estimador insesgado**, la distribución muestral está centrada en el verdadero promedio de la población, mientras que la dispersión de la distribución indica cuánta variabilidad debemos esperar al tomar muestras de sólo 50 hogares.

Para tomar conciencia de lo anterior, construiremos dos distribuciones muestrales adicionales, una basada en muestras de 10 hogares y la otra basada en muestras de 100 hogares.

Grafiquemos las 3 distribuciones juntas para ver las diferencias:

***Insertar gráfico aquí***

Cuando aumenta la cantidad de hogares, ¿qué sucede con el centro de la distribución? ¿y con la dispersión?

***Insertar respuesta aquí***

Repitamos el ejercicio pero ahora considerando el precio de las viviendas, expresado en dólares.

Veamos indicadores e histograma de los precios:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mínimo** | **Q1** | **Mediana** | **Q3** | **Máximo** | **Media** |
| COMPLETAR | COMPLETAR | COMPLETAR | COMPLETAR | COMPLETAR | COMPLETAR |

***Insertar gráfico aquí***

Tomamos una muestra de 50 hogares y calculamos el precio promedio. ¿Cuál es el promedio muestral?

***Insertar respuesta aquí***

Tomemos otra muestra aleatoria, también de 50 hogares. ¿Da muy distinto a la primer muestra?

***Insertar respuesta aquí***

Vamos a generar 5000 muestras de 50 hogares y calcular el precio promedio para cada una de ellas. Veamos el histograma de las 5000 medias muestrales:

***Insertar gráfico aquí***

Aproximadamente, ¿cuál es el centro de esta distribución?

***Insertar respuesta aquí***

Construiremos dos distribuciones muestrales adicionales, una basada en muestras de 10 hogares y la otra basada en muestras de 100 hogares.

Grafiquemos las 3 distribuciones juntas para ver las diferencias:

***Insertar gráfico aquí***

Cuando aumenta la cantidad de hogares, ¿qué sucede con el centro de la distribución? ¿y con la dispersión?

***Insertar respuesta aquí***

**Ejercicio 15**

**Fundamentos de la inferencia estadística: Intervalos de Confianza**

Este ejercicio es una adaptación del incluido en el libro OpenIntro Statistics. La versión original y en inglés del ejercicio se puede consultar en el siguiente vínculo:

<http://htmlpreview.github.io/?https://github.com/andrewpbray/oiLabs-base-R/blob/master/confidence_intervals/confidence_intervals.html>

En este ejercicio se utilizarán nuevamente datos del mercado inmobiliario de la ciudad de Ames, Iowa, EEUU. En particular, contamos con información de todas las ventas de hogares ocurridas en esa ciudad entre 2006 y 2010.

Si se tiene acceso a toda la población, pueden responderse preguntas tales como: ¿Qué tan grande es una casa promedio en Ames? o ¿cuánta variabilidad hay en el tamaño de los hogares? Si sólo se tiene acceso a una pequeña muestra de los datos, como suele ser el caso, responder a estas preguntas es más complicado. ¿Cuál es nuestra mejor estimación sobre el tamaño de una casa típica en Ames, dado que sólo conocemos el tamaño de unos pocos hogares? Este tipo de situaciones requiere utilizar las muestras para hacer inferencias sobre la población.

Utilizaremos sólo la variable que indica el tamaño del hogar (convertido a metros cuadrados). Esta será nuestra población.

Tomamos una muestra de 60 hogares. Como ya se ha visto, se puede hacer una estimación puntual del promedio poblacional mediante la media muestral. ¿Cuánto da el promedio muestral?

***Insertar respuesta aquí***

La media muestral es la mejor estimación puntual que podemos obtener, pero sería útil saber también qué tan incierta es esta estimación. Esto puede hacerse mediante un **intervalo de confianza**.

Primero calcularemos el error standard para la muestra de 60 hogares. ¿Qué resultado obtuvo?

***Insertar respuesta aquí***

Luego, calcularemos un intervalo de confianza al 95% para la media muestral, sumando y restando 1.96 errores standard a la media muestral. ¿Por qué utilizamos el valor de 1.96? ¿Se requiere algún supuesto adicional?

***Insertar respuesta aquí***

Calculamos límites inferior y superior del intervalo. ¿Cuáles son los límites del intervalo de confianza?

***Insertar respuesta aquí***

Se ha dado un paso muy grande para conocer el tamaño promedio de los hogares en Ames. Si bien no sabemos con exactitud cuál es su valor promedio, **tenemos un 95% de confianza de que se encuentra contenido en el intervalo estimado**.

¿Qué significa que tenemos 95% de confianza?

***Insertar respuesta aquí***

Para satisfacer nuestra curiosidad, ¿cuál es el verdadero valor de la media poblacional? Recuerde que esto sólo lo podemos calcular porque tenemos acceso a la población completa. No suele suceder esto en la vida real.

***Insertar respuesta aquí***

¿Se encuentra contenido en el intervalo de confianza estimado?

***Insertar respuesta aquí***

Cada estudiante de la clase debe haber obtenido un intervalo de confianza ligeramente diferente. ¿Qué proporción de esos intervalos cree que contendrán la verdadera media poblacional? ¿Por qué?

***Insertar respuesta aquí***

Vamos a simular que la clase tiene 50 estudiantes. De esta forma, vamos a obtener 50 muestras distintas, construyendo un intervalo de confianza a partir de cada una de ellas. Todas las muestras estarán conformadas por 60 hogares seleccionados al azar.

Calculemos cuántos de estos intervalos contienen al verdadero valor de la media poblacional:

***Insertar respuesta aquí***

Graficamos los intervalos de confianza. La línea vertical señala el verdadero valor de la media poblacional (generalmente desconocido).

***Insertar gráfico aquí***

Utilizando las mismas muestras, construyamos intervalos de confianza al 90% y 99%. Para ello, utilizaremos los valores críticos de 1.645 y 2.58. ¿Por qué estos valores?

***Insertar respuesta aquí***

Calculemos cuántos de estos intervalos contienen al verdadero valor de la media poblacional:

***Insertar respuesta aquí***

La cantidad de intervalos que contienen el verdadero valor de la media poblacional difiere cuando construimos un intervalo al 90%, 95% o 99%. ¿Por qué?

***Insertar respuesta aquí***

Graficamos todos los intervalos de confianza juntos para ver las diferencias. Para facilitar la visualización, ordenaremos los intervalos de acuerdo al valor de la media muestral:

***Insertar gráfico aquí***

¿Qué diferencias observa entre los tres casos? Relacione con la idea de un trade-off entre precisión y certeza.

***Insertar respuesta aquí***

Recordar que cada una de las muestras que estamos tomando se compone de 60 hogares seleccionados aleatoriamente. ¿Qué sucederá si, en lugar de 60 hogares, tomamos muestras de 120 hogares?

***Insertar respuesta aquí***

Hagamos la prueba, repitiendo el ejercicio de construir 50 intervalos para cada nivel de confianza, pero con muestras de tamaño 120 en lugar de 60.

Graficamos todos los intervalos, ordenados de acuerdo al tamaño de la media muestral:

***Insertar gráfico aquí***

Compare este gráfico con el obtenido mediante muestras de tamaño 60. ¿Qué diferencias observa en cuanto al ancho de los intervalos? ¿A qué se deben estas diferencias?

***Insertar respuesta aquí***

**Ejercicio 16**

En la primera parte del trabajo práctico se realizaron varias simulaciones de procesos aleatorios intentando ver si se obtenían resultados consistentes con los predichos por la teoría. En ese momento, no contábamos con herramientas muy formales para afirmar si el resultado obtenido estaba “cerca” o “lejos” del teórico, más allá de lo que podíamos evaluar mediante el valor z.

En este ejercicio, volveremos a evaluar los resultados de las simulaciones ya realizadas mediante la construcción de **intervalos de confianza**. Para ello, será necesario que tenga a mano la primera parte de su TP, ya que deberá introducir los resultados allí obtenidos en el código de R.

**Ejercicio 16 a)**

En el Ejercicio 5 de la primera parte del TP se simuló un mazo de naipes españolas (40 cartas) y se calculó la probabilidad teórica de que, en una mano de tres cartas, salga el 1 de espadas. La probabilidad calculada para este evento fue:

Se intentó estimar este valor mediante la simulación de 1 millón de manos, dando como resultado que la cantidad de manos que contenían al 1 de espadas fue:

***COMPLETAR CON EL RESULTADO DEL EJERCICIO 5 DE LA PRIMERA PARTE DEL TP***

De esta forma, a partir de esa muestra obtuvimos una proporción muestral () de:

***Insertar respuesta aquí***

Este resultado muestral, ¿avala o rechaza la idea de que el verdadero valor de la proporción es ? Para dar una respuesta formal a esta pregunta, utilicemos el resultado muestral para construir intervalos de confianza para la proporción poblacional. Utilizaremos niveles de confianza del 90%, 95% y 99%.

¿Qué supuestos son necesarios para construir el intervalo de confianza? ¿Se cumplen?

***Insertar respuesta aquí***

¿Qué distribución utilizará?

***Insertar respuesta aquí***

¿Cuáles son los valores críticos (z o t) para cada nivel de significatividad?

***Insertar respuesta aquí***

¿Cuál es el margen de error para cada nivel de confianza?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Cuál es el intervalo de confianza para la proporción poblacional al 90%, 95% y 99% respectivamente?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Se encuentra el valor teórico contenido por estos intervalos de confianza?

***Insertar respuesta aquí***

En base a los resultados obtenidos, ¿considera que la probabilidad teórica de obtener un 1 de espadas en una mano de tres cartas estuvo bien calculada?

***Insertar respuesta aquí***

**Ejercicio 16 b)**

En el Ejercicio 7.c) de la primera parte del TP se simuló el lanzamiento de un millón de monedas y se calculó la proporción de veces que se obtuvo una “cara” como resultado. Como se suponía que la moneda estaba equilibrada, la teoría sugiere que se debería obtener una proporción cercana a .

¿Cuál fue la proporción muestral () de caras que se obtuvo en el Ejercicio 7.c)?

***COMPLETAR CON EL RESULTADO DEL EJERCICIO 7c DE LA PRIMERA PARTE DEL TP***

***Insertar respuesta aquí***

Este resultado muestral, ¿avala o rechaza la idea de que la moneda está equilibrada? Para dar una respuesta formal a esta pregunta, utilicemos el resultado muestral para construir intervalos de confianza para la proporción poblacional. Utilizaremos niveles de confianza del 90%, 95% y 99%.

¿Qué supuestos son necesarios para construir el intervalo de confianza? ¿Se cumplen?

***Insertar respuesta aquí***

¿Qué distribución utilizará?

***Insertar respuesta aquí***

¿Cuáles son los valores críticos (z o t) para cada nivel de significatividad?

***Insertar respuesta aquí***

¿Cuál es el margen de error para cada nivel de confianza?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Cuál es el intervalo de confianza para la proporción poblacional al 90%, 95% y 99% respectivamente?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Se encuentra el valor teórico contenido por estos intervalos de confianza?

***Insertar respuesta aquí***

En base a los resultados obtenidos, ¿considera que la moneda efectivamente era equilibrada?

***Insertar respuesta aquí***

**Ejercicio 16 c)**

En el Ejercicio 7.d) de la primera parte del TP se simuló el lanzamiento de un millón de monedas y se calculó la proporción de veces que se obtuvo una “cara” como resultado. Se suponía que la moneda estaba cargada a favor de las caras, es decir, que la proporción teórica de caras que se debería obtener es mayor al 50%. Veamos si los resultados obtenidos avalan este supuesto.

¿Cuál fue la proporción muestral () de caras que se obtuvo en el Ejercicio 7.d)?

***COMPLETAR CON EL RESULTADO DEL EJERCICIO 7d DE LA PRIMERA PARTE DEL TP***

***Insertar respuesta aquí***

Este resultado muestral, ¿avala o rechaza la idea de que la moneda está cargada? Para dar una respuesta formal a esta pregunta, utilicemos el resultado muestral para construir intervalos de confianza para la proporción poblacional. Utilizaremos niveles de confianza del 90%, 95% y 99%.

¿Qué supuestos son necesarios para construir el intervalo de confianza? ¿Se cumplen?

***Insertar respuesta aquí***

¿Qué distribución utilizará?

***Insertar respuesta aquí***

¿Cuáles son los valores críticos (z o t) para cada nivel de significatividad?

***Insertar respuesta aquí***

¿Cuál es el margen de error para cada nivel de confianza?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Cuál es el intervalo de confianza para la proporción poblacional al 90%, 95% y 99% respectivamente?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

En base a los resultados obtenidos, ¿considera que la moneda efectivamente estaba cargada a favor de las caras?

***Insertar respuesta aquí***

**Ejercicio 17**

En la primera parte del trabajo práctico se realizaron varias simulaciones de procesos aleatorios intentando ver si se obtenían resultados consistentes con los predichos por la teoría. En ese momento, no contábamos con herramientas muy formales para afirmar si el resultado obtenido estaba “cerca” o “lejos” del teórico, más allá de lo que podíamos evaluar mediante el valor z.

En este ejercicio, volveremos a evaluar los resultados de las simulaciones ya realizadas mediante la realización de **test de hipotesis**. Para ello, será necesario que tenga a mano la primera parte de su TP, ya que deberá introducir los resultados allí obtenidos en el código de R.

**Ejercicio 17 a)**

En el Ejercicio 5 de la primera parte del TP se simuló un mazo de naipes españolas (40 cartas). Allí se afirmaba que la probabilidad teórica de que, en una mano de tres cartas, salga el 1 de espadas es . Veamos si los resultados obtenidos mediante una simulación de 1 millón de manos permiten respaldar esta hipótesis. Definimos las hipótesis nula y alternativa para un test bilateral:

¿Qué supuestos son necesarios para realizar el test de hipótesis? ¿Se cumplen?

***Insertar respuesta aquí***

¿Qué distribución utilizará?

***Insertar respuesta aquí***

En base a la simulación de 1 millón de manos realizada en el Ejercicio 5, ¿cuál fue la cantidad de manos que contenían al 1 de espadas?

***COMPLETAR CON EL RESULTADO DEL EJERCICIO 5 DE LA PRIMERA PARTE DEL TP***

De esta forma, a partir de esa muestra obtuvimos una proporción muestral () de:

***Insertar respuesta aquí***

¿Cuál es el z observado correspondiente a este resultado?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Cuál es el p-value de este resultado?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

En base a la respuesta anterior, considere si hay evidencia suficiente para rechazar o no la hipótesis nula. Justifique en interprete en el contexto del problema.

***Insertar respuesta aquí***

**Ejercicio 17 b)**

En el Ejercicio 7.c) de la primera parte del TP se simuló el lanzamiento de un millón de monedas y se calculó la proporción de veces que se obtuvo una “cara” como resultado. Se suponía que la moneda estaba equilibrada, es decir, que se deberían obtener resultados en torno a 50% de caras y 50% de cecas. En términos de hipótesis, este supuesto lo podemos expresar de la siguiente forma.

Veamos si los resultados obtenidos en la simulación de 1 millón de lanzamientos de monedas del Ejercicio 7.c) refuta o no la hipótesis nula.

¿Cuál fue la proporción muestral () de caras que se obtuvo en el Ejercicio 7.c)?

***COMPLETAR CON EL RESULTADO DEL EJERCICIO 7c DE LA PRIMERA PARTE DEL TP***

***Insertar respuesta aquí***

¿Qué supuestos son necesarios para realizar el test de hipótesis? ¿Se cumplen?

***Insertar respuesta aquí***

¿Qué distribución utilizará?

***Insertar respuesta aquí***

¿Cuál es el z observado correspondiente al resultado de la simulación?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Cuál es el p-value de este resultado?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

En base a la respuesta anterior, considere si hay evidencia suficiente para rechazar o no la hipótesis nula. Justifique en interprete en el contexto del problema.

***Insertar respuesta aquí***

**Ejercicio 17 c)**

En el Ejercicio 7.d) de la primera parte del TP se simuló el lanzamiento de un millón de monedas y se calculó la proporción de veces que se obtuvo una “cara” como resultado. Se suponía que la moneda estaba cargada a favor de las caras, pero vamos a hacer de cuenta que nadie nos avisó de esto. Es decir, vamos a suponer que la moneda estaba equilibrada y ver si los resultados obtenidos nos obligan a cambiar de opinión. En términos de hipótesis, este supuesto lo podemos expresar de la siguiente forma.

¿Cuál fue la proporción muestral () de caras que se obtuvo en el Ejercicio 7.d)?

***COMPLETAR CON EL RESULTADO DEL EJERCICIO 7d DE LA PRIMERA PARTE DEL TP***

***Insertar respuesta aquí***

¿Qué supuestos son necesarios para realizar el test de hipótesis? ¿Se cumplen?

***Insertar respuesta aquí***

¿Qué distribución utilizará?

***Insertar respuesta aquí***

¿Cuál es el z observado correspondiente al resultado de la simulación?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Cuál es el p-value de este resultado?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

En base a la respuesta anterior, considere si hay evidencia suficiente para rechazar o no la hipótesis nula. Justifique en interprete en el contexto del problema.

***Insertar respuesta aquí***

**Ejercicio 18**

En la primera parte del trabajo práctico se realizaron varias simulaciones de procesos aleatorios intentando ver si se obtenían resultados consistentes con los predichos por la teoría. En ese momento, no contábamos con herramientas muy formales para afirmar si el resultado obtenido estaba “cerca” o “lejos” del teórico, más allá de lo que podíamos evaluar mediante el valor z.

En este ejercicio, volveremos a evaluar los resultados de las simulaciones ya realizadas mediante la realización de **test de hipotesis**. Para ello, será necesario que tenga a mano la primera parte de su TP, ya que deberá introducir los resultados allí obtenidos en el código de R.

**Ejercicio 18 a)**

En el Ejercicio 8.c) se lanzó un millón de veces un dado supuestamente equilibrado, calculándose la proporción de veces que salieron el número 1 y el 6. Casi con seguridad, las proporciones obtenidas no fueron exactamente iguales, por lo que realizaremos un **Test de Hipótesis para diferencia de Proporciones** para ver si podemos confiar en que el dado estaba efectivamente equilibrado. Las hipótesis del test son las siguientes

¿Cuáles fueron la proporción muestrales de unos y seis () que se obtuvieron en el Ejercicio 8.c)?

***COMPLETAR CON EL RESULTADO DEL EJERCICIO 8d DE LA PRIMERA PARTE DEL TP***

***Insertar respuesta aquí***

¿Qué supuestos son necesarios para realizar el test de hipótesis? ¿Se cumplen?

***Insertar respuesta aquí***

¿Qué distribución utilizará?

***Insertar respuesta aquí***

En este test se supone que ambas proporciones son iguales, para lo cual es necesario calcular un promedio ponderado de las proporciones muestrales. ¿Cuál es el valor de ?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Cuál es el z observado correspondiente al resultado de la simulación?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Cuál es el p-value de este resultado?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

En base a la respuesta anterior, considere si hay evidencia suficiente para rechazar o no la hipótesis nula. Justifique en interprete en el contexto del problema.

***Insertar respuesta aquí***

**Ejercicio 18 b)**

En el Ejercicio 8.d) se lanzó un millón de veces un dado supuestamente cargado a favor del 6, calculándose la proporción de veces que salieron el número 1 y el 6. Casi con seguridad, se obtuvo una mayor proporción de 6 que de 1, pero requerimos de una prueba más rigurosa para saber si efectivamente esto es evidencia de que el dado estaba cargado. Para ello, realizaremos un **Test de Hipótesis para diferencia de Proporciones** con las siguientes hipótesis:

¿Cuáles fueron la proporción muestrales de unos y seis () que se obtuvieron en el Ejercicio 8.d)?

***COMPLETAR CON EL RESULTADO DEL EJERCICIO 8d DE LA PRIMERA PARTE DEL TP***

***Insertar respuesta aquí***

¿Qué supuestos son necesarios para realizar el test de hipótesis? ¿Se cumplen?

***Insertar respuesta aquí***

¿Qué distribución utilizará?

***Insertar respuesta aquí***

En este test se supone que ambas proporciones son iguales, para lo cual es necesario calcular un promedio ponderado de las proporciones muestrales. ¿Cuál es el valor de ?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Cuál es el z observado correspondiente al resultado de la simulación?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

¿Cuál es el p-value de este resultado?

***INTENTE RESOLVER ANALÍTICAMENTE. UTILICE LOS RESULTADOS DE R PARA COMPROBAR.***

***Insertar respuesta aquí***

En base a la respuesta anterior, considere si hay evidencia suficiente para rechazar o no la hipótesis nula. Justifique en interprete en el contexto del problema.

***Insertar respuesta aquí***

**Ejercicio 19**

**El "Beta" de un activo financiero**

Este ejercicio fue elaborado en base a la siguiente bibliografía:

- James Stock and Mark Watson "Introduction to econometrics". Second Edition, 2007

- Paul Newbold, William Carlson y Betty Thorne, "Estadística para administración y Economía", 8va edición, 2013

La economía financiera ha desarrollado algunas medidas y métodos de análisis para ayudar a los inversores a medir y a controlar el riesgo financiero en el desarrollo de carteras de inversión. El riesgo se puede dividir en riesgo diversificable y riesgo no diversificable. El **riesgo diversificable** es el riesgo que entrañan determinadas empresas y sectores e incluye los conflictos laborales, la nueva competencia, los cambios del mercado de consumidores y otros muchos factores. Este riesgo se puede controlar por medio de una cartera mayor e incluyendo activos cuyos rendimientos tengan **correlaciones negativas**.

El **riesgo no diversificable** es el riesgo que entraña el conjunto de la economía. Los cambios económicos provocados por los ciclos económicos, las crisis internacionales, los cambios de la demanda mundial de energía u otros factores que afectan a todas las empresas, pero no producen el mismo efecto en todas ellas. Una medida de qué tan relevante es el riesgo no diversificable para determinada empresa se obtiene por medio del **coeficiente beta**, como se analiza a continuación.

**Utilizando el modelo CAPM para el cálculo del coeficiente beta**

Una idea fundamental detrás de las finanzas es que un inversor necesita de un incentivo financiero para que esté dispuesto a tomar un riesgo. Dicho de otra manera, el retorno esperado de una inversión riesgosa (*R*), debe ser superior al retorno de una inversión segura o libre de riesgo (*R0*). Es decir, el **retorno excedente esperado** (*R - R0*) de una inversión riesgosa, como comprar acciones de determinada empresa, debe ser positivo.

El **Modelo de valoración de activos financieros** (CAPM, *Capital Asset Pricing Model* por sus siglas en inglés) formaliza esta idea. De acuerdo con el CAPM, el **retorno excedente esperado** de un activo es proporcional al retorno excedente esperado de una cartera con todos los activos disponibles: la cartera del mercado. Esto es, el CAPM dice que:

donde *Rm* es el retorno excedente esperado de la cartera del mercado y es el coeficiente de la regresión lineal de sobre . Es decir, el coeficiente de una empresa es el coeficiente de la pendiente que se obtiene cuando se realiza una regresión del rendimiento excedente de una empresa con respecto a los rendimientos excedentes de un índice de mercado:

Donde el coeficiente α se supone que toma un valor 0 (o no significativo) y es la perturbación aleatoria.

Este coeficiente indica la sensibilidad de los rendimientos de una empresa a los rendimientos totales del mercado. En la mayoría de los casos, el coeficiente será positivo, pero en algunos, los rendimientos de una empresa varían en sentido contrario al del conjunto de la economía. Si los rendimientos de la empresa siguen al mercado exactamente, tendremos . Por su parte, una acción con es menos sensible a los rendimientos del mercado y una acción con es más sensible que el resto del mercado. Un beta más alto indica que el rendimiento exigido a la inversión tiene que ser mayor. Este mayor rendimiento exigido tendría en cuenta el hecho de que en el rendimiento de las acciones influye más el riesgo de mercado no diversificable. La diversificación por medio de mayores carteras no puede tener en cuenta los cambios generales del mercado.

El cálculo de los es una práctica usual en finanzas y pueden encontrarse distintas estimaciones en la web. Por ejemplo, en el siguiente link se encuentran los estimados para distintas acciones del mercado de EEUU: <http://www.abg-analytics.com/stock-betas.shtml>

Por lo general, como inversión segura o libre de riesgo se suelen tomar los bonos de corto plazo del Tesoro de los EEUU, siendo su retorno el valor *R0* utilizado para calcular el retorno excedente. Pero dependiendo de los mercados a los que acceda el inversor, pueden tomarse otros activos.

En este ejercicio, calcularemos el de dos acciones del mercado argentino: Banco Galicia e YPF. Como cartera del mercado consideraremos al índice MERVAL, el cual está compuesto por las principales acciones del mercado local. Como activo libre de riesgo, consideraremos un plazo fijo en pesos, estimando su rendimiento a través de la tasa BADLAR publicada por el BCRA, y que es la tasa de referencia para los plazos fijos mayores a 1 millón de pesos. Trabajaremos con datos de frecuencia mensual.

Antes de comenzar a trabajar, veamos un gráfico de cada una de las series:

***Insertar gráfico aquí***

Tenga en cuenta que la BADLAR está en % (es una tasa), mientras que el resto de los activos está en pesos.

Se observa que hay datos desde 2005 hasta la actualidad, y que el mercado ha pasado a lo largo de esos años por distintos niveles de volatilidad. Trabajaremos primero con las series completas, enfocándonos luego en un período más corto para ver si cambia en algo el análisis.

Calculamos las variaciones porcentuales mensuales de cada acción. Para la tasa BADLAR, como está expresada en términos anuales, calculamos el rendimiento mensual implícito en esas tasas.

Veamos ahora un gráfico de estos rendimientos:

***Insertar gráfico aquí***

¿Qué comportamientos observa en los rendimientos de las acciones por un lado y en la tasa de plazos fijos BADLAR por el otro?

***Insertar respuesta aquí***

Calculemos ahora los **rendimientos excedentes** de cada acción. Es decir, al rendimiento de cada activo le restamos el rendimiento del activo libre de riesgo (plazo fijo).

Grafiquemos estos rendimientos excedentes:

***Insertar gráfico aquí***

¿Observa alguna diferencia entre cada activo?

***Insertar respuesta aquí***

Veamos ahora cómo se relaciona el rendimiento excedente de cada una de las acciones con el rendimiento excedente del MERVAL.

Comencemos con el Banco Galicia, calculando la siguiente regresión lineal mediante **Mínimos Cuadrados Ordinarios**:

***Insertar resultados de la regresión aquí***

En los resultados de la regresión observará una tabla con dos coeficientes: (Intercept), correspondiente al , y (rendimientos\_exc$MERVAL), correspondiente al . Bajo la columna “Estimate” encontrará el valor estimado para el coeficiente y en la última columna de la tabla está el valor-p del test de hipótesis cuya hipótesis nula es que el coeficiente es igual a 0. Si al lado del valor-p aparecen asteriscos (\*, \*\*, o \*\*\*), significa que la hipótesis nula se rechaza con un nivel de significatividad del 5%, 1% y 0,1% respectivamente.

¿Qué coeficientes resultan significativos? ¿Cuál es el valor del Beta para el Banco Galicia? ¿Cuál es el valor del R2 ajustado (Adjusted R-squared)?

***Insertar respuesta aquí***

Realice el gráfico de dispersión entre ambas variables, junto con la recta de regresión:

***Insertar gráfico aquí***

¿Considera que la regresión lineal es útil para explicar el retorno excedente del Banco Galicia en base al retorno excedente del MERVAL?

***Insertar respuesta aquí***

¿Observa algún valor claramente atípico?

***Insertar respuesta aquí***

La estimación de los Betas y de la recta de regresión pueden verse afectadas por la existencia de valores atípicos. Realicemos nuevamente la estimación por MCO, pero excluyendo previamente algunas observaciones que podrían ser atípicas.

***Insertar resultados de la regresión aquí***

¿Observa resultados muy diferentes?

***Insertar respuesta aquí***

Restrinjamos el análisis a un período donde las tasas de interés fueron relativamente estables. Por ejemplo, consideremos el período 2007-2012.

Calculemos la regresión lineal por MCO:

***Insertar resultados de la regresión aquí***

***Insertar gráfico aquí***

¿Observa resultados muy diferentes? En base a todos estos resultados, ¿qué tan estable considera que es el Beta de la acción del Banco Galicia?

***Insertar respuesta aquí***

Repitamos el ejercicio, pero ahora para YPF. Comencemos estimando la siguiente regresión para todo el período, incluyendo todas las observaciones.

***Insertar resultados de la regresión aquí***

¿Qué coeficientes resultan significativos? ¿Cuál es el valor del Beta para YPF? ¿Cuál es el valor del R2 ajustado? Compare sus resultados con los obtenidos para el Banco Galicia.

***Insertar respuesta aquí***

Realice el gráfico de dispersión entre ambas variables, junto con la recta de regresión:

***Insertar gráfico aquí***

¿Considera que la regresión lineal es útil para explicar el retorno excedente de YPF en base al retorno excedente del MERVAL?

***Insertar respuesta aquí***

¿Observa algún valor atípico?

***Insertar respuesta aquí***

Como ya dijimos, la estimación de los Betas y de la recta de regresión pueden verse afectadas por la existencia de valores atípicos. Realicemos nuevamente la estimación por MCO, pero excluyendo previamente algunas observaciones que parecen ser atípicas.

***Insertar resultados de la regresión aquí***

¿Observa resultados muy diferentes?

***Insertar respuesta aquí***

Restrinjamos el análisis a un período donde las tasas de interés fueron relativamente estables. Por ejemplo, consideremos el período 2007-2012.

Calculemos la regresión lineal por MCO:

***Insertar resultados de la regresión aquí***

***Insertar gráfico aquí***

En base a todos estos resultados, ¿qué tan estable considera que es el Beta de la acción del YPF?

***Insertar respuesta aquí***

Si tuviera que considerar invertir en acciones del Banco Galicia o en YPF, ¿qué acción considera más riesgosa? ¿Con cuál esperaría un retorno mayor?

***Insertar respuesta aquí***