Universidad Torcuato Di Tella

Métodos Estadísticos Aplicados a Negocios

Trabajo Práctico 1

Agustín Alba Chicar
ag.albachicar@gmail.com

Joaquín Gonzalez
joagonzalez@gmail.com

16 de abril de 2019

Índice

1. Instrucciones

El trabajo puede ser grupal (con un máximo de 4 alumnos). Deberán entregar un informe final a través del campus virtual en el cual adjuntarán el Script en `R` para poder replicar sus resultados.

Tienen tiempo para entregar el TP hasta el lunes 15 de abril a las 23:55hs. Una entrega tardía implica desaprobar el trabajo práctico.

2. Consignas

Utilice la base de datos `BlackFriday.csv` que contiene una muestra de más de 500.000 transacciones realizadas en una tienda de retail en un Black Friday. La tienda quiere tener un mejor conocimiento de la compra de los clientes frente a diferentes productos. La base de datos incluye las siguientes variables:

- `Use_ID`: código del cliente
- `Product_ID`: código del producto comprado
- `Gender`: M (Male), F (Female)
- `Age`: edad del cliente
- `Occupation`: código de la ocupación del cliente
- `City_Category`: código de la ciudad (A, B, C)
- `Stay_In_Current_City_Years`: cantidad de años que el cliente vive en dicha ciudad
- `Marital Status`: 1 (casado/a), el resto de los códigos indican otros estados civiles.
- `Product_Category_1`
- `Product_Category_2`
- `Product_Category_3`
- `Purchase`: monto de la compra en dólares

Se pide:

(1) '**(5 puntos)** Describa la base de datos y obtenga los principales estadísticos descriptivos. Comente.

(2) '**(5 puntos)** Construya un intervalo de confianza de la proporción de mujeres que realizan compras en el Black Friday.

(3) '**(10 puntos)** Estime intervalos de confianza del monto medio gastado por un individuo en el Black Friday, pero segmentando por edad. Intente expresarlo gráficamente. Comente. ¿Existen diferencias en el monto gastado de acuerdo a la edad?

(4) '**(10 puntos)** ¿Cuál es el producto más comprado? (Identifique a partir del ID). Pruebe si la proporción de mujeres que compran dicho producto es igual al 27 %. Grafique la curva de potencia del test. Comente.

(5) '**(10 puntos)** ¿Hay evidencia suficiente para probar que un individuo gasta más de 9300 dólares en promedio? ¿Y más de 9400 dólares? Construya el estadístico y el pvalor en `R`. Concluya.

(6) '**(10 puntos)** El dueño de la tienda afirma que las mujeres gastan más dinero que los hombres. ¿Encuentra evidencia empírica que valide dicha afirmación?

(7) '**(10 puntos)** ¿Hay evidencia suficiente respecto de que las personas casadas gastan más dinero en el Black Friday? Justifique.

(8) '**(10 puntos)** Pruebe si existen diferencias significativas en el monto gastado entre las tres ciudades (A, B y C).

(9) '**(10 puntos)** ¿Hay evidencia suficiente respecto de que las mujeres casadas que tienen entre 26-35 años gastan más dinero en un Black Friday que los hombres casados del mismo grupo etario?

(10) '**¡Vamos por el 10! (20 puntos)** Usando los datos de esta muestra diseñe una prueba para evaluar si los siguientes estimadores de la proporción poblacional (p) son insesgados, eficientes y consistentes. Justifique. (Hint: extraiga al azar muestras pequeñas y grandes a partir de alguna variable de la base de datos).

3. Desarrollo

3.1. Consigna 1

3.1.1. Generalidades

La base de datos representa una muestra de las transacciones (más de 500.000) realizadas durante un BlackFriday. A su vez, posee información de los consumidores y de los productos transados en dicho día.

Los datos presentan *missings (NA)* por lo que al momento de trabajar con los mismos deberán ser tenidos en cuenta.

3.1.2. Compras

Se computan los estadísticos más importantes sobre las columnas relacionadas con ventas.

Cuadro 1: Propiedades de las compras.

Característica	**Media muestral**	**Desvío estándar muestral**
Compras	9333,86 U$D	4981,02 U$D

3.1.3. Productos comercializados

Se distinguen 99 productos más una categoría extra que nuclea a todos los que no pueden ser distinguidos ("(Other)"). De esta forma, podemos mencionar:

- P00002142 es el producto menos transado, con 728 unidades.
- P00265242 es el producto más transado, con 1858 unidades.
- La media de productos transados es 991 unidades (sin considerar la categoría *(Other)*) con un desvío muestral de 252 unidades.

Por otro lado, los productos poseen tres tipos de categorías, donde cada categroría presenta su propio set de tipos. En función de las categorías, podemos determinar lo siguiente:

Se debe destacar para las categorías 2 y 3, los *missings* fueron removidos para considerar los máximos y mínimos, así como la media de unidades y desvíos.

3.1.4. Clientes

A nivel cliente, podemos distinguir su género, condición civil (sólo si están casados, los otros tipos se conglomeran en un tipo *no-casado*) y su rango etario. En las figuras 2 a 3 mostramos la composición.

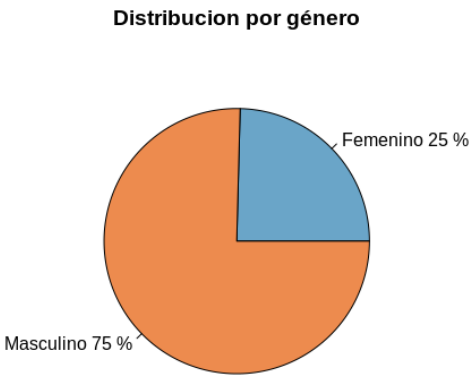

Figura 1: Distribución por género de los clientes.

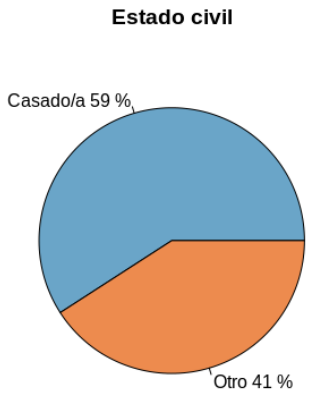

Figura 2: Distribución por estado civil de los clientes.

Cuadro 2: Propiedades de los productos.

Categoría	**Cantidad**	**Mínimo de productos por categoría**	**Máximo de productos por categoría**	**Media muestral de productos por categoría**	**Desvío estándar muestral de productos por categoría**
Categoría 1	18	Tipo 9, 0,075 %	Tipo 5, 27,64 %	29865u.	48524u.
Categoría 2	18	Tipo 7, 0,114 %	Tipo 8, 11,73 %	29865u.	39427u.
Categoría 3	16	Tipo 3, 0,111 %	Tipo 16, 5,98 %	33598u.	91068u.

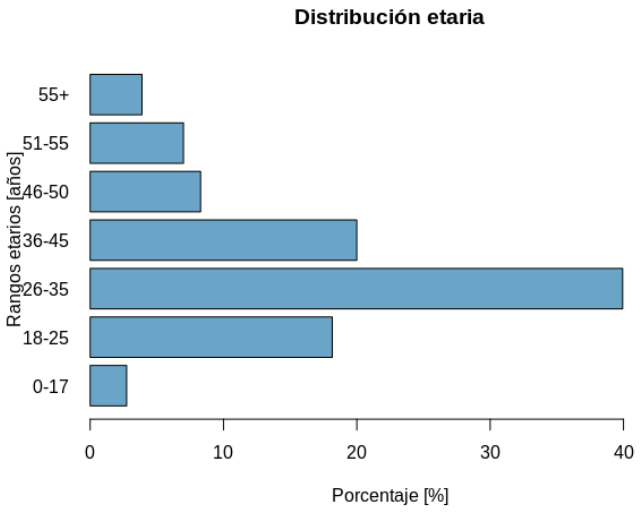

Figura 3: Distribución por rango etario de los clientes

3.2. Consigna 2

Inspeccionando la columna *Gender* no encontramos *missings* por lo que procedemos a utilizar la totalidad de las muestras. Computamos la proporción de clientes con género femenino como:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{i=n} x_{f_i} \tag{1}$$

Donde:

- $\hat{p}$ es el estimador de la proporción.
- n es el número de muestras.
- x_{f_i} vale 1 si es un caso favorable y 0 cuando no.

A su vez, presentamos la varianza muestral para el estimador de la proporción:

$$\hat{s^2_{\hat{p}}} = \frac{\hat{p}(1-\hat{p})}{n} \tag{2}$$

Finalmente, vasta adoptar un margen de significación para la medición de la proporción de clientes con género femenino. Adoptamos un $\alpha = 5\,\%$, y conociendo que la distribución de los estimadores proporción se comportan con una característica normal, construimos el intervalo de confianza de la siguiente manera:

$$IC = \left[\hat{p} - t_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{s^2_{\hat{p}}}}{n}}; \hat{p} + t_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{s^2_{\hat{p}}}}{n}} \right] \tag{3}$$

Donde $t_{1-\frac{\alpha}{2}}$ es el valor de probabilidad acumulada en una distribución t-Student normalizada con $n-1$ grados de libertad.

Utilizando R, computamos dichos valores resultando así:

$$\hat{p} = 24,59127\,\% \tag{4}$$

$$IC = [24.,59111\,\%; 24,59142\,\%] \tag{5}$$

3.3. Consigna 3

Para obtener la información solicitada se procede a realizar en R el siguiente procedimiento:

- Filtrar la base para quedarse con las columnas: *Purchase*, *Age*.
- Computar promedios, desvíos e intervalos de confianza para cada uno de los segmentos etarios (previa definición α).

Producto de dicho procedimiento, obtenemos la imagen 5 (ver tabla 3.3)

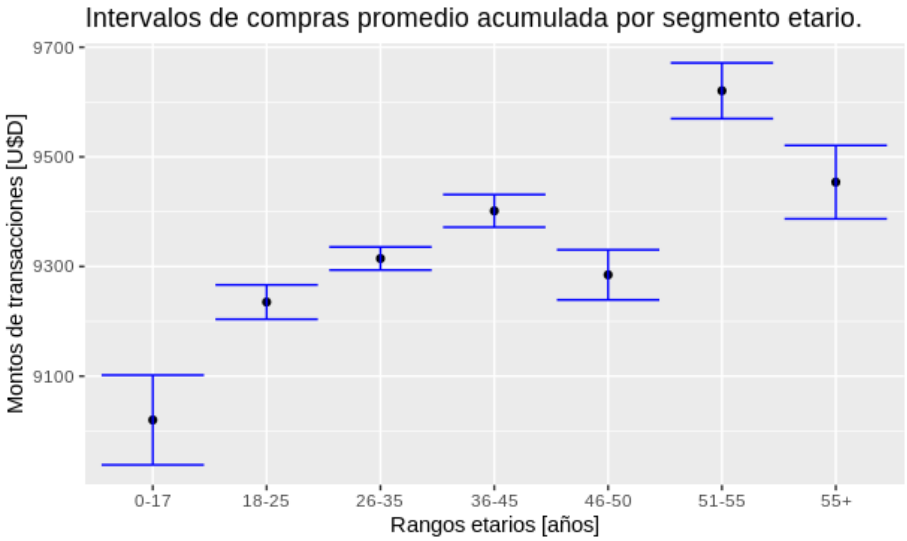

Figura 4: Intervalos de compra promedio según rango etario.

Cuadro 3: Intervalos al 95 % ($\alpha = 5\,\%$) de montos de compra promedio en función de los rangos etarios.

Rango etario	**Monto promedio**	**Monto mínimo**	**Monto máximo**
0-17 años	9020U$D	8938U$D	9102U$D
18-25 años	9235U$D	9204U$D	9267U$D
26-35 años	9315U$D	9294U$D	9336U$D
36-45 años	9401U$D	9372U$D	921834U$D
46-50 años	9285U$D	9239U$D	9431U$D
51-55 años	9621U$D	9570U$D	9672U$D
+55 años	9454U$D	9387U$D	9521U$D

Comparamos entonces las medias de los intervalos y observamos que en valores absolutos los rangos 18-25 años, 26-35 años y 36-45 años presentan el mayor promedio de compra. Por otro lado, es curioso destacar que los intervalos 46-55 y 51-55 presentan en sí menores promedios aunque cada uno encierra una rango etario más pequeño (5 años cada uno). En sí, son generaciones distintas y cada una presenta hábitos de consumo distintos a la vez que capacidad de afrontar gastos, no resulta curioso que los segmentos etarios que presentan mayor consumo son los que fácilmente podemos asociar con la población económicamente activa y más joven.

Por último, se desarrollan múltiples test de hipótesis para comparar los gastos entre todos los rangos etarios. De esta manera, se podrá observar con el nivel de significancia elegido, en que casos hay suficiente evidencia empírica como para rechazar la suposición de que los gastos son iguales. Esto permitirá abordar la evidencia gráfica del punto anterior con mayor rigor estadístico.

Las hipótesis nula y alternativa se listan a continuación:

- *Hipótesis nula*: La media de gastos del rango etario i es igual a la del rango etario j tal que $i \neq j$ y $1 < i, j < 7$ ($length(rangos\_etarios) = 7$).
- *Hipótesis alternativa*: La media de gastos del rango etario i es distinta de la media de gastos del rango etario j tal que $i \neq j$ y $1 < i, j < 7$ ($length(rangos_etarios) = 7$).

Se puede observar en la tabla 3.3 que los tests que comparan rangos etarios donde hay una superposición considerable en los intervalos de confianza, no poseen evidencia empírica suficiente como para indicar que son diferentes. Los resultados de los tests fueron computados realizando la siguiente comparación $p - valor < \alpha$ para cada una de las iteraciones.

Cuadro 4: Tests donde no se rechaza la hipótesis nula.

Edad i	**Edad** j
18-25 años	45-50
26-35 años	45-50
36-45 años	55+

Cuadro 5: Datos del producto con mas compras.

Product_ID	**Cantidad**
P00265242	1858

3.4. Consigna 4

Para obtener la información solicitada se procede a realizar en R el siguiente procedimiento:

- Filtrar la base para quedarse con las columnas: *Product_ID*, *Gender*, *Purchase*
- Reducir la tabla sumando la cantidad de veces que fue comprado cada producto, unificando el volumen de las filas. Se filtra el producto con mas compras.
- Para el producto mas comprado hallado anteriormente, se obtiene la proporción por genero de esas compras.
- Se computa error muestral y el estadístico de proporción $\hat{p}$.
- Para analizar si la proporción de mujeres que compran ese producto igual al 27 % se plantea una prueba de hipótesis.

Las hipótesis nula y alternativa se listan a continuación:

- *Hipótesis nula*: proporción de mujeres que compran el producto = 27 %.
- *Hipótesis alternativa*: proporción de mujeres que compran el producto != 27 %.

Al tratarse de un estadístico de proporción, utilizamos distribución normal con la siguiente expresión:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1) \tag{6}$$

Utilizando R, computamos dichos valores resultando así:

$$\hat{p} = 0{,}2717976 \tag{7}$$

$$IC = [0{,}2516727; 0{,}2926454] \tag{8}$$

Observamos que $\hat{p}$ cae dentro del intervalo de confianza. No existe evidencia suficiente para refutar la hipótesis nula.

Calculamos la potencia del test 1-β. Es razonable que la potencia aumente a medida que nos alejamos del valor esperado. 1-β es la capacidad de rechazar la hipótesis nula siendo esta falsa (evitar errores tipo II), y mientras mas alejados del valor real estemos en términos de las proporciones obtenidas de la muestra, mas difícil es y mas potencia tiene el test.

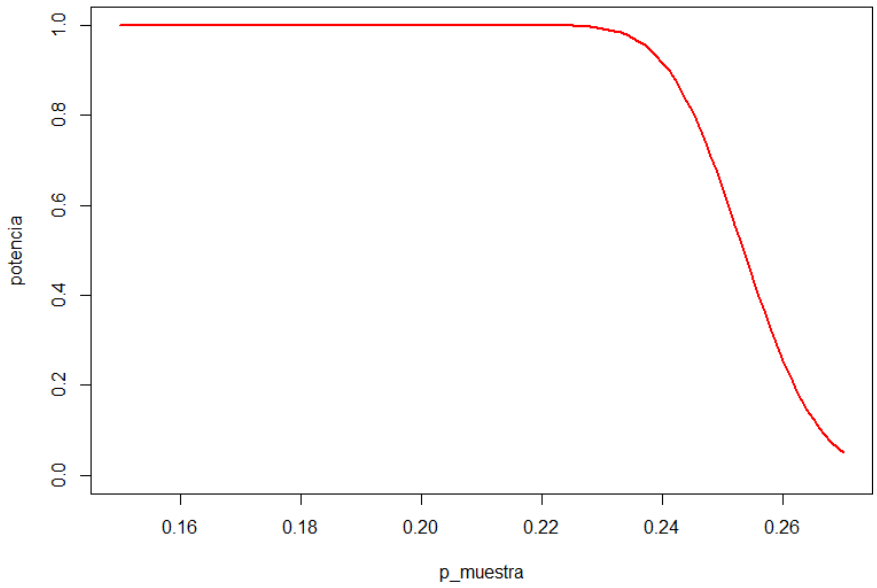

Figura 5: Potencia del test para distintos valores del estadístico proporción $\hat{p}$.

3.5. Consigna 5

Para obtener la información solicitada se procede a realizar en R el siguiente procedimiento:

- Filtrar la base para quedarse con las columnas: *User_ID*, *Gender*, *Purchase*
- Se toma la media de lo gastado por un individuo. Aquí se realiza una aclaración: Si bien el enunciado habla de promedio por individuo, si se agrupan gastos por User_ID y se toma la media, el valor el muy superior a 9300 U$D lo que genera un resultado trivial al realizar un análisis. Por esta razón se toma la media total de la columna *Purchase* que da un valor de 9333.86 U$.
- Se computa error muestral y el estadístico de media junto con su varianza y p-valor.
- Para analizar si la media de gasto por individuo es mayor a los umbrales indicados, se realiza una prueba de hipótesis

Las hipótesis nula y alternativa se listan a continuación:

- *Hipótesis nula*: media de gasto por usuario es menor o igual a 9300U$ / 9400U$
- *Hipótesis alternativa*: media de gasto por usuario es mayor a 9300U$ / 9400U$

Se trata de un test unilateral a una cola. Como la varianza poblacional es desconocida, debe estimarse y el estadístico sigue una distribución t-Student.

Para el caso del primer test, se obtiene un p-valor = 3.113e-07, menor que el nivel de significancia elegido $\alpha = 5\,\%$. Se concluye que existe evidencia suficiente para rechazar la hipótesis nula.

Para el caso del segundo test, p-valor = 1 es mayor que α. Se concluye que no existe evidencia suficiente para rechazar la hipótesis nula.

3.6. Consigna 6

Para obtener la información solicitada se procede a realizar en R el siguiente procedimiento:

- Filtrar la base para quedarse con las columnas: *Gender*, *Purchase*
- Se promedian los consumos de usuarios según su género.
- Se realiza un test de hipótesis de diferencia de medias.
- *Hipótesis nula*: la diferencia de medias de consumo entre hombres y mujeres es menor o igual a cero.
- *Hipótesis alternativa*: la diferencia de medias de consumo entre hombres y mujeres es mayor a cero.

Se trata de un test de diferencia de medias con varianzas poblacionales desconocidas y distintas. Se utiliza la distribución t-Student y el estadísitco es:

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_0}{s_p \sqrt{\frac{S_x^2}{n_X} + \frac{S_x^2}{n_Y}}} \sim t_{\mu_0} \tag{9}$$

Este test también es conocido como *Test de Welch* y se computa en *R* utilizando la función t.test().

Se obtienen los siguientes valores:

Cuadro 6: Resultados significativos del test de hipótesis.

Variable	**Valor**
z_t	45.673
α	0.05
$GL\,(z_t)$	238460
$p - valor$	$2{,}2^{-16}$

Se concluye que existe evidencia suficiente para rechazar la hipótesis nula ya que el p-valor computado es menor que α. Es decir, no podemos afirmar estadísticamente que las mujeres gastas mas dinero que los hombres para esta muestra de la población.

3.7. Consigna 7

La consigna se plantea como un test de hipótesis de diferencias de medias con medias y desvíos poblacionales distintos y desconocidos. Dicho eso, la metodología de resolución se basa en seguir los siguientes pasos:

- A partir de la base de datos, se obtienen los consumos y su estado civil.
- Luego, se computan los promedios de consumos según su estado civil (*casados* y *otros*). Además se agregan los desvíos muestrales de cada media con el número de muestras.
- Se genera el estadístico en función de la hipótesis nula y alternativa. Se ejecuta el test.
- Se provee la conclusión.

Las hipótesis nula y alternativa se listan a continuación:

- *Hipótesis nula*: la diferencia de medias de consumo entre la población casada y la población con otro estado civil es menor o igual a cero.
- *Hipótesis alternativa*: la diferencia de medias de consumo entre la población casada y la población con otro estado civil es mayor a cero.

Definimos para el test, un valor de significancia α del 5 %.
Confeccionamos entonces el estimador del test:

$$z_t = \frac{(\bar{x_c} - \bar{x_o}) - 0}{\sqrt{\frac{s_c^2}{n_c} + \frac{s_o^2}{n_o} + 2Cov\left(x_c, x_o\right)}} \tag{10}$$

Donde:

- z_t es el estimador de distribución t-Student.
- x_c es la media muestral de consumos de gente casada.
- x_o es la media muestral de consumos de gente con otro estado civil.
- s_c es el desvío muestral de consumos de gente casada.
- s_o es la desvío muestral de consumos de gente con otro estado civil.
- n_c es el número de consumos de gente casada.
- n_o es el número de consumos de gente con otro estado civil.
- $Cov\left(x_c, x_o\right)$ es la covarianza entre las medias muestrales de la gente casada y la gente con otro estado civil.

Un detalle importante, es que el término de covarianza se anula puesto que no existe correlación entre las muestras. Las personas pertenecen a grupos distintos que se asumen sin relación alguna. Esto simplifica el estudio del error muestral del estimador.

Como se mencionó antes, el estimador z_t es un estimador que posee una distribución t-Student. Falta mencionar los grados de libertad de dicha distribución. El cómputo se realiza con la siguiente expresión:

$$GL\left(z_t\right) = \frac{\left(\frac{s_c^2}{n_c} + \frac{s_o^2}{n_o}\right)^2}{\frac{\left(\frac{s_c^2}{n_c}\right)^2}{n_c - 1} + \frac{\left(\frac{s_o^2}{n_o}\right)^2}{n_o - 1}} \tag{11}$$

Finalmente, utilizando `R`, obtenemos los siguientes valores:

Cuadro 7: Resultados significativos del test de hipótesis.

Variable	**Valor**
Media de consumos de los casados	9334.63U$D
Media de consumos de los no casados	9333.33U$D
$GL\left(z_t\right)$	537580
$p - valor$	0.462
t_c	-1.644

Donde t_c es el valor de t que hace que la probabilidad acumulada de la distribución sea $1 - \alpha$ (95 %) desde t_c hasta $+\infty$.

Como $\alpha < p - valor$ podemos concluir, entonces, que no existe suficiente evidencia estadística que el promedio de consumos de la población casada es mayor al promedio de consumos de la población con otro estado civil, o sea, para rechazar la hipótesis nula.

Nota: se realizó un test similar al de Welch pero asumiendo varianzas muestrales iguales y no se obtuvo diferencia alguna en el resultado.

3.8. Consigna 8

La consigna se plantea como un test de hipótesis de diferencias de medias con medias y desvíos muestrales distintos y desconocidos. Dicho eso, la metodología de resolución se basa en seguir los siguientes pasos:

- A partir de la base de datos, se obtienen los consumos y la categoría de ciudades a la que pertenecen.
- Luego, se obtienen en vectores distintos los valores de las compras según la categoría de ciudad (A, B y C).
- Se genera el estadístico en función de la hipótesis nula y alternativa. Se ejecuta el test tomando a las ciudades de a pares.
- Se provee la conclusión.

Las hipótesis nula y alternativa se listan a continuación:

- *Hipótesis nula*: la diferencia de medias de consumo entre la ciudad X y la ciudad Y son iguales.
- *Hipótesis alternativa*: la diferencia de medias de consumo entre la ciudad X y la ciudad Y es distinta.

Definimos para el test, un valor de significancia α del 5 %.
De la etapas de filtrado, se obtienen los valores que figuran en la tabla 3.8:

Cuadro 8: Consumos promedios según la ciudad.

Categoría de ciudad	**Monto consumido promedio**
A	8958,01U$D
B	9198,66U$D
C	9844,44U$D

Confeccionamos entonces el estimador del test:

$$z_t = \frac{(\bar{x_x} - \bar{x_y}) - 0}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} + 2Cov\,(x_x, x_y)}} \tag{12}$$

Donde:

- z_t es el estimador de distribución t-Student.
- x_x es la media muestral de consumos de la ciudad con categoría X.

- x_y es la media muestral de consumos de la ciudad con categoría Y.
- s_x es el desvío muestral de consumos de la ciudad con categoría X.
- s_y es la desvío muestral de consumos de la ciudad con categoría Y.
- n_x es el número de consumos de la ciudad con categoría X.
- n_y es el número de consumos de la ciudad con categoría Y.
- $Cov\left(x_x, x_y\right)$ es la covarianza entre las medias muestrales de la gente de la ciudad con categoría X y de la gente de la ciudad con categoría Y.

Un detalle importante, es que el término de covarianza se anula puesto que no existe correlación entre las muestras. Las personas pertenecen a grupos distintos que se asumen sin relación alguna. Esto simplifica el estudio del error muestral del estimador.

Como se mencionó antes, el estimador z_t es un estimador que posee una distribución t-Student. Falta mencionar los grados de libertad de dicha distribución. El cómputo se realiza con la siguiente expresión:

$$GL\left(z_t\right) = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}} \tag{13}$$

Al realizar el test de hipótesis con el método de Welch, para las tres ciudades tomadas de a dos vemos que en todos los casos de compación el $p-valor$ resulta menor a $2,2{,}10^{-}16$ lo cual es considerablemente menor respecto a $\alpha/2$ (2,5 %). De esta forma, podemos rechazar la hipótesis nula en los tres casos.

Nota: se realizó un test similar al de Welch pero asumiendo varianzas muestrales iguales y no se obtuvo diferencia alguna en el resultado.

3.9. Consigna 9

La consigna se plantea como un test de hipótesis de diferencias de medias con medias y desvíos poblacionales distintos y desconocidos. Dicho eso, la metodología de resolución se basa en seguir los siguientes pasos:

- A partir de la base de datos, se obtienen los consumos, su estado civil, género y su edad. Se filtra por estado civil (casados/as) y por rango etario (26-35 años).
- Luego, se generan dos series según el género (F y M) con los datos de consumo.
- Se genera el estadístico en función de la hipótesis nula y alternativa. Se ejecuta el test.
- Se provee la conclusión.

Las hipótesis nula y alternativa se listan a continuación:

- *Hipótesis nula*: la diferencia de medias de consumo entre las mujeres casadas entre 26-35 años es menor o igual a la de los hombres casados entre 26-35 años.
- *Hipótesis alternativa*: la diferencia de medias de consumo entre las mujeres casadas entre 26-35 años es mayor a la de los hombres casados entre 26-35 años.

Definimos para el test, un valor de significancia α del 5 %.
De la etapas de filtrado, se obtienen los valores que figuran en la tabla 3.9:

Cuadro 9: Consumos promedios según el estado civil.

Género	**Monto consumido promedio**
Femenino	8988.08U$D
Masculino	9418.42U$D

Confeccionamos entonces el estimador del test:

$$z_t = \frac{(\bar{x_f} - \bar{x_m}) - 0}{\sqrt{\frac{s_f^2}{n_f} + \frac{s_m^2}{n_m} + 2Cov\left(x_f, x_m\right)}} \tag{14}$$

Donde:

- z_t es el estimador de distribución t-Student.
- x_f es la media muestral de consumos de gente de género femenino en el rango etario 26-35 años y casadas.
- x_m es la media muestral de consumos de gente de género masculino en el rango etario 26-35 años y casados.
- s_f es el desvío muestral de consumos de gente de género femenino en el rango etario 26-35 años y casadas.
- s_m es la desvío muestral de consumos de gente de género masculino en el rango etario 26-35 años y casados.
- n_f es el número de consumos de gente de género femenino en el rango etario 26-35 años y casadas.
- n_m es el número de consumos de gente de género masculino en el rango etario 26-35 años y casados.
- $Cov\left(x_f, x_m\right)$ es la covarianza entre las medias muestrales de la gente de género femenino en el rango etario 26-35 años y casadas y, la gente de género masculino en el rango etario 26-35 años y casados.

Un detalle importante, es que el término de covarianza se anula puesto que no existe correlación entre las muestras. Las personas pertenecen a grupos distintos que se asumen sin relación alguna. Esto simplifica el estudio del error muestral del estimador.

Como se mencionó antes, el estimador z_t es un estimador que posee una distribución t-Student. Falta mencionar los grados de libertad de dicha distribución. El cómputo se realiza con la siguiente expresión:

$$GL\left(z_t\right) = \frac{\left(\frac{s_f^2}{n_f} + \frac{s_m^2}{n_m}\right)^2}{\frac{\left(\frac{s_f^2}{n_f}\right)^2}{n_f - 1} + \frac{\left(\frac{s_m^2}{n_m}\right)^2}{n_m - 1}} \tag{15}$$

Finalmente, utilizando `R`, realizamos un test de hipótesis con el método de Welch y asumiendo que ambas varianzas muestrales son iguales. En ambos casos, obtenemos que con un $p - valor$ aproximadamente $1,0$ (por rendondeo) no hay evidencia suficiente para rechazar la hipótesis nula.

3.10. Consigna 10

Para resolver lo solicitado en el ejercicio se realiza el siguiente procedimiento:

- Filtrar la base para quedarse con las columnas: *Gender*
- Se toman muestras aleatorias de distintos tamaños para la columna de datos filtrada y se computan los estimadores de la consigna.
- El proceso anterior se replica n-veces para distintos valores de $n = 10, 20, 50, 100$. De esta manera se podrán computar los valores de media y varianza para los estimadores.
- Se realizan los tests de eficiencia, insesgadez y consistencia para los dos estimadores estudiados.

$$Estimador1 : \hat{p} = \frac{X}{n} \tag{16}$$

$$Estimador2 : \hat{p} = \frac{X}{n-c} \tag{17}$$

Análisis de insesgadez: Se considera insesgado el estimador que cumple

$$sesgo(\theta) = E(\hat{\theta}) - \theta = 0 \tag{18}$$

Análisis de eficiencia: Se prefiere estimador con menor error cuadrático medio.

$$ECM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] \tag{19}$$

Análisis de consistencia: Se considera consistente al estimador cuyo valor muestral tiende al valor real a medida que el tamaño n aumenta.

A continuación, se adjuntan gráficos con los resultados obtenidos para ambos estimadores.

Conclusiones:

Analizando la figura 6, puede observarse que el estimador p_1 es insesgado mientras que el estimador p_2 no lo es ya que $E(p_1) - p_0 = 0$ y $E(p_2) - p_0 \neq 0$. Esta conclusión se corrobora analizando el gráfico de la figura 8, donde se comparan los sesgos de manera directa.

La figura 7 permite concluir que ambos estimadores son consistentes, ya que su desvío estándar tiende a cero a valores cada vez mas grandes de n. Esto hace más probable que el estimador se .ªcerque.ªl parámetro.

En la figura 9 se comparan los errores cuadráticos medios de ambos estimadores, se concluye que el estimador p_1 es mas eficiente que p_2.

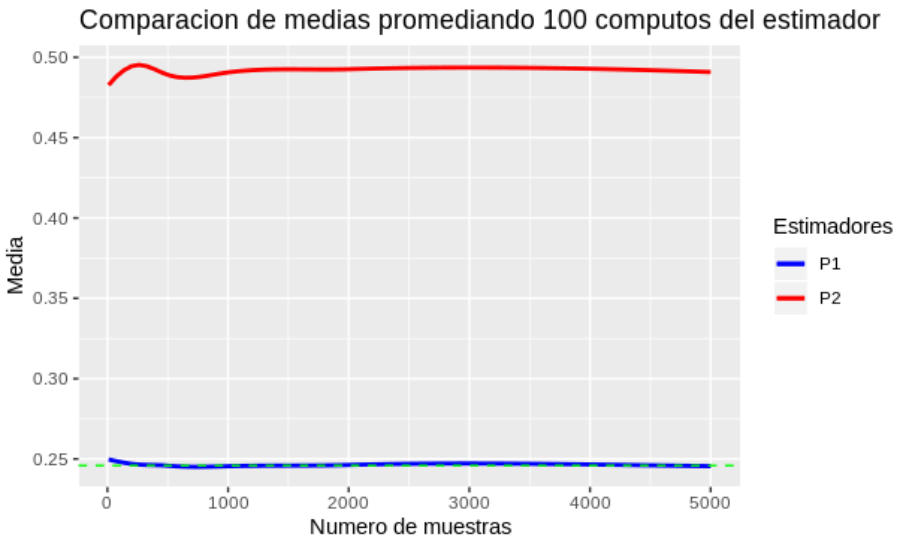

Figura 6: Evolución de la media con distintos tamaño de muestras para 100 promedios.

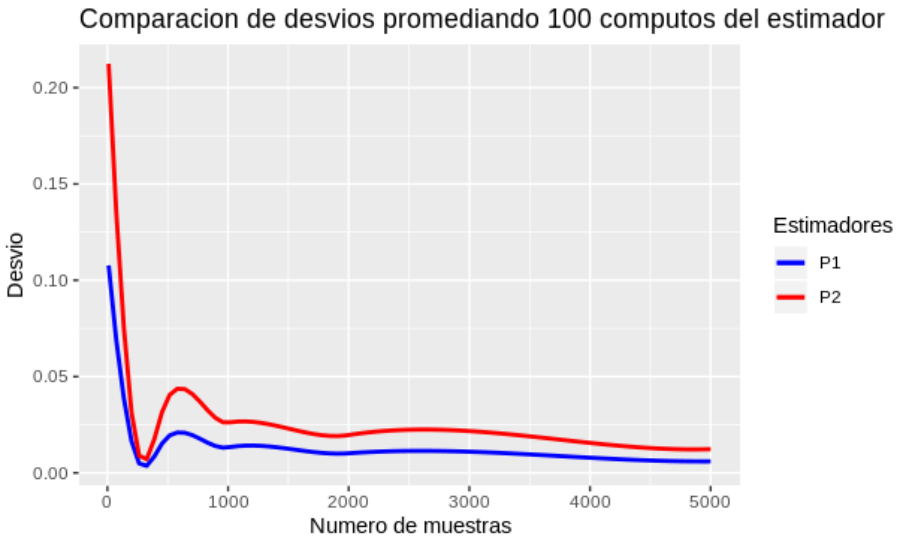

Figura 7: Evolución de los desvíos con distintos tamaño de muestras para 100 promedios.

Figura 8: Evolución del sesgo con distintos tamaño de muestras para 100 promedios.

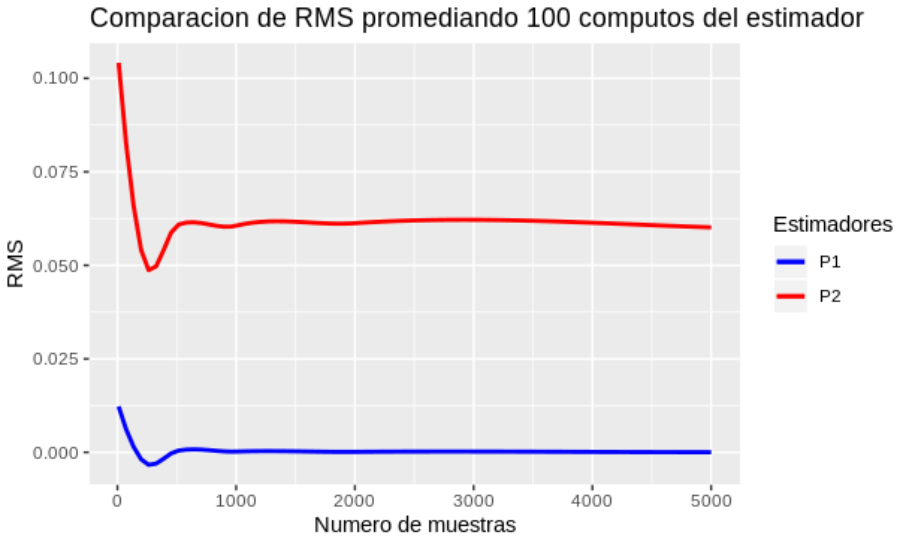

Figura 9: Evolución de RMS con distintos tamaño de muestras para 100 promedios.

A. Uso del script en R

Para el correcto uso del script en R adjunto, se deben instalar los archivos que figuran en el encabezado y modificar la variable `path_to_workspace` con la ruta al directorio que contiene el script. Luego, el script asumirá que la base de datos `BlackFriday.csv` se encuentra en el mismo directorio.

El script depende de las siguientes librerías:

- `dplyr`
- `ggplot2`

Cada una de las consignas se encuentra comentada y se asume que la ejecución de cada una de ellas dependen únicamente que la base de datos `black_friday` se encuentre en el entorno de desarrollo. Notar que previo a cargar la base todas las variables en el espacio de trabajo serán borradas.