Tarea 04 Prueba de Hipotesis

Rodrigo Alan Garcia Perez

2024-09-09

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(nullabor) # paquete de pruebas visuales
library(tidyverse)
propinas <- read_csv("propinas.csv")
```

```
## Rows: 244 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): fumador, dia, momento
## dbl (3): cuenta_total, propina, num_personas
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(propinas)
```

```
## Rows: 244
## Columns: 6
## $ cuenta_total <dbl> 16.99, 10.34, 21.01, 23.68, 24.59, 25.29, 8.77, 26.88, 15~
## $ propina      <dbl> 1.01, 1.66, 3.50, 3.31, 3.61, 4.71, 2.00, 3.12, 1.96, 3.2~
## $ fumador      <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No~
## $ dia          <chr> "Dom", "Dom", "Dom", "Dom", "Dom", "Dom", "Dom", "Dom", "~
## $ momento      <chr> "Cena", "Cena", "Cena", "Cena", "Cena", "Cena", "Cena", "~
## $ num_personas <dbl> 2, 3, 3, 2, 4, 4, 2, 4, 2, 2, 2, 4, 2, 4, 2, 2, 3, 3, 3, ~
```

Prueba de sospechosos (prueba de hipótesis visual) ¿Las propinas son diferentes entre cena y comida (momento)

1. Crea permutaciones

¿Cómo se ve la tabla perms_momento?

```
set.seed(2178827)
perms_momento <- lineup(null_permute("momento"), propinas, n = 16)
```

```
## decrypt("ZYpj CyAy TJ HX6TATXJ fi")
```

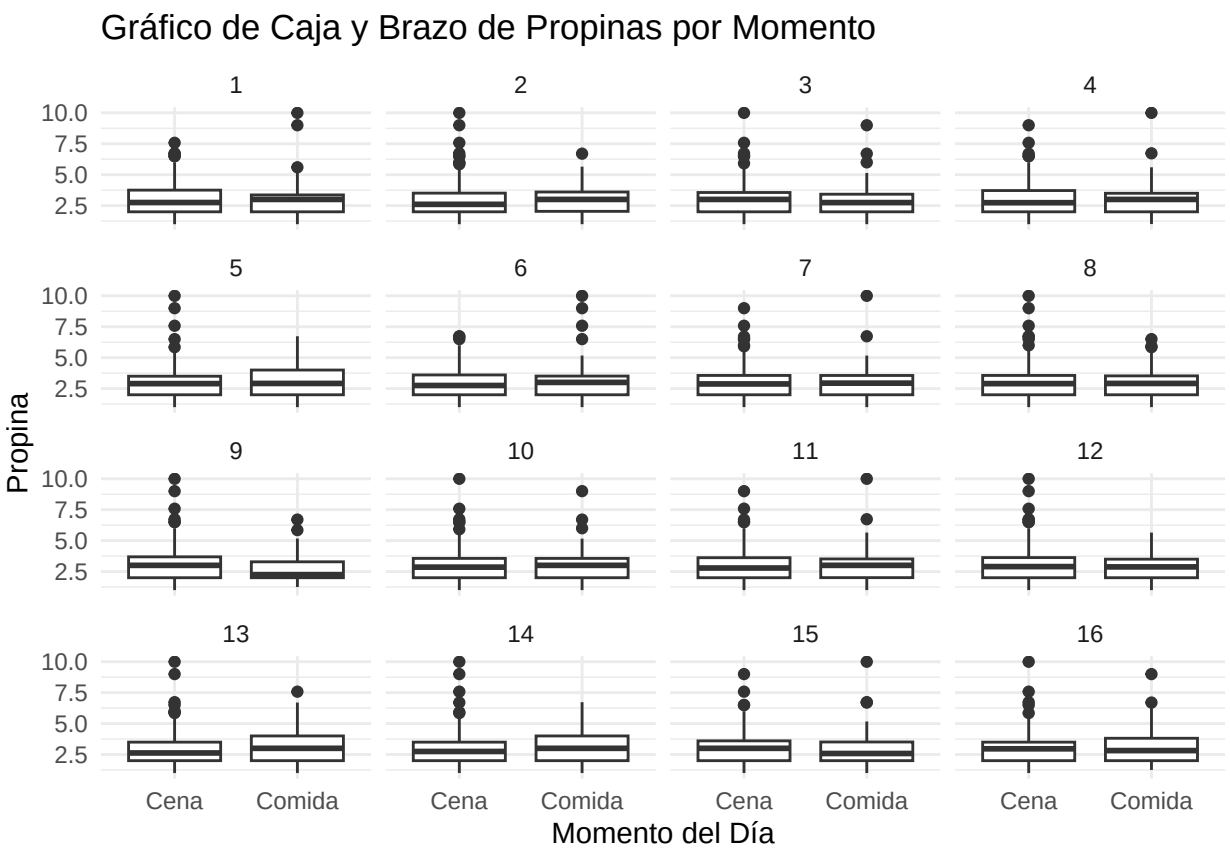
```
glimpse(perms_momento)
```

```
## Rows: 3,904
## Columns: 7
## $ cuenta_total <dbl> 16.99, 10.34, 21.01, 23.68, 24.59, 25.29, 8.77, 26.88, 15~
## $ propina      <dbl> 1.01, 1.66, 3.50, 3.31, 3.61, 4.71, 2.00, 3.12, 1.96, 3.2~
## $ fumador      <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No~
## $ dia          <chr> "Dom", "Dom", "Dom", "Dom", "Dom", "Dom", "Dom", "Dom", "~
## $ momento      <chr> "Comida", "Cena", "Cena", "Cena", "Cena", "Cena", "Cena",~
## $ num_personas <dbl> 2, 3, 3, 2, 4, 4, 2, 4, 2, 2, 2, 4, 2, 4, 2, 2, 3, 3, 3, ~
## $ .sample      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
```

La tabla perms_momento se creó a partir de la tabla original propinas, pero se hicieron permutaciones aleatorias de la columna “momento”.perms_momento contiene múltiples versiones (16 en total) de los datos originales, donde cada versión corresponde a una permutación diferente de los valores en “momento”.

Haz una grafica de caja y brazo de momento contra propinas separando en páneles las permutaciones

```
ggplot(perms_momento, aes(x=momento, y=propina)) +
  geom_boxplot() +
  facet_wrap(~ .sample) +
  labs(
    title = "Gráfico de Caja y Brazo de Propinas por Momento",
    x = "Momento del Día",
    y = "Propina"
  ) +
  theme_minimal()
```



Puedes identificar los datos verdaderos? Usa el comando decrypt que salió arriba para averiguar dónde están los datos.

Sin revisar la respuesta correcta, diría que los datos correctos se encuentran en el panel número 12

```
decrypt("ZYpj CyAy TJ HX6TATXJ fi")
```

```
## [1] "True data in position  9"
```

Qué tanta evidencia crees que aporta este análisis en contra de la hipótesis que las propinas son similares en niveles en comida y cena)

Si hubieramos escogido la respuesta correcta, esto hubiera tenido una significancia de $1/16 = 0.0625$ lo cual habría sido bastante evidencia en contra de la hipotesis nula.

2. Relaciones lineales

¿Está relacionado el tamaño de la cuenta con el tamaño de la propina?

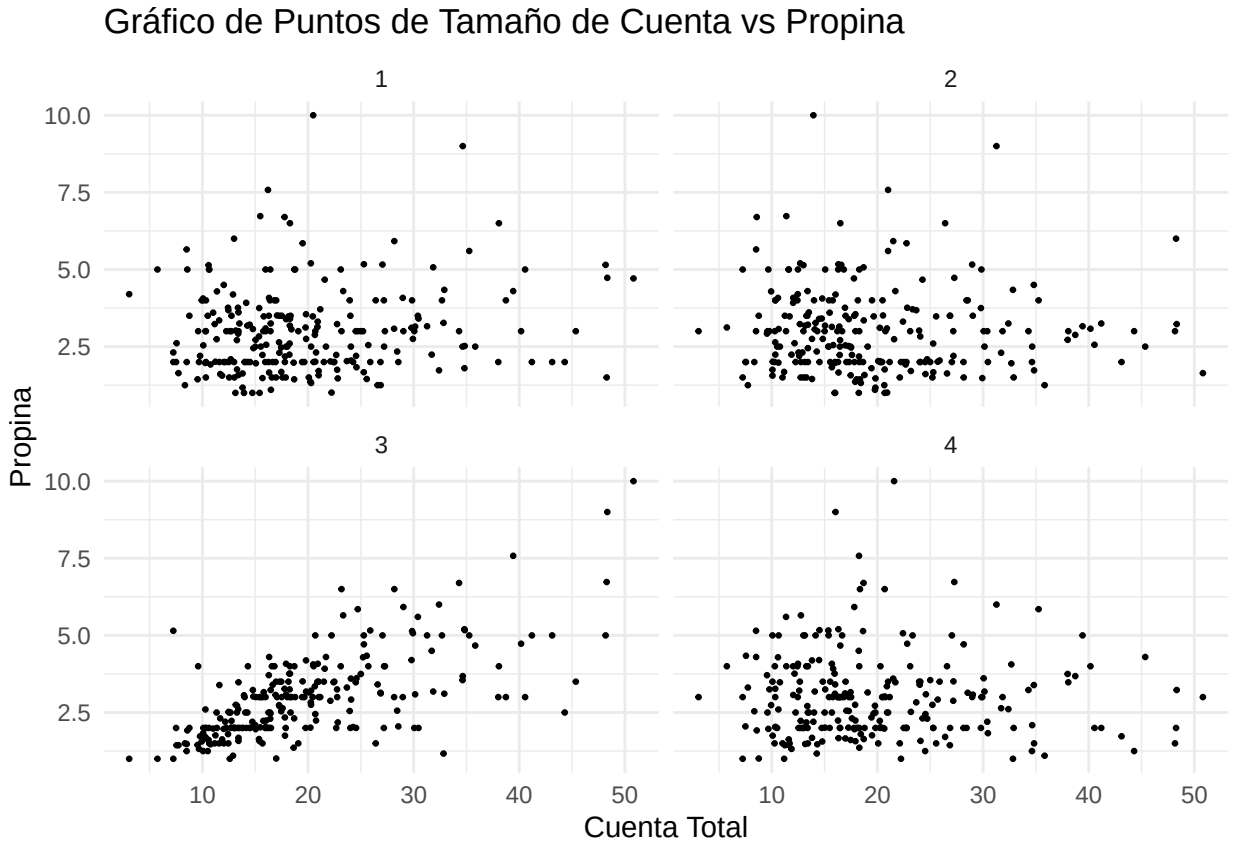
```
perms_propina <- lineup(null_permute("propina"), propinas, n = 4)
```

```
## decrypt("ZYpj CyAy TJ HX6TATXJ fI")
```

Haz una grafica de puntos de tamaño de cuenta vs propina separando en páneles las permutaciones

```
ggplot(perms_propina, aes(x=cuenta_total, y=propina)) +
  geom_point(size = 0.5) + facet_wrap(~ .sample) +
  labs(
    title = "Gráfico de Puntos de Tamaño de Cuenta vs Propina",
    x = "Cuenta Total",
```

```
    y = "Propina"
  ) +
  theme_minimal()
```



Cuánta evidencia tienes en contra de que no están relacionados?

A mi observación, yo diria que el panel 3 es donde se encuentran los datos verdaderos. De ser cierto tendría una significancia de 0.25 en contra de la hipotesis nula que señala que el total de la cuenta y la propina no esta relacionada.

```
decrypt("ZYpj CyAy TJ HX6TATXJ fI")
```

```
## [1] "True data in position  3"
```

3. Haz una prueba de diferencia de medias para comparar la propina en cena vs en comidas

```
# cinco mil permutaciones
momento_propina_tbl <- propinas %>% select(momento, propina)
perms_propina <- lineup(null_permute("propina"),
                        momento_propina_tbl, n = 5000)
```

```
## decrypt("ZYpj CyAy TJ HX6TATXJ fUfi")
```

```
# resumimos y calculamos diferencia
# revisa lo que obtienes en cada paso del siguiente código
resumen <- perms_propina %>%
```

```
  group_by(momento, .sample) %>%
  summarise(media = mean(propina)) %>%
  pivot_wider(names_from = momento, values_from = media) %>%
  mutate(dif_cena_comida = Cena - Comida)
```

```
## `summarise()` has grouped output by 'momento'. You can override using the
## `.groups` argument.
```

```
glimpse(resumen)
```

```
## Rows: 5,000
## Columns: 4
## $ .sample         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ Cena            <dbl> 3.036023, 2.980682, 3.139148, 2.950000, 2.992386, 2.96~
## $ Comida          <dbl> 2.900588, 3.043824, 2.633676, 3.123235, 3.013529, 3.08~
## $ dif_cena_comida <dbl> 0.135434492, -0.063141711, 0.505471257, -0.173235294, ~
```

Ahora grafica distribución (histograma, geom_histogram) de simulaciones contra el valor en los datos

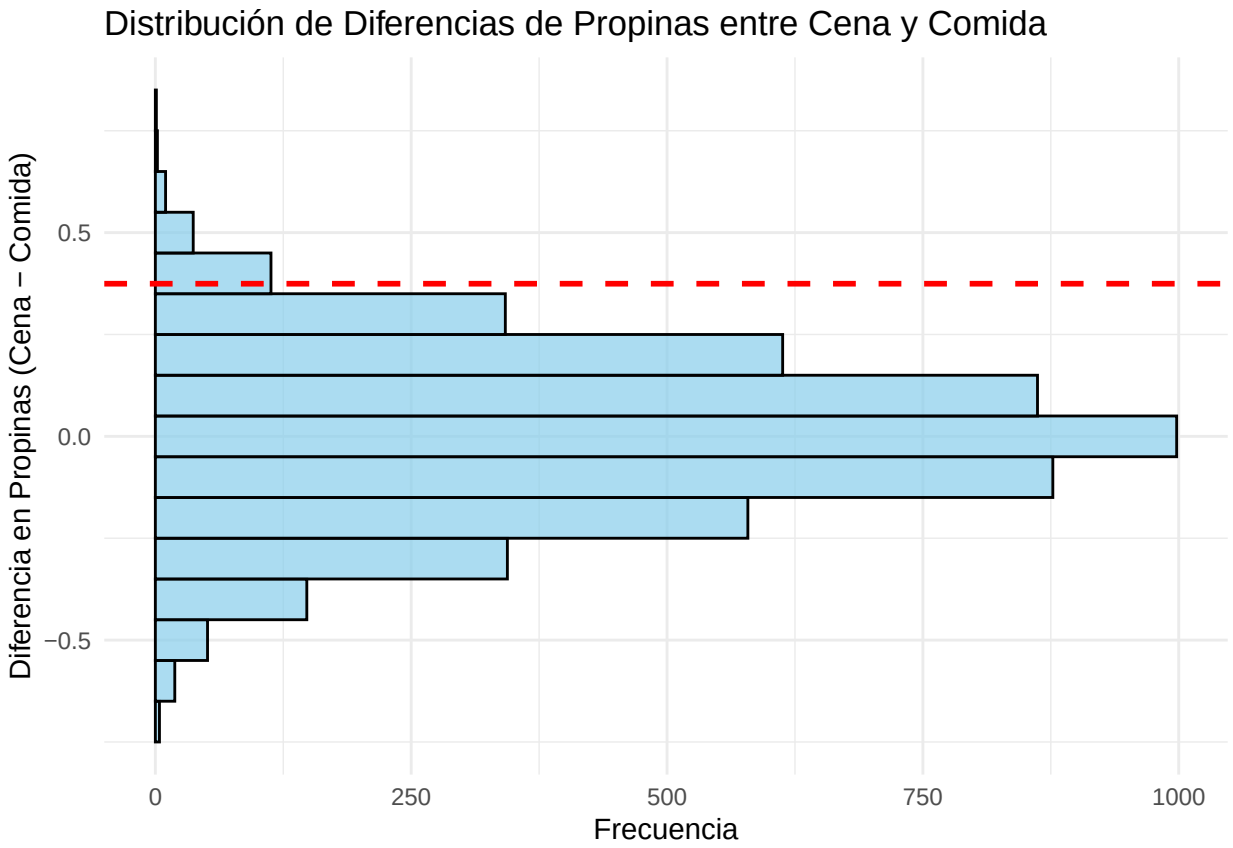
```
dif_obs <- momento_propina_tbl %>%
  group_by(momento) %>%
  summarise(media = mean(propina)) %>%
  pivot_wider(names_from = momento, values_from = media) %>%
  mutate(dif_cena_comida = Cena - Comida) %>%
  pull(dif_cena_comida)
dif_obs
```

```
## [1] 0.3745822
```

```
ggplot(resumen, aes(x=dif_cena_comida)) +
         geom_histogram(binwidth = 0.1, fill = "skyblue", color = "black", alpha = 0.7) +
         geom_vline(xintercept = dif_obs,
             color = "red", linetype = "dashed", size = 1) +

        labs(
          title = "Distribución de Diferencias de Propinas entre Cena y Comida",
          x = "Diferencia en Propinas (Cena - Comida)",
          y = "Frecuencia"
        ) +
        coord_flip() +
        theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



La diferencia observada no es muy extrema con respecto a la distribución nula.

Y el valor p (dos colas)

```
dist_ref <- ecdf(resumen$dif_cena_comida)
valor_p <- 2 * min(dist_ref(dif_obs), (1 - dist_ref(dif_obs)))
valor_p
```

```
## [1] 0.0508
```

¿cuál es tu conclusión?

Un valor de p de 0.0508 sugiere que, bajo la hipótesis nula de que no hay diferencia significativa en las propinas entre “Cena” y “Comida”, la probabilidad de observar una diferencia tan extrema como la que se ha observado (o más extrema) es del 5.08%. No hay suficiente evidencia para rechazar la hipótesis nula al nivel de significancia de 0.05. Sin embargo, está muy cerca del umbral, lo que puede justificar una mayor investigación o la obtención de más datos.