```
library(tidyverse)

library(patchwork)

library(readr)

## set working directory

setwd("/Users/thomas/OneDrive - INSTITUTO TECNOLOGICO AUTONOMO DE MEXICO/2016_ITAM/2023_MaestriaCienciasDeDatos/2023-3_FundamentoEstadisticos/tarea_01")

## Lee los datos

tips <- read_csv("tips.csv")

glimpse(tips)


## Recodificar nombres y niveles

propinas <- tips %>%

  rename(cuenta_total = total_bill,

         propina = tip, sexo = sex,

         fumador = smoker,

         dia = day, momento = time,

         num_personas = size) %>%

  mutate(sexo = recode(sexo, Female = "Mujer", Male = "Hombre"),

         fumador = recode(fumador, No = "No", Si = "Si"),

         dia = recode(dia, Sun = "Dom", Sat = "Sab", Thur = "Jue", Fri = "Vie"),

         momento = recode(momento, Dinner = "Cena", Lunch = "Comida")) %>%

  select(-sexo) %>%

  mutate(dia  = fct_relevel(dia, c("Jue", "Vie", "Sab", "Dom")))

propinas



## 1. Calcula percentiles de la variable propina

## junto con mínimo y máxmo
```

```
quantile(propinas$propina, probs = seq(0, 1, 0.05))


## 2. Haz una gráfica de cuantiles de la variable propina
propinas <- propinas %>%
  mutate(orden_propina = rank(cuenta_total, ties.method = "first"),
         f = orden_propina / n() )
## aquí tu código
# ggplot

ggplot(propinas, aes(x = cuenta_total)) +
  geom_histogram()

## 3. Haz un histograma de la variable propinas
## Ajusta distintos anchos de banda
# ggplot con binwidth=[0.1 0.5 5]
ggplot(propinas, aes(x = cuenta_total)) +
  geom_histogram(binwidth=5)
```

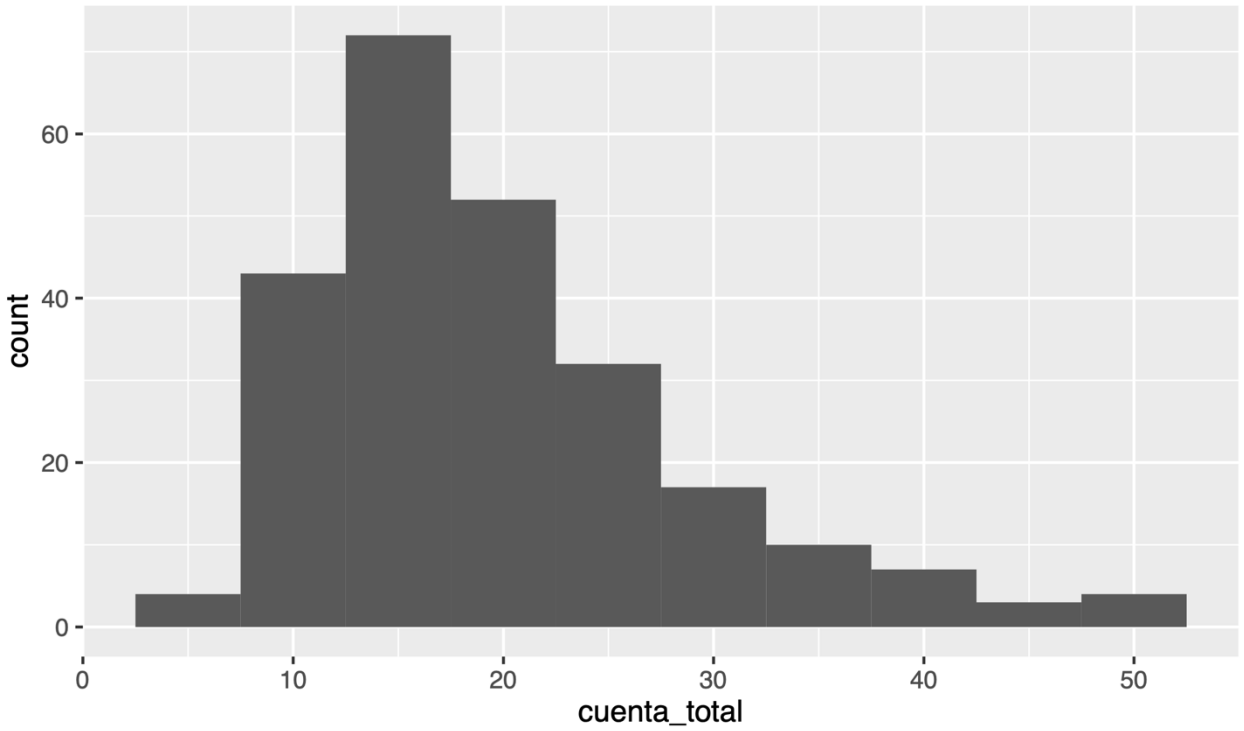


*Ilustración 1: histogram with bandwidth 5*

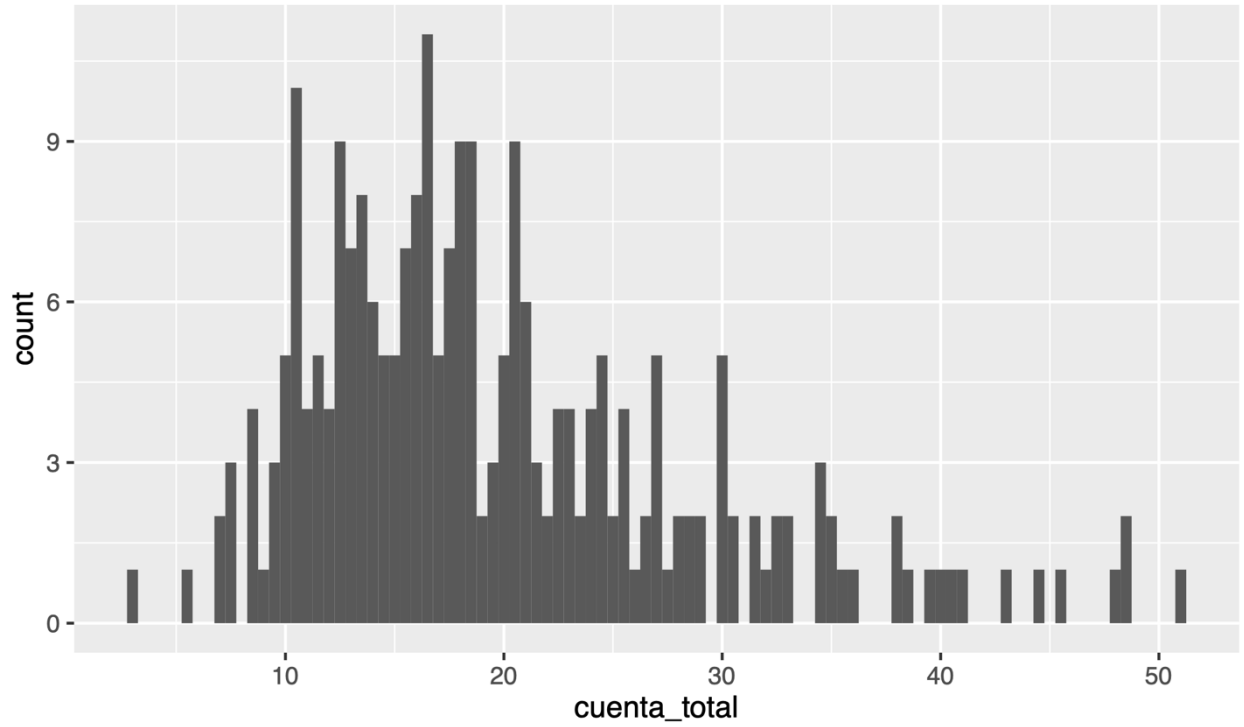


*Ilustración 2: Histogram with bandwidth 1*

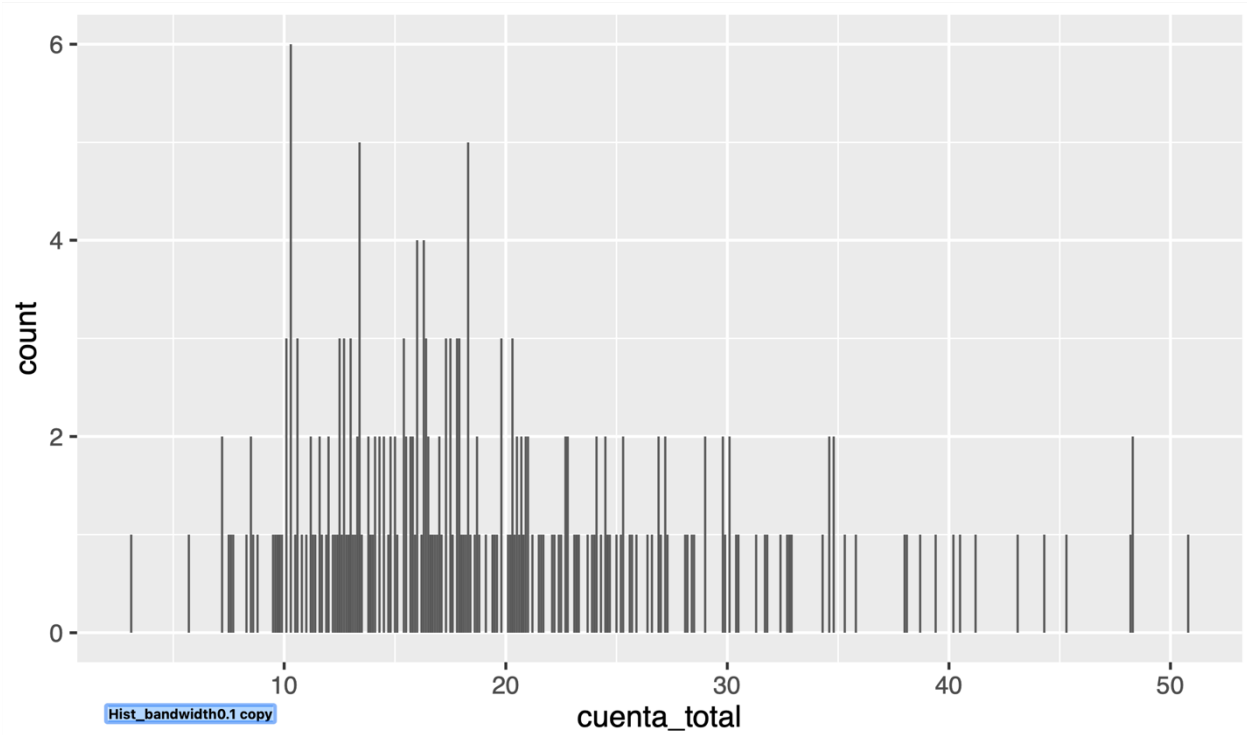


*Ilustración 3: Histogram with bandwidth 0.1*

## 4. Haz una gráfica de cuenta total contra propina

## ggplot

```
ggplot(propinas) +
  geom_point(aes(x = cuenta_total, y = propina, color=num_personas))
```

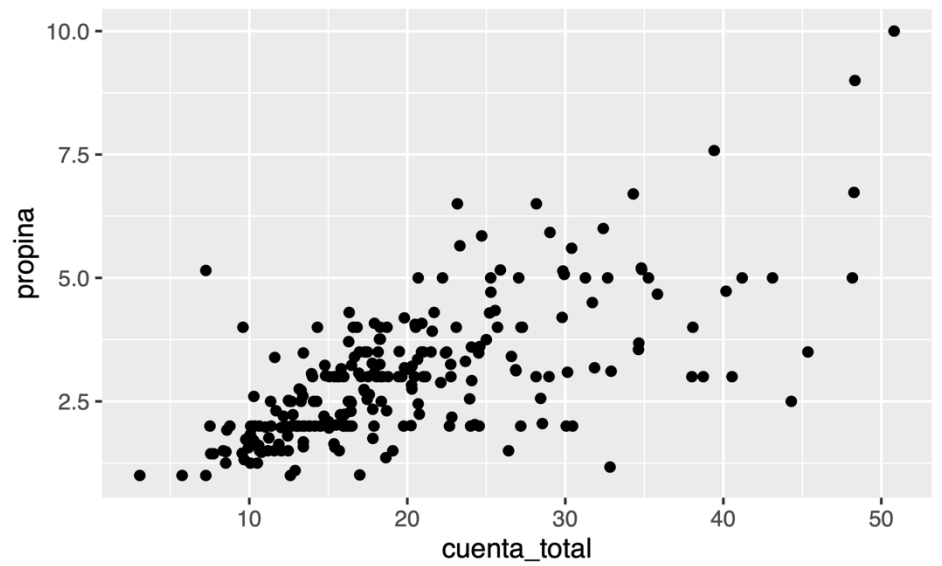


*Ilustración 4: cuenta vs. propina*

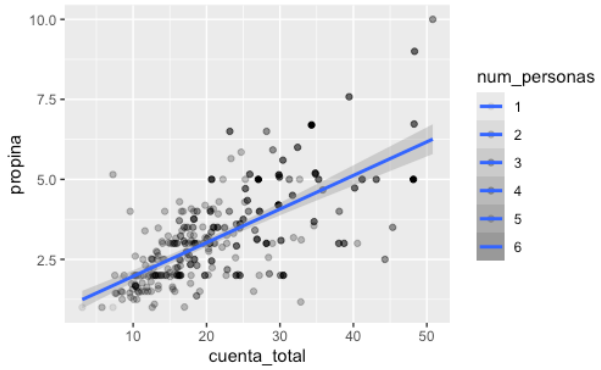


*Ilustración 5: cuenta vs propina usando alpha*

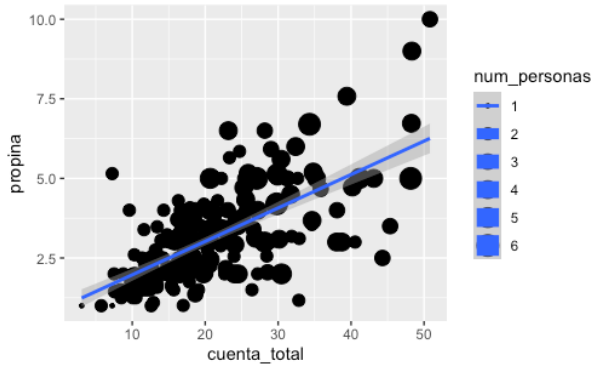


*Ilustración 6: cuenta vs propina usando size*

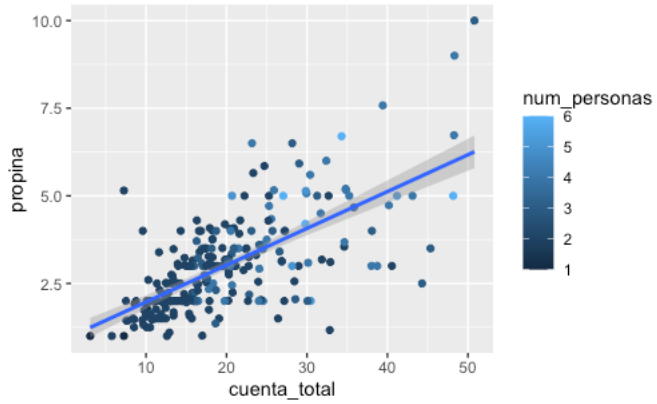


*Ilustración 7: cuenta vs propina usando coloue*

```
ggplot(propinas,aes(x = cuenta_total, y = propina, color=num_personas)) +
  geom_point() +
  geom_smooth(method = "glm")
```

## 5. Calcula propina en porcentaje de la cuenta total

```
propinas <- propinas %>%
```

```
  mutate(pct_propina = propina/cuenta_total)
## calcula algunos cuantiles de propina en porcentaje
quantile(propinas$pct_propina, probs = seq(0, 1, 0.05))



## 6. Haz un histograma de la propina en porcentaje. Prueba con
##  distintos anchos de banda.
## binwidth = [default, 0.1, 0.01]
ggplot(propinas) +
  geom_histogram(aes(x=pct_propina), binwidth = 0.01) +
  facet_wrap(~momento)
```

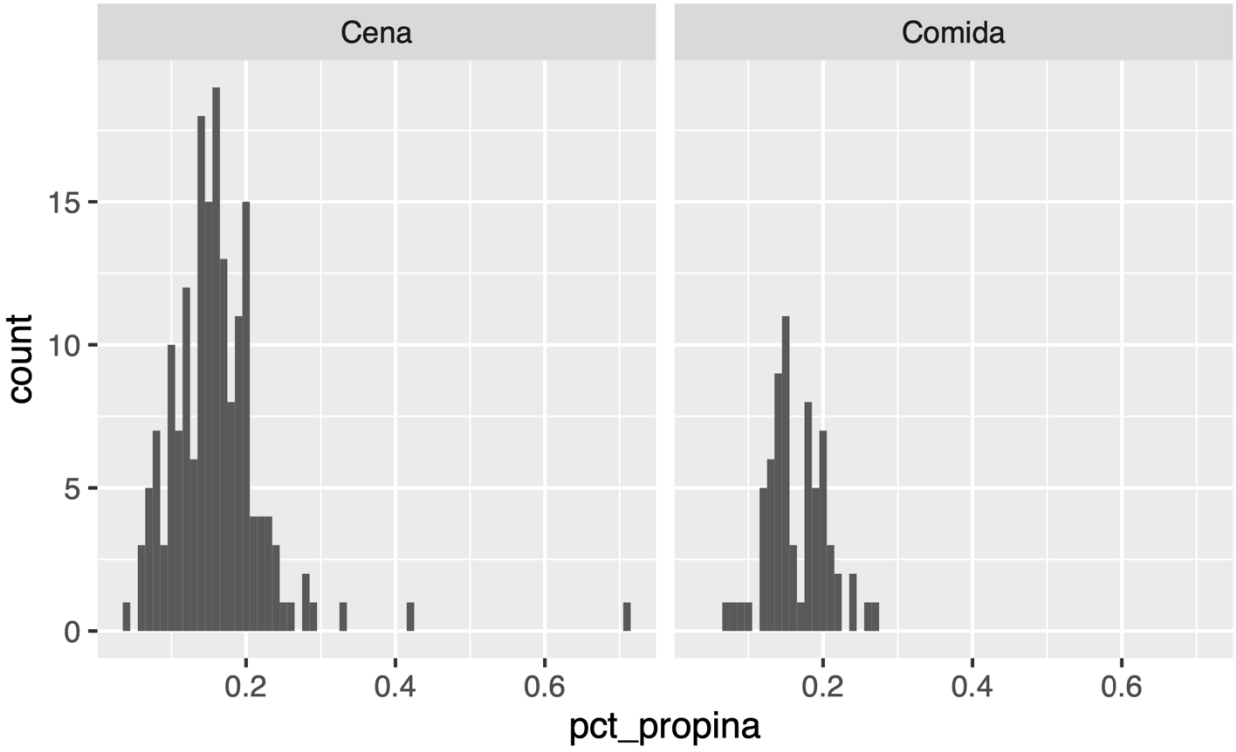


*Ilustración 8: Histogram with bandwidth 0.01*

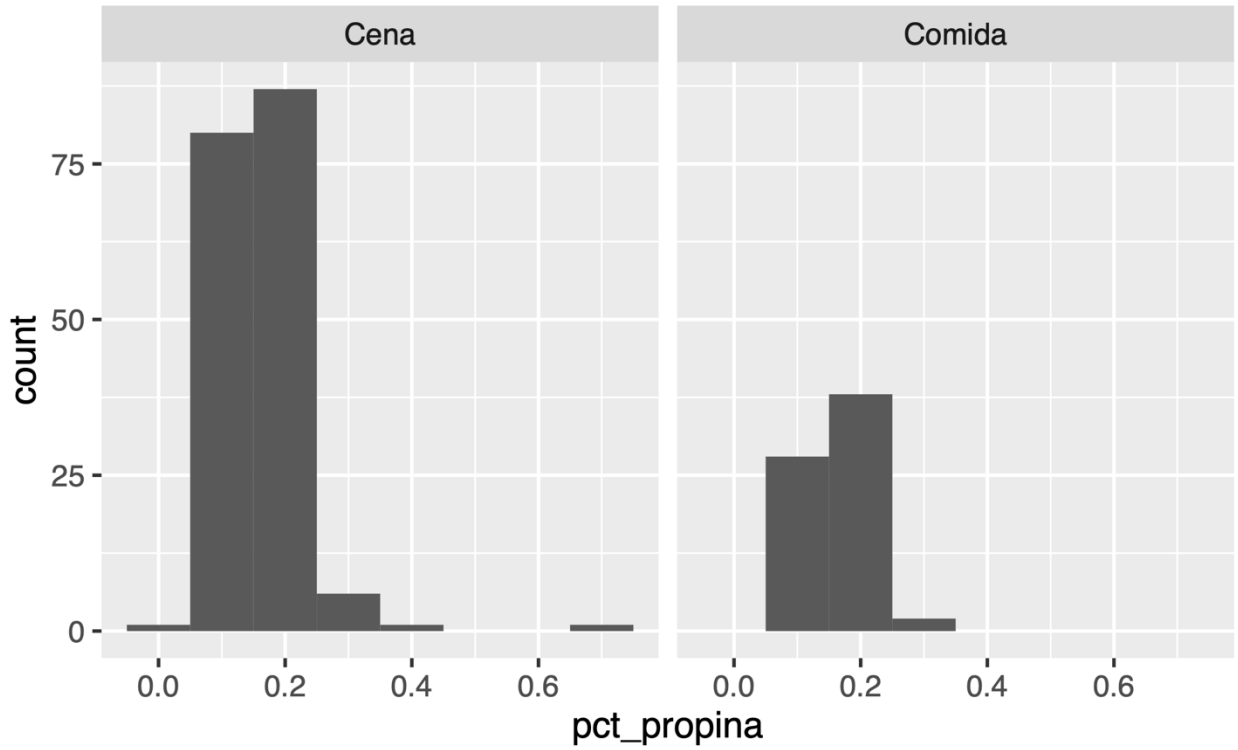


*Ilustración 9: Histogram with bandwidth 0.1*

## 7. Describe la distribución de propina en pct. ¿Hay datos atípicos?

max(propinas$pct_propina)

mean(propinas$pct_propina)

median(propinas$pct_propina)

sd(propinas$pct_propina)

## La distribucion de los porcentages de propina parece ser una distribución normal.

## con un valor promedio de 16% y und mediano de 15.5% y una devacion estandard de 6.1%

atypical <- propinas$pct_propina>0.3

propinas$pct_propina[atypical == TRUE]

## hay dos datos atípicos: mas que 30% (32.6%, 71% 41.7%)

## Lo interesante aqui es que tambien el valor de la cuenta total es relativamente bajo y

## el numero de personas: numero de personas: 1 2 2, y la cuenta total:$3.07 $7.25 $9.60.

## tambien la hora, las tres son en la tarde: puede ser que solamente tomaron algo.

```
propinas$num_personas[atypical == TRUE]

propinas$cuenta_total[atypical == TRUE]

propinas$momento[atypical == TRUE]


##8. Filtra los casos con porcentaje de propina muy altos.

## ¿Qué tipos de cuentas son? ¿Son cuentas grandes o chicas? --> arriba en 7.)

filter(propinas, pct_propina>0.3)

## cuenta_total propina fumador dia   momento num_personas orden_propina     f pct_propina

## <dbl>   <dbl> <chr>   <fct> <chr>         <dbl>        <int> <int>       <dbl>

##   1       3.07    1    Yes     Sab   Cena             1           1     1       0.326

##   2       7.25    5.15 Yes     Dom   Cena             2           4     4       0.710

##  3       9.6     4    Yes     Dom   Cena             2          14    14       0.417


## 9. Haz una diagrama de caja y brazos para

## propina en dolares dependiendo del momento (comida o cena)

## ¿Cuál parece más grande? ¿Por qué? Haz otras gráficas si es necesario.

ggplot(propinas) +

  geom_boxplot(aes(x=momento, y= pct_propina), outlier.colour = "red")


## parece mas grande el boxplot de la cena. Una razon puede ser que las propinas de la cena

## contiene los atípicos

## quitando las propinas mayor a 30% y vuelvo a usar ggplot con esos datos.

propinas_filtered <-filter(propinas, pct_propina<0.3)


ggplot(propinas_filtered) +

  geom_boxplot(aes(x=momento, y= pct_propina), outlier.colour = "red")
```

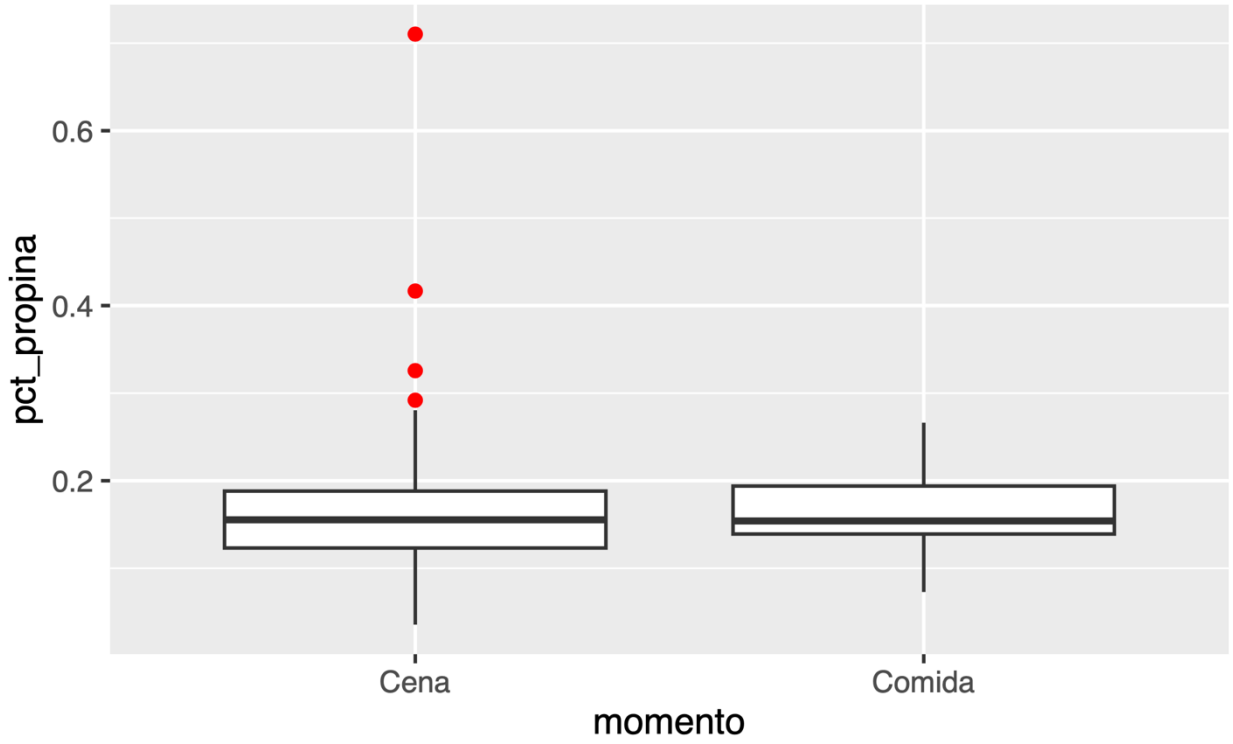


*Ilustración 10: Boxplot propina, incluyendo atipicos*

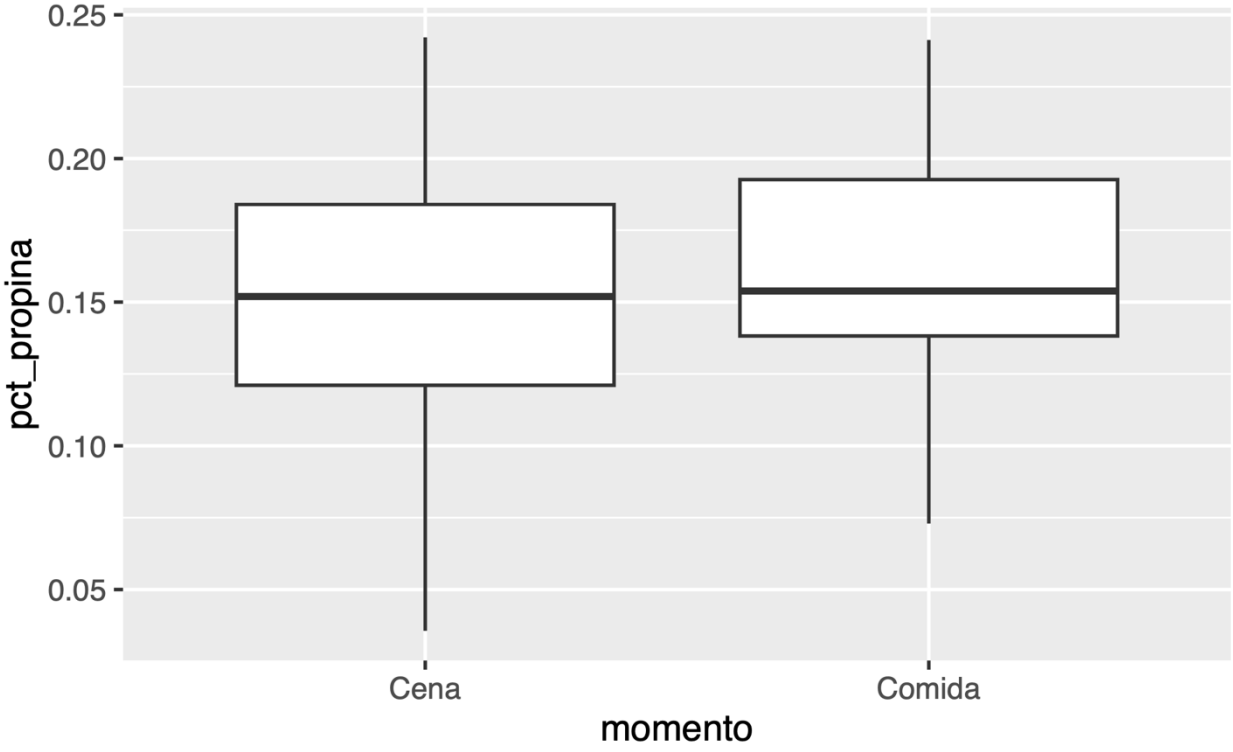


*Ilustración 11:Boxplot propina, sin atipicos*