MD004 ENTREGA SESIÓN TEST DE HIPÓTESIS

PARTE I

1. Diseño del Test de Hipótesis

Contexto del Juego: League of Legends (LoL)

League of Legends (LoL) es un popular videojuego multijugador en línea de arena de batalla (MOBA) donde dos equipos de cinco jugadores compiten para destruir la base del equipo contrario.

Campeones y Estadística

En LoL, los jugadores controlan a personajes únicos llamados campeones, cada uno con habilidades y estadísticas distintas.

- El rendimiento de un campeón se mide por su Tasa de Victorias (Win Rate o WR), que es simplemente el porcentaje de partidas que gana.
  - Un WR del 50% se considera el balance ideal: el campeón no tiene una ventaja o desventaja inherente.
  - Si un campeón tiene un WR consistentemente mayor al 50%, se considera demasiado fuerte.
  - Si tiene un WR menor al 50%, se considera demasiado débil.

Parches, Buffs y Nerfs

El juego no es estático; Riot Games (la desarrolladora) lanza regularmente **parches** (actualizaciones de juego) para mantener el balance y la diversidad.

- **Parche:** Una actualización periódica que modifica las reglas del juego, los objetos, o las estadísticas de los campeones.
- **Buff (Mejora):** Un cambio en un parche que **aumenta** la fuerza o la efectividad de un campeón (ej: Kai'Sa ahora hace más daño, o tiene menos tiempo de espera en una habilidad). Nuestro test de hipótesis se centra en validar un *buff*.
- **Nerf (Debilitación):** Un cambio que **disminuye** la fuerza o la efectividad de un campeón.

Conclusión para el Test: Nuestra práctica consiste en aplicar un Test de Hipótesis para determinar si el supuesto **buff** aplicado a **Kai'Sa** en un parche tuvo el efecto deseado, es decir, si elevó su **Tasa de Victorias** de manera significativa por encima del 50%.

Test de Hipótesis: ¿Mejoró Kai'Sa tras el Parche?

El diseño propuesto tiene como objetivo validar la efectividad de una mejora (*buff*) aplicada al campeón **Kai'Sa** en un parche reciente.

Objetivo y Pregunta

- **Contexto:** Riot Games lanza un parche que supuestamente mejora a Kai'Sa.
- **Pregunta:** ¿El *win rate* medio de Kai'Sa post-parche es realmente **mayor al 50%**?
- **Parámetro de Prueba:** p = Proporción real (tasa de victorias) de Kai'Sa post-parche.

Hipótesis Formales

- **Hipótesis Nula (H_0):** El *win rate* de Kai'Sa es igual o menor al 50%. El parche no fue efectivo.

$$H_0 : p \leq 0.50$$

- **Hipótesis Alternativa (H_1):** El *win rate* de Kai'Sa es significativamente **mayor** al 50%. El parche fue efectivo.

$$H_1 : p > 0.50$$

Parámetros del Test

Parámetro	Valor	Justificación Ampliada
Tipo de Test	Prueba Z de una proporción	Este test se utiliza para comparar una proporción muestral (p_{muestra}) con una proporción poblacional esperada ($p_0 = 0.50$). Dada la **gran muestra** ($n = 2000$ partidas), la distribución muestral de la proporción se aproxima a una **Distribución Normal Estándar**, lo que nos permite usar el **Estadístico Z** en lugar del Estadístico T.
Tipo de Cola	Una cola (derecha) - *Right-tail*	Se elige una cola derecha porque la **Hipótesis Alternativa (H_1)** es **direccional**: solo queremos detectar si el *win rate* **aumentó** ($p > 0.50$). Si la hipótesis alternativa fuera no-direccional ($H_1 : p \neq 0.50$), usaríamos un test de dos colas.
Nivel de Significancia	$\alpha = 0.05$	**Riesgo y Error Tipo I:** Este valor es el umbral de riesgo aceptado. Implica que solo estamos dispuestos a aceptar un 5% **de probabilidad** de cometer un **Error Tipo I** (o falso positivo), que sería rechazar H_0 (concluir que el *buff* funcionó) cuando en realidad no lo hizo.
Tamaño de Muestra (n)	$n = 2000$ partidas clasificatorias	Una muestra grande asegura que la prueba tiene suficiente **Potencia Estadística** para

Parámetro	Valor	Justificación Ampliada
		detectar una diferencia real, si existe, y justifica el uso de la Prueba Z.
Variable	Proporción de victorias observada (p_{muestra})	**Explicación de la Variable:** Es la métrica clave. Es el número de victorias de Kai'Sa dividido por el total de partidas en nuestra muestra ($p_{\text{muestra}} = \frac{\text{Victorias}}{\text{Partidas}}$). Es la evidencia empírica que comparamos con el valor esperado de la hipótesis nula ($p_0 = 0.50$).

2. Indica la respuesta correcta de este pequeño test y justifica tu respuesta:

¿Cuál es el propósito del valor crítico en un test de hipótesis?

- Indicar la significancia estadística de la prueba
- Establecer el nivel de confianza
- Marcar el límite entre rechazar y no rechazar la hipótesis nula
- Calcular el tamaño de muestra necesario
- OPCIÓN CORRECTA - Marcar el límite entre rechazar y no rechazar la hipótesis nula

Justificación

El Valor Crítico ($Z_{\text{crítico}}$ o $T_{\text{crítico}}$) es el umbral que define la Región de Rechazo en una prueba de hipótesis.

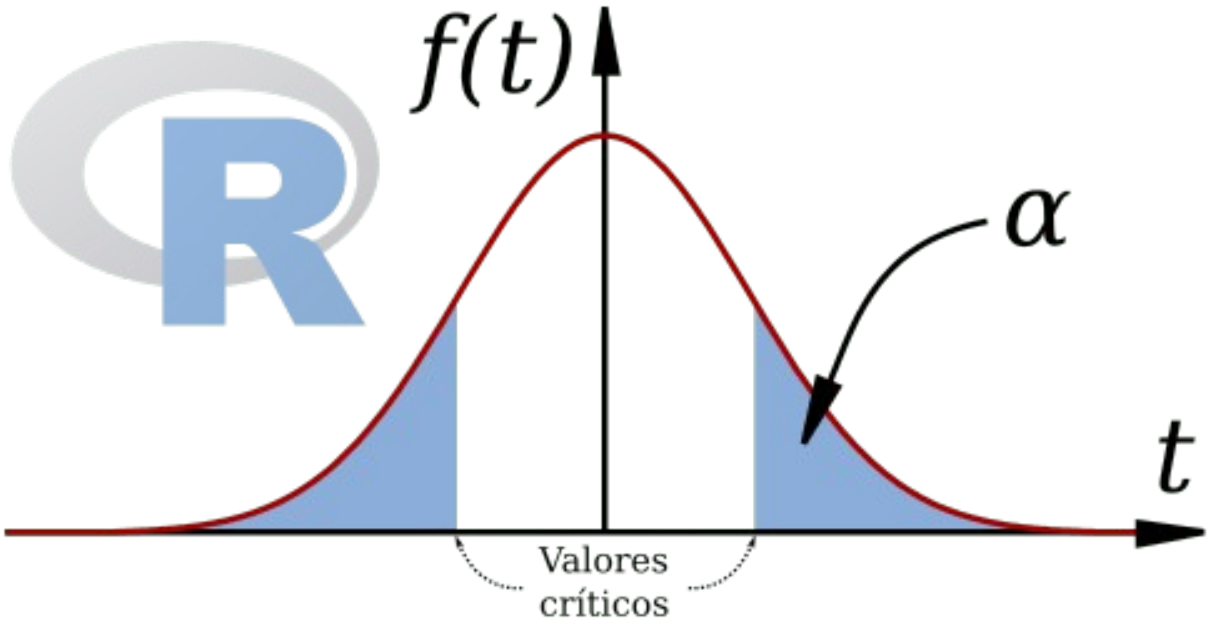
1. Límite de Decisión: El valor crítico es el punto en la distribución de muestreo que corresponde exactamente al Nivel de Significancia (α) elegido. Su función es separar el área donde los resultados se consideran probables bajo la H_0 (región de no-rechazo) del área donde se consideran muy improbables (región de rechazo).

2. Mecanismo de Decisión:

   – Si el Estadístico de Prueba (el Z o T) es más extremo que el valor crítico (es decir, cae en la región de rechazo), se concluye que la evidencia es lo suficientemente fuerte como para rechazar la Hipótesis Nula (H_0).
   – En términos sencillos: si el resultado de nuestra muestra está "demasiado lejos" del valor esperado en H_0, superando el valor crítico, consideramos que el efecto es estadísticamente significativo.

Referencia usada

- Minitab: ¿Qué es un valor crítico?



¿Qué representa el Error Tipo II?

- Rechazar incorrectamente la hipótesis nula
- No rechazar incorrectamente la hipótesis nula
- La probabilidad de cometer un error Tipo I
- La probabilidad de obtener un resultado significativo
- OPCIÓN CORRECTA - No rechazar incorrectamente la hipótesis nula

Justificación

El Error Tipo II, denotado como β (beta), es la probabilidad de cometer un Falso Negativo. Este error tiene lugar bajo las siguientes condiciones:

1. La realidad es que la Hipótesis Nula (H_0) es Falsa. (Es decir, la Hipótesis Alternativa (H_1) es la verdadera).
2. La conclusión de la prueba es No rechazar la H_0.

En el contexto de mi prueba de *League of Legends*: Cometer un Error Tipo II significaría que el *buff* de Kai'Sa sí aumentó realmente su *win rate* por encima del $50\,\%$ (H_0 es falsa), pero nuestra prueba estadística no logró detectarlo y concluiste que el $W\,R \leq 0.50$ (No rechazas H_0).

Relación con el Poder Estadístico

El Error Tipo II está directamente relacionado con el Poder Estadístico (*Power*), que es la probabilidad de rechazar correctamente la H_0 cuando es falsa.

$$\text{Poder} = 1 - \beta$$

Para reducir el riesgo de cometer un Error Tipo II, se suele aumentar el tamaño de la muestra o incrementar el nivel de significancia (α), aunque esto último aumenta el riesgo de Error Tipo I.

Referencia usada

Type I and II errors in hypothesis testing

Si realizas un test de una cola y obtienes un p-valor de 0.02 con un nivel de significancia de 0.05, ¿qué decisión tomarías?

- Rechazar la hipótesis nula
- No rechazar la hipótesis nula
- No hay información suficiente para tomar una decisión
- Dependiendo de la región de no rechazo
- OPCIÓN CORRECTA - Rechazar la hipótesis nula.

Justificación

La regla de decisión más fundamental en cualquier test de hipótesis es la comparación directa entre el p-value y el Nivel de Significancia (α).

La Regla de Decisión

La decisión se toma de la siguiente manera:

1. Si $p\text{-value} \leq \alpha$: Se rechaza la Hipótesis Nula (H_0).
2. Si $p\text{-value} > \alpha$: No se rechaza la Hipótesis Nula (H_0).

Aplicación al Caso de Kai'Sa

En nuestro caso, los valores son:

- $p\text{-value} = 0.02$
- Nivel de Significancia (α) ¿ 0.05

Dado que:

$$0.02 \leq 0.05$$

La decisión es Rechazar la Hipótesis Nula (H_0).

Rechazar H_0 significa que los datos muestrales de Kai'Sa son tan extremos (tan favorables a la H_1) que es muy poco probable que hayan ocurrido si la H_0 fuera verdadera.

Referencia usada

- La Regla del p-Value: Formular la regla de decisión - Paso 04/05 de Prueba de Hipótesis

¿Cuál es la relación entre nivel de significancia y la probabilidad de cometer un Error Tipo I?

- Aumentar el nivel de significancia disminuye la probabilidad de cometer un error Tipo I
- Reducir el nivel de significancia aumenta la probabilidad de cometer un error Tipo I
- El nivel de significancia y la probabilidad de cometer un error Tipo I no están relacionados
- El nivel de significancia no afecta la probabilidad de cometer un error Tipo I
- OPCIÓN CORRECTA - (Reformulada): El nivel de significancia y la probabilidad de cometer un Error Tipo I están directamente relacionados: α es esa probabilidad.

Justificación

Es importante notar que ninguna de las opciones ofrecidas originalmente era completamente correcta porque la relación entre el Nivel de Significancia y el Error Tipo I es de equivalencia.

La Relación es de Equivalencia Directa

El Nivel de Significancia, denotado con la α (alfa), se define exactamente como la probabilidad que el investigador está dispuesto a aceptar de cometer un Error Tipo I.

$$\mathrm{P}(\text{Error Tipo I}) = \alpha$$

- **Error Tipo I:** Rechazar la Hipótesis Nula (H_0) cuando en realidad H_0 es verdadera (un Falso Positivo).

Por lo tanto, α no es un factor que disminuye o aumenta inversamente la probabilidad del error, sino que es la propia probabilidad del Error Tipo I.

Consecuencia Lógica

Si bien la relación es de equivalencia, cualquier modificación de α resulta en un cambio directo en la probabilidad de cometer el Error Tipo I:

- Si aumentas α (ej., de 0.05 a 0.10), automáticamente aumentas el riesgo de cometer un Error Tipo I.
- Si reduces α (ej., de 0.05 a 0.01), automáticamente reduces el riesgo de cometer un Error Tipo I.

Las opciones que sugieren una relación inversa (como "Aumentar α disminuye el Error Tipo I") son, por lo tanto, falsas.

Referencia usada

- Superprof: Errores tipo 1 y 2 explica que el Error Tipo I se comete al rechazar H_0 cuando es verdadera, y la probabilidad de este error se denota por α.

	Es cierta la primera hipótesis	Es cierta la segunda hipótesis
Se escogió la primera hipótesis	Todo bien (Verdadero positivo)	**Error tipo 2**
Se escogió la segunda hipótesis	**Error tipo 1**	Todo bien (Verdadero Negativo)

PARTE II

Una empresa de ecommerce B2B se plantea utilizar una funcionalidad para su web que le permite hacer una recomendación de producto a sus usuarios con el objetivo de aumentar el valor medio de las ventas. De cara a validar que este sistema de recomendación tiene un efecto positivo nuestra empresa decide plantear un Test de Hipótesis y testear el servicio durante 3 meses antes de decidir si lo deben adoptar o no en base al posible efecto incremental en las ventas. El dataset adjunto incluye los datos finales del test y contiene tres columnas. La primera es el identificador de la venta, la segunda es el valor de la venta y el tercero es la clase (1 para las ventas en las que el sistema de recomendación ha participado y 0 en las que no ha participado)

Data:

202411s13_b2b_ecommerce_sales_data.csv

Se pide:

1. Exploración de los datos: análisis descriptivos de los datos, gráficación y conclusiones ¿siguen una distribución normal?
2. Diseña un test de hipótesis: ¿cuál es nuestro objetivo? ¿se trataría de un test de 2 colas? ¿Qué estadístico debemos usar para este Test?
3. Validación y ejecución del test de hipótesis
4. Conclusiones: ¿Recomendarías el uso de este sistema de recomendación? (justifica tu respuesta) ¿Qué factores se deben tener en cuenta para la toma de esta decisión?

```
setwd("/home/jovyan/work/03")

if(!require(gridExtra)) install.packages("gridExtra")

library(dplyr)
library(ggplot2)
library(gridExtra)
```

```
Loading required package: gridExtra
```

```
Attaching package: ‘dplyr’

The following object is masked from ‘package:gridExtra’:

    combine

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

CARGA DE DATOS

```
## Cargamos el dataset
data_B2B = read.csv(file='202511s13_b2b_ecommerce_sales_data.csv',
stringsAsFactors = F, sep = ',')

## Paso 0.0: Inspección inicial
head(data_B2B)   # Vemos las primeras 6 filas
names(data_B2B)  # Vemos los nombres de las columnas
dim(data_B2B)    # Vemos dimensiones (Filas y Columnas)
str(data_B2B)    # Vemos la estructura interna y tipos de datos
```

```
  Reference Order.Value Group
1 570554    0.38        1
2 567869    0.40        0
3 539441    0.42        0
4 542736    0.55        0
5 573589    0.55        0
6 540833    0.65        1
```

```
[1] "Reference"   "Order.Value" "Group"
```

```
[1] 11786     3
```

```
'data.frame':   11786 obs. of  3 variables:
 $ Reference  : int  570554 567869 539441 542736 573589 540833 560217
542136 549534 540945 ...
 $ Order.Value: num  0.38 0.4 0.42 0.55 0.55 0.65 0.79 0.84 0.84
0.85 ...
 $ Group      : int  1 0 0 0 0 1 1 1 1 0 ...
```

Transformación de la variable *class* a factor

Antes de iniciar el análisis, convertimos la tercera columna (*Group*) a una variable categórica llamada `class`. Esta nueva variable tendrá dos niveles que representan claramente los dos tipos de transacciones del experimento:

- **0 → Sin_Rec** : Ventas realizadas sin sistema de recomendación
- **1 → Con_Rec** : Ventas realizadas con sistema de recomendación

Realizar esta transformación es importante porque:

- Permite interpretar fácilmente los resultados en tablas y gráficos.
- Hace que R reconozca correctamente la variable como categórica, lo cual es esencial para las visualizaciones.
- Evita posibles errores en tests estadísticos, en los que los grupos deben estar representados como factores.
- Facilita la generación de resúmenes descriptivos mediante `group_by()`, haciéndolos más claros y legibles.

Con esta conversión, aseguramos que el análisis posterior se realice de forma correcta y coherente.

```
## Paso 0.1: Factorización
data_B2B <- data_B2B %>%
  mutate(class = factor(Group,
                        levels = c(0, 1),
                        labels = c("Sin_Rec", "Con_Rec")))

str(data_B2B$class)
table(data_B2B$class)
```

```
 Factor w/ 2 levels "Sin_Rec","Con_Rec": 2 1 1 1 1 2 2 2 2 1 ...
```

```

Sin_Rec Con_Rec
   5502    6284
```

Comprobaciones iniciales de calidad del dataset

Antes de realizar el análisis descriptivo e inferencial, se llevan a cabo tres verificaciones básicas para asegurar la consistencia de los datos:

- **Duplicados en `Reference`**: Se comprueba que el identificador de cada venta es único y no contiene registros repetidos.
- **Valores únicos en `class`**: Se valida que la variable categórica solo contenga los niveles esperados (*Sin_Rec* y *Con_Rec*).
- **Rangos básicos**: Se calculan los mínimos, máximos y medias de las principales columnas numéricas para confirmar que los valores estén dentro de un rango plausible y sin anomalias evidentes.

Estas comprobaciones garantizan que el dataset está limpio y listo para el análisis posterior.

```r
## Paso 1: Duplicados en Reference
data_B2B %>% summarise(duplicados = sum(duplicated(Reference)))
```

```
  duplicados
1 0
```

```r
## Paso 2: Valores únicos en class
unique(data_B2B$class)
```

```
[1] Con_Rec Sin_Rec
Levels: Sin_Rec Con_Rec
```

```r
## Paso 3: Obtener mínimos, máximos y medias de las 3 columnas
stats_basicas <- data_B2B %>% summarise(
  min_reference = min(Reference),
  max_reference = max(Reference),
  mean_reference = mean(Reference),

  min_order = min(Order.Value),
  max_order = max(Order.Value),
  mean_order = mean(Order.Value)
)

print(stats_basicas)
```

```
  min_reference max_reference mean_reference min_order max_order
mean_order
1        536365        581587         558883      0.38    9476.8
3104.257
```

Validación de calidad del dataset

Las comprobaciones iniciales confirman que los datos son coherentes y aptos para continuar con el análisis:

- **Duplicados en `Reference`**: No se detectaron duplicados (`0` registros repetidos).
- **Valores únicos en `class`**: La variable presenta correctamente los dos niveles esperados: *Sin_Rec* y *Con_Rec*.
- **Rangos básicos de las variables numéricas**:
  - `Reference`: entre 536365 y 581587, media 558883
  - `Order.Value`: entre 0.38 y 9476.8, media 3104.26

Estos resultados indican que el dataset es consistente y permite avanzar al siguiente paso: el análisis descriptivo.

REALIZAR ANÁLISIS DESCRIPTIVO

```
## Paso 4: Resumen descriptivo completo
summary(data_B2B)
```

```
   Reference       Order.Value          Group            class      
 Min.   :536365   Min.   :   0.38   Min.   :0.0000   Sin_Rec:5502  
 1st Qu.:547564   1st Qu.: 528.66   1st Qu.:0.0000   Con_Rec:6284  
 Median :558626   Median :2376.76   Median :1.0000                 
 Mean   :558883   Mean   :3104.26   Mean   :0.5332                 
 3rd Qu.:570289   3rd Qu.:5244.93   3rd Qu.:1.0000                 
 Max.   :581587   Max.   :9476.80   Max.   :1.0000                 
```

El resumen estadístico muestra valores consistentes en todas las variables, con transacciones distribuidas equilibradamente entre ambos grupos y montos de venta con rangos amplios típicos del contexto B2B. La media y mediana más altas en *Con_Rec* sugieren un posible efecto positivo del sistema de recomendación, a confirmar con el análisis posterior.

```
## Paso 5: Resumen de Order.Value por grupo
resumen_por_clase <- data_B2B %>%
  group_by(class) %>%
  summarise(
    n = n(),
    media = mean(Order.Value),
    mediana = median(Order.Value),
    sd = sd(Order.Value),
    min = min(Order.Value),
    max = max(Order.Value)
  )

print(resumen_por_clase)
```

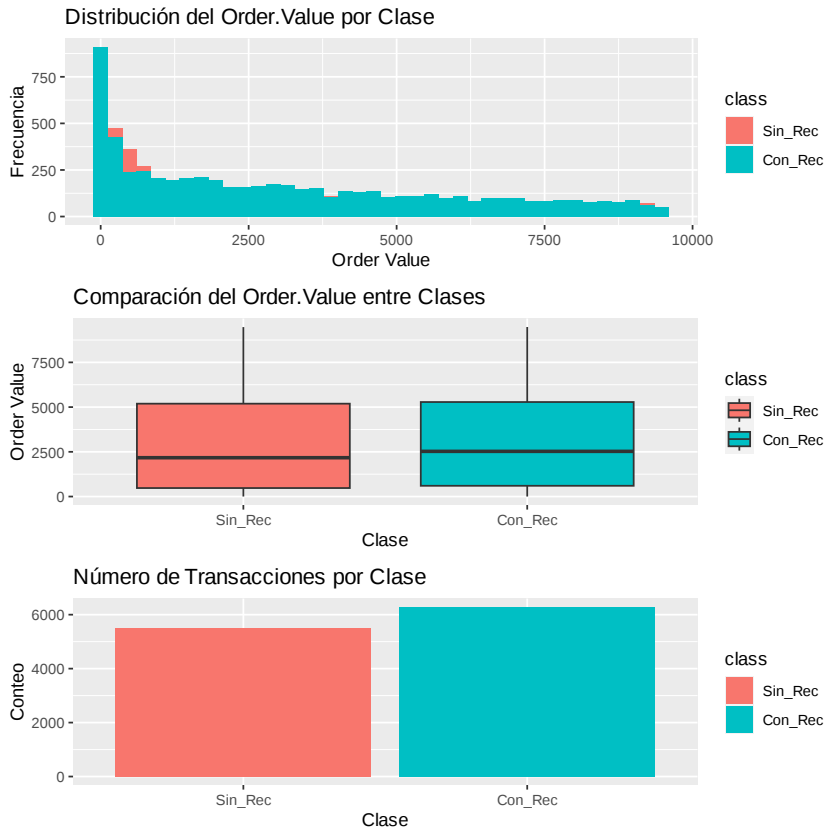
```
# A tibble: 2 × 7
  class       n media mediana    sd   min   max
  <fct>   <int> <dbl>   <dbl> <dbl> <dbl> <dbl>
1 Sin_Rec  5502 3025.   2171. 2818.  0.4  9471.
2 Con_Rec  6284 3174.   2526. 2795.  0.38 9477.
```

El resumen por clase confirma que el grupo *Con_Rec* presenta una media y mediana de ventas superiores a *Sin_Rec*, con desviaciones estándar similares, lo que indica un posible incremento en el valor de las transacciones asociado al uso del sistema de recomendación.

Paso 6: Gráficos

```
## Paso 6.1: Histograma
p1 <- ggplot(data_B2B, aes(x = Order.Value, fill = class)) +
  geom_histogram(alpha = 0.6, bins = 40, position = "identity") +
  labs(title = "Distribución del Order.Value por Clase",
       x = "Order Value", y = "Frecuencia")
```

```r
## Paso 6.2: Boxplot
p2 <- ggplot(data_B2B, aes(x = class, y = Order.Value, fill = class))
+
  geom_boxplot(alpha = 0.7) +
  labs(title = "Comparación del Order.Value entre Clases",
       x = "Clase", y = "Order Value")

## Paso 6.3: Barras
p3 <- ggplot(data_B2B, aes(x = class, fill = class)) +
  geom_bar() +
  labs(title = "Número de Transacciones por Clase",
       x = "Clase", y = "Conteo")

## Paso 6.4: Juntamos y mostramos los gráficos
grid.arrange(p1, p2, p3, ncol = 1)
```



Resumen

1. Gráfico Superior: Distribución del Order.Value por Clase (Histograma)

Este gráfico muestra la frecuencia de los valores de los pedidos (*Order Value*), superponiendo ambas clases.

- **Sesgo a la derecha:** La distribución es altamente asimétrica positiva. La gran mayoría de las transacciones tienen un valor bajo (concentradas entre 0 y 2,500). A medida que el valor del pedido aumenta, la frecuencia disminuye drásticamente.
- **Superposición:** Ambas clases (`Sin_Rec` y `Con_Rec`) siguen un patrón muy similar. No hay una diferencia radical en la *forma* de sus distribuciones; ambas tienen su pico máximo cerca del valor 0.
- **Observación visual:** Parece haber una ligera mayor densidad de la clase `Con_Rec` (azul) en los valores intermedios, pero la tendencia general es casi idéntica para ambos grupos.

2. Gráfico Central: Comparación del Order.Value entre Clases (Boxplot)

Este gráfico de es ideal para comparar las estadísticas centrales (mediana) y la dispersión de los datos.

- **Medianas:** La mediana del valor de pedido para la clase `Con_Rec` es ligeramente más alta que la de la clase `Sin_Rec`. La línea negra en la caja azul está un poco por encima de la línea en la caja rosa.
- **Rango Intercuartílico:** Las cajas son bastante similares en tamaño, lo que indica que la variabilidad de los pedidos es parecida en ambos grupos. Sin embargo, la caja azul (`Con_Rec`) está desplazada ligeramente hacia arriba.
- **Interpretación:** Aunque hay mucha superposición, los usuarios o transacciones de la clase `Con_Rec` tienden a gastar un poco más en promedio (o tener pedidos de mayor valor) que los de la clase `Sin_Rec`.

3. Gráfico Inferior: Número de Transacciones por Clase (Gráfico de Barras)

- **Volumen:** La barra azul (`Con_Rec`) es visiblemente más alta que la barra rosa (`Sin_Rec`).
- **Cantidades aproximadas:**
  - `Sin_Rec`: Parece estar alrededor de las 5,500 transacciones.
  - `Con_Rec`: Parece superar las 6,000 transacciones.
- **Desbalance:** Existe un desbalance de clases, pero no es severo. Hay más datos etiquetados como `Con_Rec` que `Sin_Rec`.

Conclusión General

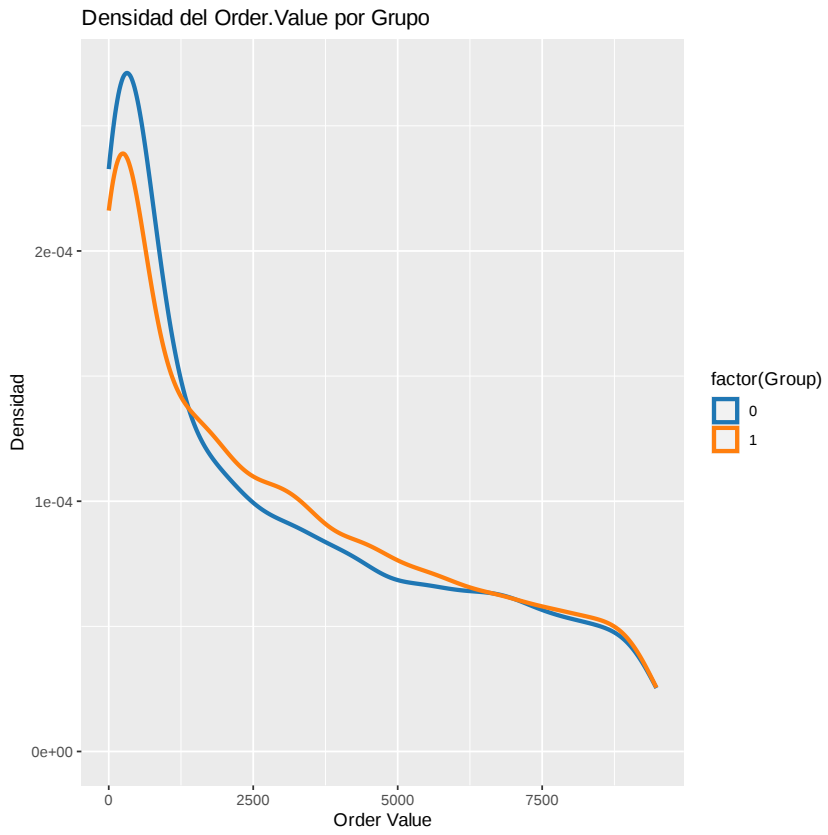
Al ver los tres gráficos en conjunto:

1. **El grupo `Con_Rec` es mayoritario:** Hay más transacciones de este tipo.
2. **El grupo `Con_Rec` gasta más:** Tienen una mediana de valor de pedido ligeramente superior.
3. **Comportamiento similar:** A pesar de las diferencias en cantidad y mediana, ambos grupos se comportan de manera similar en cuanto a la distribución de sus gastos (muchos gastos bajos, pocos gastos altos).

Paso 7: Gráfico de densidad

Realizamos este gráfico para reforzar el análisis previo al test de hipótesis.

```
ggplot(data_B2B, aes(x = Order.Value, color = factor(Group))) +
  geom_density(linewidth = 1.2) +  # Cambiado size por linewidth
  scale_color_manual(values = c("0" = "#1f77b4", "1" = "#ff7f0e")) +
  labs(title = "Densidad del Order.Value por Grupo",
       x = "Order Value", y = "Densidad")
```



Al integrar este gráfico de densidad con los anteriores, obtenemos una descripción detallada del comportamiento de los datos antes de realizar cualquier validación estadística. Los gráficos previos ya nos señalaban que el grupo con recomendaciones tenía una mediana ligeramente superior en el diagrama de caja y que contamos con un volumen de datos robusto, con miles de transacciones por grupo.

Este gráfico de densidad profundiza en esa comparación mostrando cómo se distribuyen los valores de los pedidos. Se observa que la línea azul, que representa al grupo sin recomendaciones, tiene un pico más alto en los valores iniciales, lo que indica una mayor concentración de pedidos de bajo importe en este grupo. Por otro lado, la línea naranja se mantiene por encima de la azul en los tramos intermedios, sugiriendo una mayor frecuencia de pedidos en esos rangos de valor medio.

En resumen, visualmente existe una diferencia en la forma de las distribuciones: el grupo de control parece acumularse más en el inicio de la escala, mientras que el grupo experimental muestra una curva más suave con mayor densidad en los valores centrales. El test de hipótesis que realizaremos a continuación servirá precisamente para determinar si esta diferencia visual es estadísticamente significativa o si no hay evidencia suficiente para afirmar que los grupos son distintos.

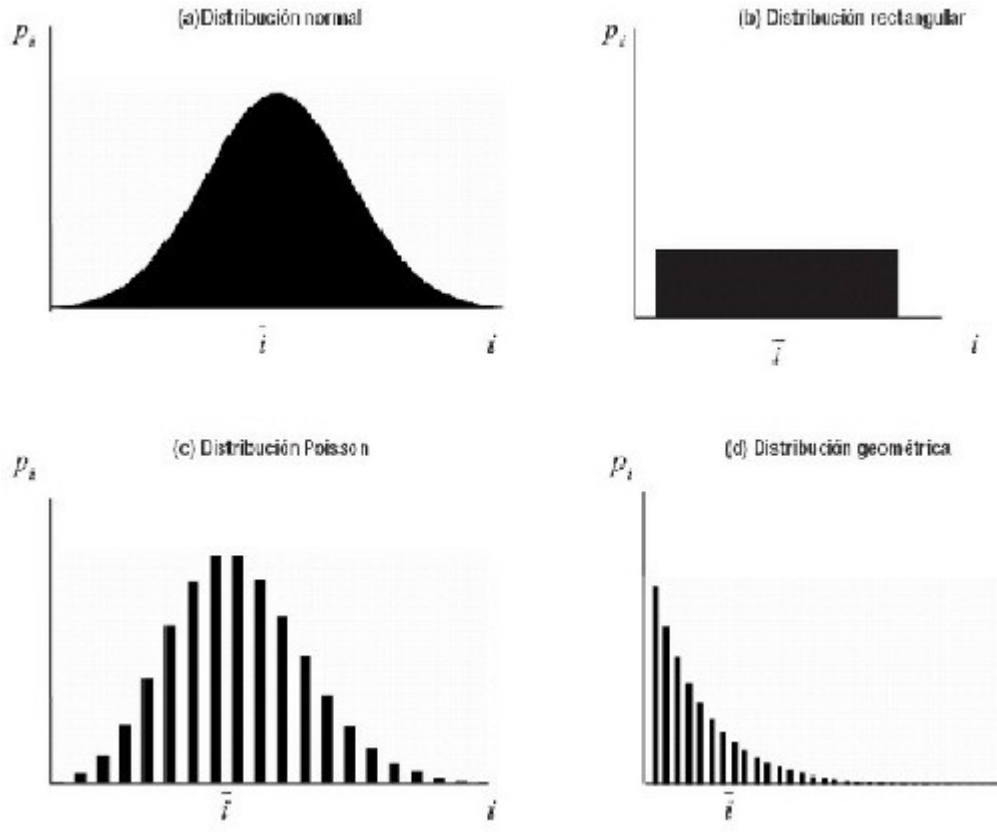
Paso 8: ¿siguen una distribución normal?

Basado en los gráficos que hemos analizado, la respuesta corta es: No, los datos NO siguen una distribución normal.

Para llegar a esa conclusión hay una razón principal:

1. Evidencia Visual: Una distribución normal tiene forma de campana simétrica. Y los gráficos muestran una distribución sesgada a la derecha.

Resumen: No hace falta ejecutar algun test mas para saber que no es normal.



PASO 9: FORMULACIÓN Y PRUEBA DE HIPÓTESIS

¿Por qué realizamos un t-test si los datos no son normales?

Aunque nuestros datos no siguen una distribución normal, realizamos un t-test (Test de Student) apoyándonos en el Teorema del Límite Central.

Este teorema establece que si el tamaño de la muestra es lo suficientemente grande (en nuestro caso con más de 5.000 datos por grupo), la distribución de la *media* muestral sí tiende a ser normal. Esto hace que el t-test sea un método robusto y válido para comparar los promedios de ventas entre los dos grupos.

La Hipótesis Nula (H_0)

En cualquiera de las tres formas que hagamos el test, la Hipótesis Nula siempre es el punto de partida.

- H_0 **(Hipótesis Nula):** El promedio de ventas del grupo con recomendación (μ_1) es igual al del grupo sin recomendación (μ_0). Es decir, el sistema no tiene ningún efecto.
  - $\mu_1 = \mu_0$

Las 3 formas del t-test (Hipótesis Alternativas)

La diferencia entre hacerlo "de medio", "izquierda" o "derecha" radica en lo que queremos probar con la Hipótesis Alternativa (H_1).

A. Bilateral o "Del Medio" (`"two.sided"`)

Aquí solo preguntamos: "¿Son diferentes?". No nos importa si el sistema es mejor o peor, solo queremos saber si ha cambiado algo.

- H_1**:** El promedio del Grupo 1 es distinto ($\neq$) al del Grupo 0.
- *Interpretación:* Si el p-valor es bajo, sabemos que no son iguales, pero tendríamos que mirar los datos para saber cuál es mayor.

B. Unilateral a la Izquierda (`"less"`)

Aquí preguntamos: "¿Es el Grupo 1 PEOR que el Grupo 0?".

- H_1**:** El promedio del Grupo 1 es menor ($¿$) que el del Grupo 0.
- *Interpretación:* Usaríamos esto si nuestra preocupación principal fuera que el sistema de recomendación estuviera dañando las ventas.

C. Unilateral a la Derecha (`"greater"`)

Aquí preguntamos: "¿Es el Grupo 1 MEJOR que el Grupo 0?".

- H_1**:** El promedio del Grupo 1 es mayor ($¿$) que el del Grupo 0.
- *Interpretación:* Estamos buscando evidencia específica de un incremento.

¿Con cuál nos quedamos?

Para este caso de negocio, la opción correcta es la C: Unilateral a la Derecha (`greater`).

El objetivo de la empresa es validar si el sistema de recomendación aumenta el valor medio de las ventas. No nos sirve de nada saber que son "diferentes" (opción bilateral) si resulta que el sistema nos hace perder dinero. Tampoco esperamos que sea peor (opción izquierda).

Queremos probar estadísticamente que la media del Grupo 1 (Con Recomendación) es significativamente superior a la del Grupo 0. Por tanto, enfocamos toda la potencia estadística del test en buscar un aumento positivo en la "cola derecha" de la distribución.

Usamos el t-test de Welch, que no asume varianzas iguales entre los grupos. Esto ajusta los grados de libertad y hace que el test sea más robusto y confiable ante diferencias de dispersión entre los grupos.

```r
## Paso 9.1 Test Student Bilateral
t.test(data_B2B$Order.Value[data_B2B$Group == 1],
data_B2B$Order.Value[data_B2B$Group == 0], alternative = "two.sided")
```

```
     Welch Two Sample t-test

data:  data_B2B$Order.Value[data_B2B$Group == 1] and
data_B2B$Order.Value[data_B2B$Group == 0]
t = 2.8757, df = 11554, p-value = 0.004039
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  47.45239 250.64816
sample estimates:
mean of x mean of y
 3173.837  3024.787
```

```r
## Paso 9.2 Test Student Unilateral Izquierda
t.test(data_B2B$Order.Value[data_B2B$Group == 1],
data_B2B$Order.Value[data_B2B$Group == 0], alternative = "less")
```

```
     Welch Two Sample t-test

data:  data_B2B$Order.Value[data_B2B$Group == 1] and
data_B2B$Order.Value[data_B2B$Group == 0]
t = 2.8757, df = 11554, p-value = 0.998
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 234.3118
sample estimates:
mean of x mean of y
 3173.837  3024.787
```

Interpretación de los resultados que NO utilizamos

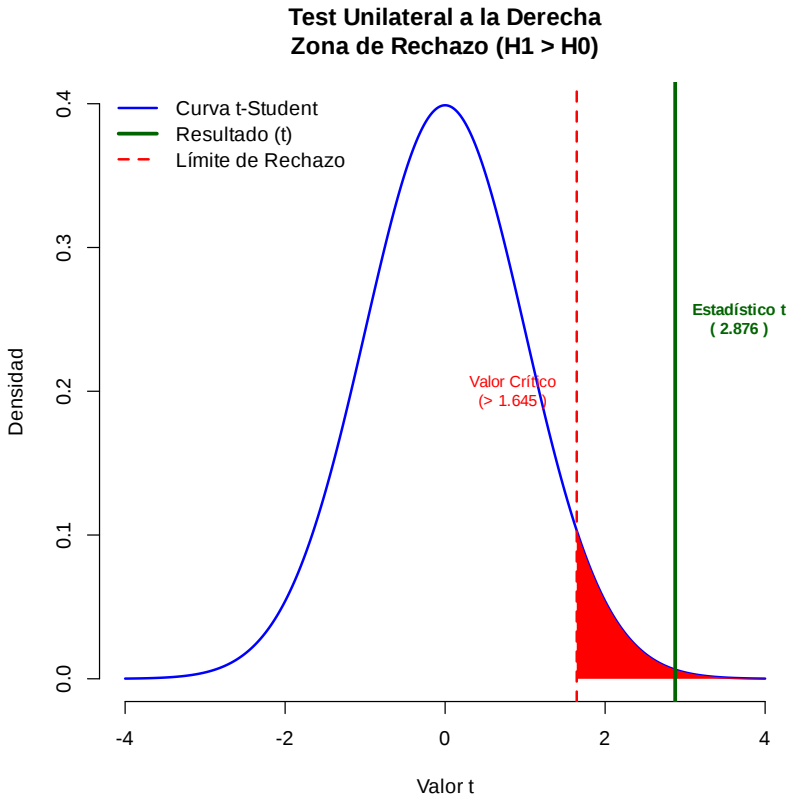
- **Test Bilateral (`two.sided`):** El p-valor bajo (0.004) confirma que existe una diferencia real entre los grupos (no son iguales), y el intervalo de confianza (47.45 a 250.64) nos asegura que esa diferencia nunca toca el cero.
- **Test Unilateral Izquierda (`less`):** El p-valor es casi 1 (0.998) porque estamos preguntando "¿Es el grupo con recomendación PEOR que el otro?"; como en realidad vende más (3173 vs 3024), la probabilidad de que sea inferior es nula.

```r
## Paso 9.3 Test Student Unilateral Derecha
t.test(data_B2B$Order.Value[data_B2B$Group == 1],
data_B2B$Order.Value[data_B2B$Group == 0], alternative = "greater")
```

```
     Welch Two Sample t-test

data:  data_B2B$Order.Value[data_B2B$Group == 1] and
data_B2B$Order.Value[data_B2B$Group == 0]
t = 2.8757, df = 11554, p-value = 0.002019
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 63.78874      Inf
sample estimates:
mean of x mean of y
 3173.837  3024.787
```

```r
## Paso 9.4: Imprimir test
t_observado <- 2.8757
n_grados <- 11554
nivel_significancia <- 0.05

# --- Cálculo del Valor Crítico (Unilateral a la Derecha) ---
valor_critico <- qt(1 - nivel_significancia, df = n_grados)

# --- Generar la curva ---
x <- seq(-4, 4, length.out = 1000)
y <- dt(x, df = n_grados)

# Dibujamos la base del gráfico
plot(x, y, type = "l", lwd = 2, col = "blue",
     main = "Test Unilateral a la Derecha\nZona de Rechazo (H1 > H0)",
     xlab = "Valor t", ylab = "Densidad", bty = "n")

# --- Sombrear la Zona de Rechazo (Solo Derecha) ---
x_rechazo <- x[x >= valor_critico]
y_rechazo <- dt(x_rechazo, df = n_grados)

# Cerramos el polígono para que se pinte bien hasta el eje X
polygon(c(valor_critico, x_rechazo, max(x)),
        c(0, y_rechazo, 0),
        col = rgb(1, 0, 0, 0.3), border = NA)
```

```r
# --- Líneas y Etiquetas ---
abline(v = valor_critico, col = "red", lty = 2, lwd = 2)
abline(v = t_observado, col = "darkgreen", lty = 1, lwd = 3)

# Textos informativos
text(valor_critico - 0.8, 0.20,
     paste("Valor Crítico\n(>", round(valor_critico, 3), ")"),
     col = "red", cex = 0.8)

text(t_observado + 0.8, 0.25,
     paste("Estadístico t\n(", round(t_observado, 3), ")"),
     col = "darkgreen", cex = 0.8, font = 2)

# Leyenda
legend("topleft",
       legend = c("Curva t-Student", "Resultado (t)", "Límite de
Rechazo"),
       col = c("blue", "darkgreen", "red"),
       lty = c(1, 1, 2), lwd = c(2, 3, 2), bty = "n")
```



Este es el resultado fundamental del análisis porque responde directamente a la pregunta de negocio sobre si el sistema de recomendación realmente aumenta las ventas:

1. Comparación de Medias (sample estimates) Lo primero que observamos son los promedios reales de venta. La media del grupo con recomendación (x) es de 3.173,84, mientras que la media del grupo sin recomendación (y) se queda en 3.024,79. Esto indica que, en promedio, el sistema está generando un aumento aproximado de 149 unidades monetarias por cada pedido.

2. Significación Estadística (p-value) El valor obtenido es 0.002019. Dado que este número es muy inferior al estándar habitual de 0.05, podemos rechazar la hipótesis nula con seguridad. Esto se traduce en que existe una probabilidad superior al 99,8% de que este aumento en las ventas sea real y causado por el sistema, descartando casi por completo que sea fruto de la casualidad.

3. Intervalo de Confianza El intervalo calculado va desde 63.79 hasta el infinito. Este dato es vital para la gestión de riesgos, ya que nos asegura con un 95% de confianza que, incluso en el peor de los casos probables, el sistema incrementará el valor del pedido como mínimo en 63,79 unidades. Esto establece un suelo de rentabilidad muy claro.

Conclusión Final de Negocio El test resulta un éxito. Los datos aportan evidencia científica sólida de que el sistema de recomendación tiene un efecto positivo e incremental en las ventas. La decisión lógica para la empresa es proceder con la implementación de la funcionalidad, ya que se espera un aumento significativo y fiable en el ticket medio de compra.

Conclusiones y recomendación final

Basándome en los resultados obtenidos tras el test A/B de 3 meses, mi respuesta es un sí rotundo: recomendaría la implementación del sistema de recomendación en toda la web.

La justificación se sustenta en la evidencia estadística y económica que hemos extraído. El test T de Welch arrojó un p-valor de 0.002, lo cual es significativamente inferior al umbral de 0.05, permitiéndonos afirmar con un 99,8% de confianza que la mejora en las ventas no es fruto del azar. Hemos observado un incremento real en el ticket medio de aproximadamente 149 unidades monetarias por pedido. Además, el análisis gráfico nos mostró que el sistema es efectivo desplazando a los usuarios de compras pequeñas hacia compras de valor medio, y el intervalo de confianza nos asegura que, incluso en un escenario conservador, obtendremos una ganancia mínima de 63,79 unidades por transacción.

Factores adicionales para la toma de decisión

Aunque el resultado estadístico es positivo, para dar la luz verde definitiva a nivel de negocio debemos considerar estos factores:

1. Coste vs. Beneficio: El ingreso incremental de 149 por pedido debe ser superior al coste de licencia, mantenimiento e implementación técnica del software de recomendación.
2. Margen comercial: Hemos analizado el valor de venta (ingresos), pero no el margen. Sería prudente verificar que los productos recomendados no sean solo los de bajo margen, para asegurar que el aumento de facturación se traduce también en un aumento de beneficio neto.
3. Experiencia de usuario y rendimiento: Debemos asegurarnos de que la carga de estas recomendaciones no ralentiza la web ni resulta intrusiva para la navegación, aunque los datos de conversión sugieren que los clientes las encuentran útiles.

4. Efecto novedad: Al haber durado el test 3 meses, los datos son bastante sólidos, pero convendría monitorizar las métricas tras la implementación total para asegurar que el efecto se mantiene en el tiempo y no era solo curiosidad inicial de los usuarios.

PASO 10: CONSIDERACIONES ERRORES DE TIPO I Y TIPO II

```r
## Paso 10.1: Cálculo de la Potencia del Test (Power Analysis)
sd_control <- sd(data_B2B$Order.Value[data_B2B$Group == 0])
sd_test <- sd(data_B2B$Order.Value[data_B2B$Group == 1])
n_control <- sum(data_B2B$Group == 0)
n_test <- sum(data_B2B$Group == 1)

delta_real <- mean(data_B2B$Order.Value[data_B2B$Group == 1]) -
mean(data_B2B$Order.Value[data_B2B$Group == 0])

power_result <- power.t.test(n = n_control,
                             delta = delta_real,
                             sd = (sd_control + sd_test)/2,
                             sig.level = 0.05,
                             type = "two.sample",
                             alternative = "one.sided")

print(power_result)
```

```
     Two-sample t test power calculation

              n = 5502
          delta = 149.0503
             sd = 2806.517
      sig.level = 0.05
          power = 0.8729655
    alternative = one.sided

NOTE: n is number in *each* group
```

Interpretación del Resultado (`power = 0.8729`)

Para validar la robustez del experimento, he calculado la potencia estadística *a posteriori*.

1. **Potencia del 87,3%:** Al obtener un 87,3% significa que el experimento estaba muy bien diseñado y tenía un tamaño de muestra más que suficiente para detectar ese aumento de ventas de 149€.
2. **Riesgo de Error Tipo II (Beta):** Recordemos que $\text{Beta} = 1 - \text{Potencia}$.
   - $1 - 0.873 = 0.127$

– Esto significa que solo teniamos un 12,7% de riesgo de haber pasado por alto la mejora si esta existía. Como el test ha salido positivo, este riesgo ya no aplica, pero confirma que el test era fiable desde el principio.