# Tarea 1

### Taller de Econometría Aplicada I

Esteban Degetau

2023-09-12

## 1. Análisis descriptivo

a. Usando el paquete `DataExplorer`, obtuvimos las estadísticas mostradas en la Tabla 1, dodne se puede observar que los datos de bebidas de Starbucks tienen 15 variables para 1147 observaciones.

b. Estadística descriptiva en la Tabla 2 con media y desviación estándar entre paréntesis, así como mediana y rango intercuartílico entre corchetes.

c. Análisis gráfico. Histograma en la Figura 1, Boxplot en la Figura 2, y el diagrama de dipersión en la Figura 3.

d. El efecto de agregar crema batida sobre las calorías se puede ver en la Figura 4. En el panel (a) se puede ver la diferencia en medianas universal, mientras que en el panel (b) se puede ver la diferencia en promedios por tipo tamaño de bebida, estimado por OLS.

e. Cinco conclusiones. (i) En Starbucks se pueden pedir 1,147 bebidas distintas, sin incluir las bebidas por temporada (Tabla 1). (ii) La bebida promedio de Starbucks tiene 228 calorías, 6.2 gramos de grasa y 35 gramos de azúcar (Tabla 2). (iii) La bebida *venti* mediana tiene más cafeína, más calorías, más azúcar y más grasa que las bebdias medianas *grande* y *tall* (Figura 2). (iv) Las bebidas con más calorías suelen tener más grasa, más colesterol y más azúcar, pero no necesariamente más cafeína (Figura 3). (v) Incluir crema batida en una bebida *grande* aumenta la cantidad de calorías en promedio de 200 a 400 KCal (Figura 4).

## 2. Estimación puntual

a. Estimación con `MASS` de la distribución de las variables azúcar y cafeína. En la Figura 5 con parametrización exponencial y en la Figura 7 con normal. Notar que ninguna parametrización le hace justicia a la verdadera distribución de los datos.

b. Tabla de resumen de estimadores. En la Tabla 3 se muestran los parámtros obtenidos de la distribución exponencial, y en la Tabla 4 de la distribución normal.

c. La Figura 6 confirma que la distribución exponencial no se ajusta muy bien a los datos de azúcar ni de cafeína. De hecho, los datos se ajustan mejor a una distribución normal, salvo por los valores en la cola izquierda, como se puede ver en la Figura 8. El la Figura 5 se puede ver el histograma comparado con la distribución exponencial, y en la Figura 7 con la distribución normal.

**Conclusión**: A pesar de que uno pensaría que una distribución exponencial se ajustaría mejor a los datos, por tener muchas observaciones en 0 (i.e. muchas bebidas sin cafeína o sin azúcar), la distribución normal se ajusta mejor a los datos al interior de la distribución (i.e. para valores mayores a 0) en cada caso.

## 3. Estimación por intervalo

a. Intervalo de 95% de confianza para la media de la variable de calorías (medidas en KCal). De acuerdo con la Tabla 5, 95 por ciento de las estimaciones de la media estarán dentro de un intervalo [220.43, 236.36].

b. Por tamaño de bebida, de acuerdo con la Tabla 6, las bebidas *venti* tienen en promedio más calorías que cualquier otro tamaño, con 95% de confianza. Las bebidas *solo* y las bebidas *triple* tienen en promedio, una cantidad comparable de calorías. Notar que solo hay una bebida con tamaño *1 shot*, por lo que no hay una varianza que nos arroje una desviación estándar válida y no podemos calcular el error estándar (se) ni el intervalo de confianza.

c. Representación gráfica de los intervalos de confianza por tamaño en la Figura 9.

d. Cocnlusión. Las bebidas con más calorías son las *venti*, que tienen en promedio 50 calorías más que las bebidas *grande* y 100 más que las *tall*. Hay unos tamaños con muy pocas calorías, como *triple*, *solo*, *quad* entre otras con menos de 50 calorías.

## 4. Pruebas de Hipótesis

a. A un nivel de significancia de 5%, escribimos para $i$ y $j$ tipos de leche distintos.

$$H_0 : \mu_i = \mu_j \quad vs \quad H_1 : \mu_i \neq \mu_j$$

En particular, podemos estar interesados en determinar si las bebidas con leche entera (*whole*) tienen en promedio una cantidad distinta de calorías, grasa y azúcar que las bebidas con leche *nonfat*. Adicionalmente, podríamos estar interesados en plantear la hipótesis alternativa de que las bebidas con leche ($i$) entera tienen en promedio más calorías, grasas y azúcar que las

bebidas con leche nonfat ($j$). Entonces escribiríamos $H_1 : \mu_i > \mu_j$. Sin embargo, esta hipótesis alternativa de un solo lado genera una región de rechazo más grande en el lado derecho de la distribución, para un mismo nivel de significancia. Entonces, en la prueba de un solo lado podríamos rechazar estimaciones que en la prueba de ambos lados no rechazaríamos. Para robustecer nuestras conclusiones, minimizando errores tipo 1, en esta sección nos limitaremos a pruebas de hipótesis bilaterales.

b. En la Tabla 7 se estima la prueba de hipótesis de diferencia de medias de calorías grasa y azúcar, entre leche entera y nonfat por OLS. Se presentan los estadísticos $t$ en paréntesis. Con un nivel de significancia de 5%, enocntramos que las bebidas con leche entera tienen más calorías y grasas que las bebidas con leche nonfat, pero no podemos rechazar la hipótesis nula de que tengan la misma cantidad de azúcares, puesto que enocntramos una $t = 1.045 < 1.96$.

c. En la Tabla 8 presentamos los valores-p de las pruebas de hipótesis. Consistente con los resultados de la Tabla 7, obtenemos valores p inferiores al 5% de nuestra significancia estadística para las medias de grasa y calorías, pero encontramos uno mayor a 5% para la prueba del azúcar.

d. Interpretación. Las bebidas con leche entera tienen en promedio más calorías que las bebidas con leche nonfat. Las bebidas con leche entera tienen en promedio más grasa que las bebidas con leche nonfat. No encontramos evidencia de que las bebidas con leche entera tengan en promedio más azúcar que las bebidas con leche nonfat.

e. Como vimos antes, el tamaño de la bebida y el uso de crema batida tienen un efecto en calorías. Rápidamente, en la Tabla 9 se puede ver la significancia de estas variables en la determinación de calorías.

## 5. Estimación por remuestreo

a. El coeficiente de correlación entre cafeína y calorías dentro de las bebidas de Starbucks es -0.075. En la Figura 10 se puede ver la relación negativa entre las variables.

b. En la Tabla 10 se puede ver la varianza del coeficiente de correlación entre cafeína y calorías, estimada con el 100% y el 50% de la muestra (con reemplazo) y 1,000 repeticiones en Bootstrap. Se puede ver que la varianza bajo las submuestras de 50% es mayor.

c. En la Figura 11 se puede ver la densidad de las correlaciones entre cafeína y calorías por tamaño de la submuestra por Bootstrap. Se puede apreciar que el ejercicio con 50% de la muestra tiene mayor varianza.

d. La varianza de la correlación entre cafeína y calorías utilizando Jackknife fue de $10^{-6}$, tres órdenes de magnitud inferior que ambas estimaciones por Bootstrap. En la Figura 12 se puede apreciar la distribución de las etimaciones de la correlación entre cafeína y

calorías por Jackknife. Notar que las densidad al rededor de la moda es mucho mayor que en Bootstrap.

e. Utilizando la varianza obtenida por Bootstrap con el totalidad de la muestra, sobre el coeficiente de correlacion entre cafeína y calorías $\rho$, evaluamos la hipótesis:

$$H_0 : \rho = 0 \quad vs \quad H_1 : \rho \neq 0$$

Obtenemos un estadístico $t$ = -70.3412903, que indica que podemos rechazar la hiótesis nula de que el coeficiente de correlación entre cafeína y calorías sea cero, incluso con un nivel de significancia del 1%, al evaluar respecto de la distribución empírica de los estimadores obtenida por Bootstrap con la totalidad de la muestra. La Tabla 11 muestra los valores de $|t|$ a partir de los cuales podemos rechazar $H_0$.

Tabla 1: Descripción de los datos, bebidas de Starbucks

| Estadístico | Valor |
|---|---|
| rows | 1147 |
| columns | 15 |
| discrete_columns | 5 |
| continuous_columns | 10 |
| all_missing_columns | 0 |
| total_missing_values | 0 |
| complete_rows | 1147 |
| total_observations | 17205 |
| memory_usage | 151064 |

## Tablas

Tabla 2: Bebidas de Starbucks. Estadística descriptiva

| **Característica** | **N = 1,147** |
|---|---|
| serv_size_m_l | |
| Media (DE) | 461 (172) |
| Mediana [RIQ] | 473 [354, 591] |
| calories | |
| Media (DE) | 228 (138) |
| Mediana [RIQ] | 220 [130, 320] |
| total_fat_g | |
| Media (DE) | 6.2 (6.0) |
| Mediana [RIQ] | 4.5 [1.0, 10.0] |
| cholesterol_mg | |
| Media (DE) | 15 (18) |
| Mediana [RIQ] | 5 [0, 30] |
| sugar_g | |
| Media (DE) | 35 (22) |
| Mediana [RIQ] | 34 [18, 49] |
| caffeine_mg | |
| Media (DE) | 92 (78) |
| Mediana [RIQ] | 75 [30, 150] |

Tabla 3: Estimación de parámetros con distribución exponencial

| Parámetro | Azúcar | Cafeína |
|---|---|---|
| Tasa | 0.029 | 0.011 |

Tabla 4: Estimación de parámtros con distribución normal

| Parámetro | Azúcar | Cafeína |
|---|---|---|
| Media | 34.995 | 91.855 |
| Desviación Estándar | 22.449 | 78.075 |

Tabla 5: Intervalo de 95% de confianza para la media de calorías de las bebidas de Starbucks

| media | sd | n | se | lower | upper |
|---|---|---|---|---|---|
| 228.39 | 137.67 | 1147 | 4.06 | 220.43 | 236.36 |

Tabla 6: Intervalo de 95% de confianza para la media de las calorías de bebidas de Starbucks por tamaño

| size | media | sd | n | se | lower | upper |
|---|---|---|---|---|---|---|
| 1 scoop | 27.50 | 3.54 | 2 | 2.50 | 22.60 | 32.40 |
| 1 shot | 5.00 | NA | 1 | NA | NA | NA |
| doppio | 16.43 | 8.52 | 7 | 3.22 | 10.12 | 22.74 |
| grande | 247.91 | 119.44 | 334 | 6.54 | 235.10 | 260.72 |
| quad | 27.86 | 9.94 | 7 | 3.76 | 20.49 | 35.22 |
| short | 116.44 | 73.88 | 123 | 6.66 | 103.38 | 129.50 |
| solo | 10.00 | 9.13 | 7 | 3.45 | 3.24 | 16.76 |
| tall | 182.29 | 89.70 | 318 | 5.03 | 172.43 | 192.15 |
| trenta | 182.62 | 54.95 | 21 | 11.99 | 159.12 | 206.12 |
| triple | 22.14 | 8.09 | 7 | 3.06 | 16.15 | 28.14 |
| venti | 320.14 | 150.76 | 320 | 8.43 | 303.62 | 336.66 |

Tabla 7: Diferencia de medias entre bebidas con leche entera y nonfat (categoría omitida), apara calorías, grasas y azúcares

| Dependent Variables: | calories | sugar_g | total_fat_g |
|---|---|---|---|
| Model: | (1) | (2) | (3) |
| *Variables* | | | |
| Constant | 225.7*** | 38.09*** | 3.518*** |
| | (26.88) | (26.97) | (9.625) |
| milkwhole | 66.28*** | 2.174 | 6.613*** |
| | (5.361) | (1.045) | (12.28) |
| *Fit statistics* | | | |
| Observations | 412 | 412 | 412 |
| Dependent variable mean | 256.29 | 39.097 | 6.5680 |
| F-test | 28.736 | 1.0925 | 150.91 |

*IID co-variance matrix, t-stats in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Tabla 8: Diferencia de medias entre bebidas con leche entera y nonfat, con valores p

| **Característica** | **nonfat**, N = 222 | **whole**, N = 190 | **Difference** | **p-valor** |
|---|---|---|---|---|
| calories | 226 (120) | 292 (131) | -66 | <0.001 |
| sugar_g | 38 (21) | 40 (21) | -2.2 | 0.3 |
| total_fat_g | 3.5 (4.7) | 10.1 (6.2) | -6.6 | <0.001 |

[1] Media (DE)

[2] t de Student

Tabla 9: Diferencia de medias de calorías para tamaño y uso de crema batida, solo para bebidas tamaños *tall*, *grande* y *venti*

| Dependent Variable: | calories | |
|---|---|---|
| Model: | (1) | (2) |
| *Variables* | | |
| Constant | 247.9*** | 201.9*** |
| | (6.705) | (4.021) |
| sizetall | -65.62*** | |
| | (9.601) | |
| sizeventi | 72.22*** | |
| | (9.586) | |
| whip | | 183.6*** |
| | | (7.835) |
| *Fit statistics* | | |
| Observations | 972 | 972 |
| Dependent variable mean | 250.22 | 250.22 |
| F-test | 101.00 | 549.18 |

*IID standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Tabla 10: Varianza de la correlación entre cafeína y calorías, estimada con diferentes tamaños de submuestra en Bootstrap

| subsample | var |
|---|---|
| 100% | 0.0011 |
| 50% | 0.0023 |

Tabla 11: Cuantiles de la distribución de coeficientes de correlación entre cafeína y calorías con Bootstrap y la totalidad de la muestra. Indican el nivel a partir del cual rechazar la hipótesis nula del inciso 5.e

| 90% | 95% | 99% |
|---|---|---|
| 70.30997 | 70.32166 | 70.33901 |

## Figuras

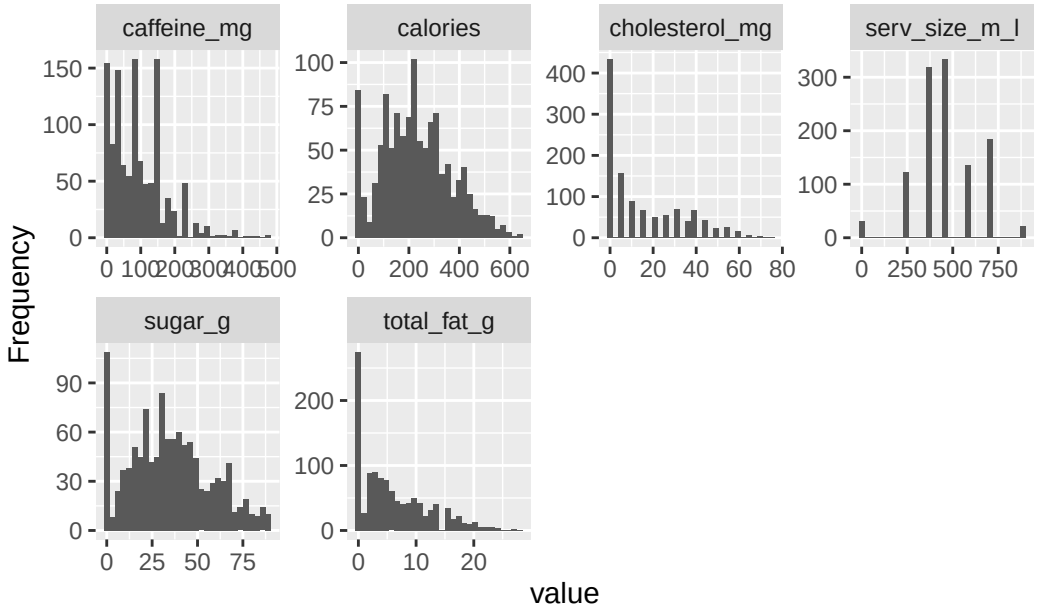



Figura 1: Histogramas

```
package 'gclus' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\LENOVO\AppData\Local\Temp\RtmpUHQsmg\downloaded_packages
```

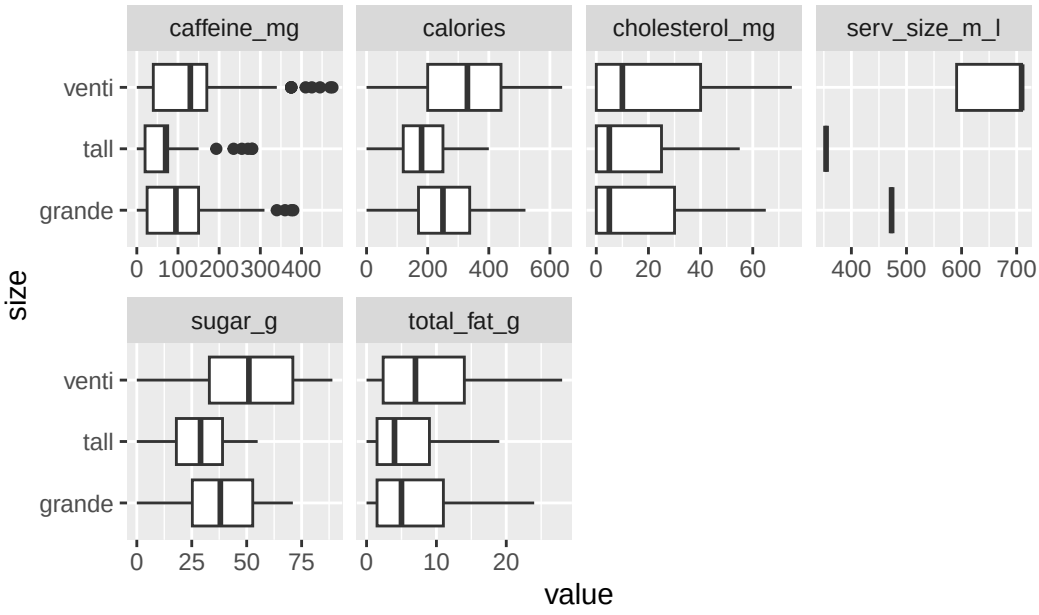



Figura 2: Boxplot por tamaño de bebida

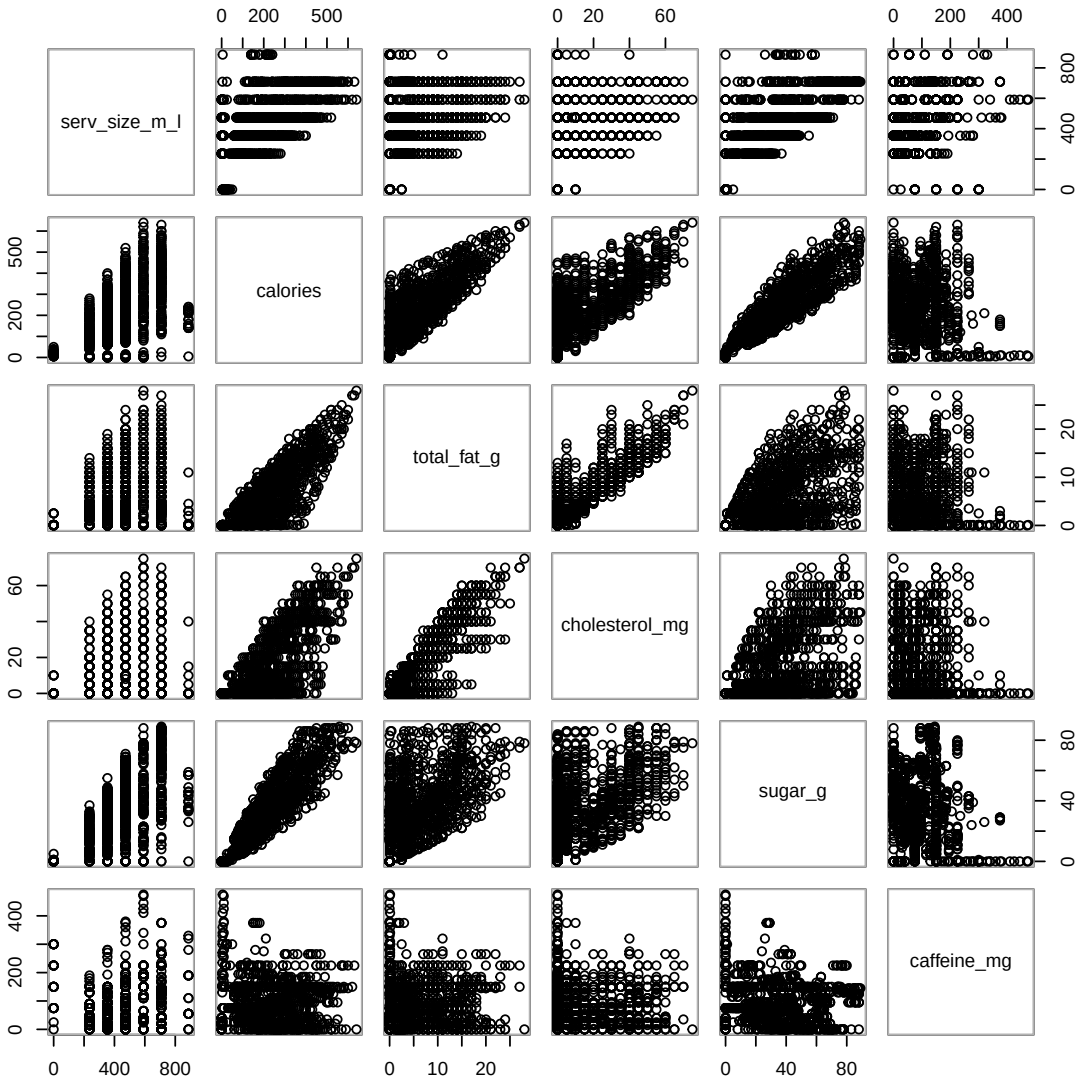
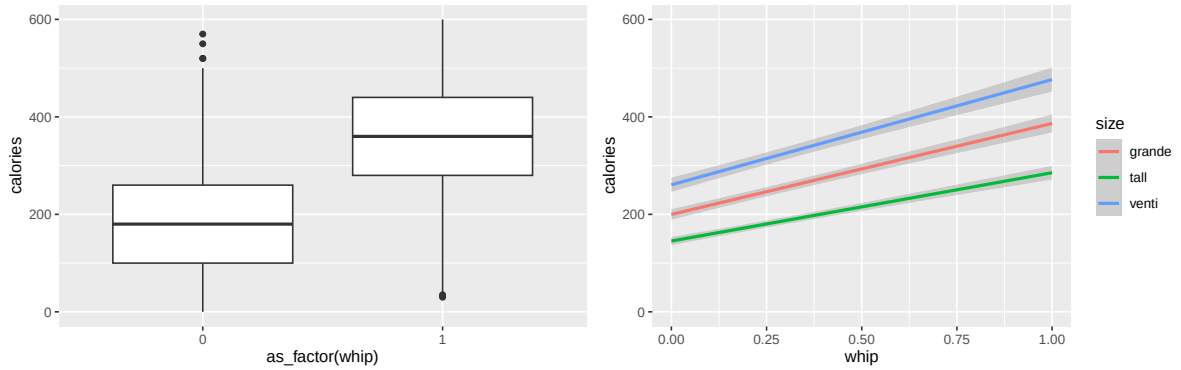
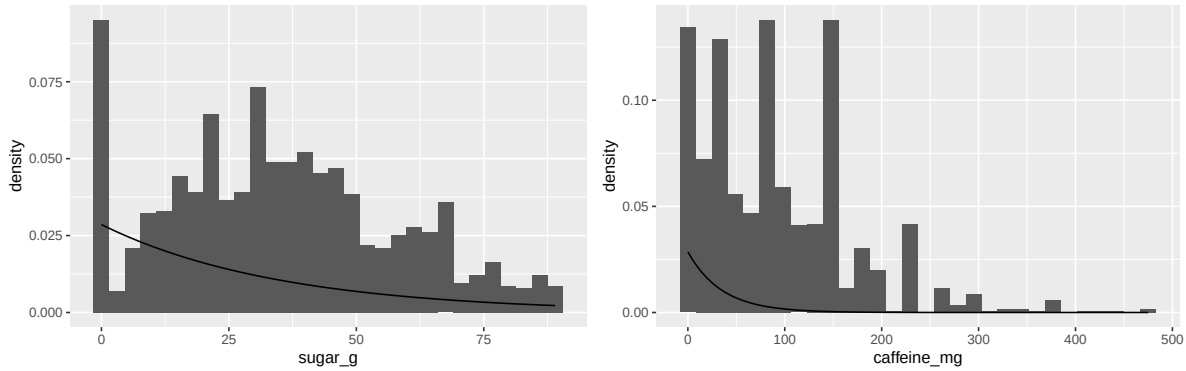



Figura 3: Diagrama de dispersión



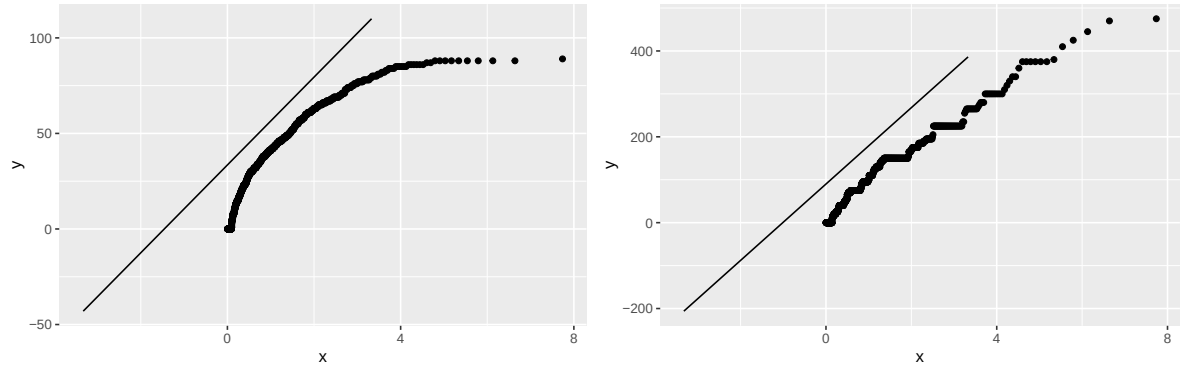

(a) Diferencia en medianas (b) Diferencia en medias

Figura 4: Crema batida y calorías



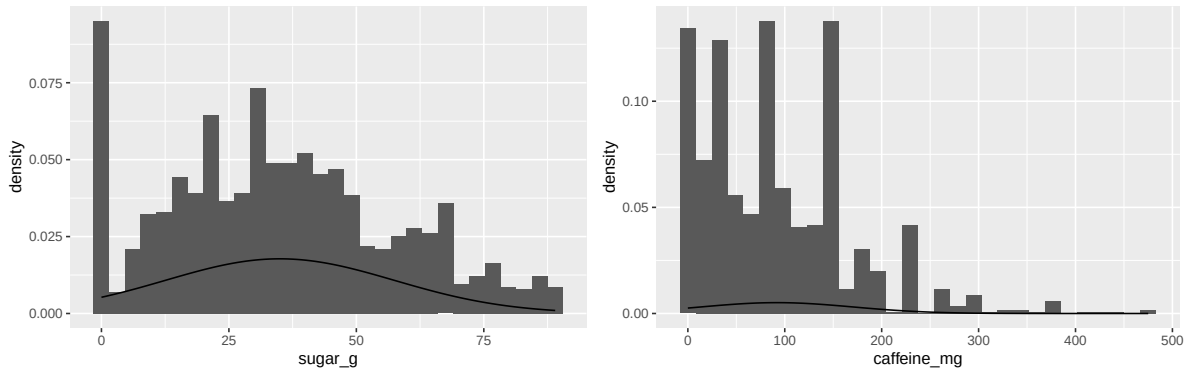

(a) Azúcar (b) Cafeína

Figura 5: Parametrización con distribución exponencial



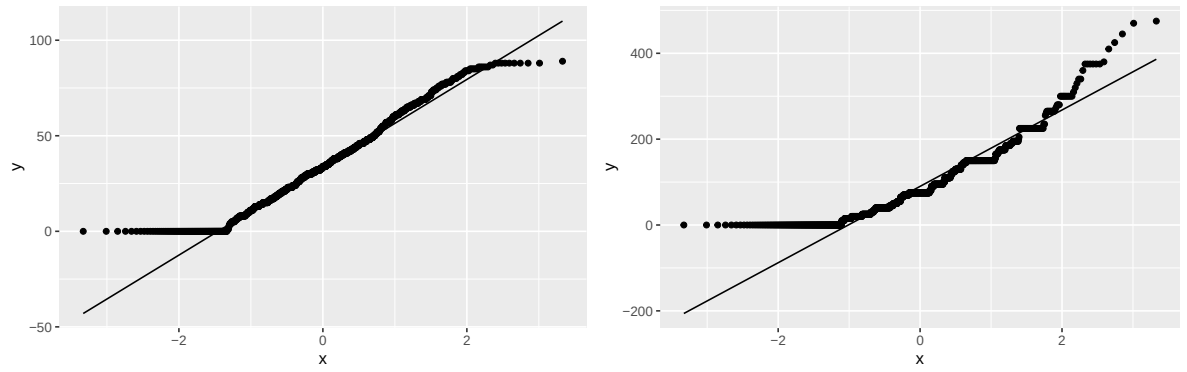

(a) Azúcar (b) Cafeína

Figura 6: Cuantil-cuantil exponencial




(a) Azúcar (b) Cafeína

Figura 7: Parametrización con distribución Normal




(a) Azúcar (b) Cafeína

Figura 8: Cuantil-cuantil normal

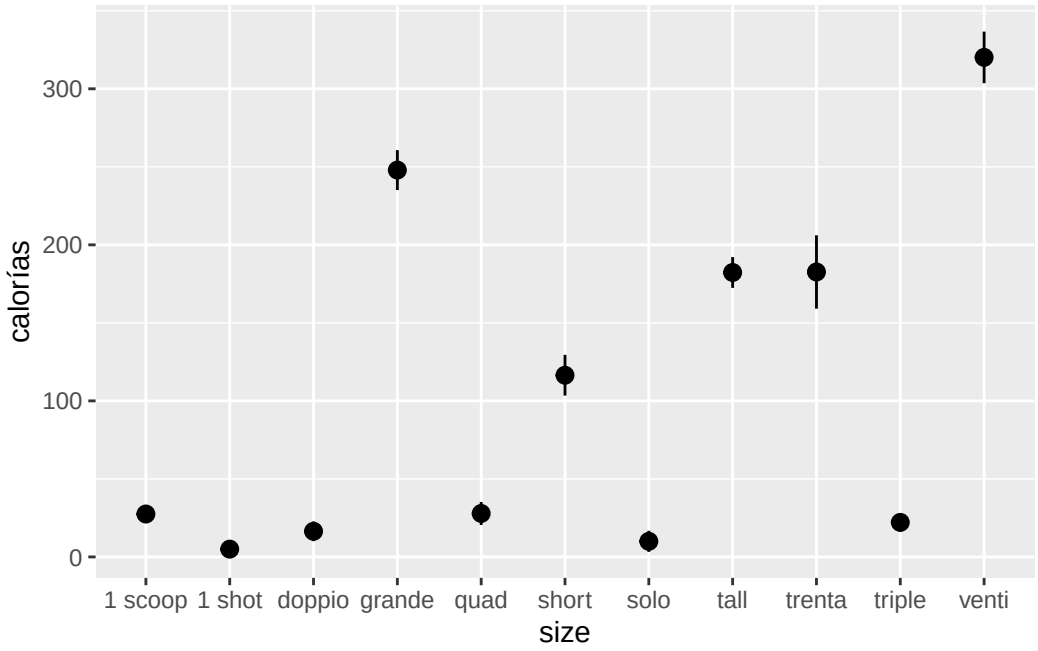



Figura 9: Calorías promedio por tamaño de bebida e intervalos de 95% de confianza

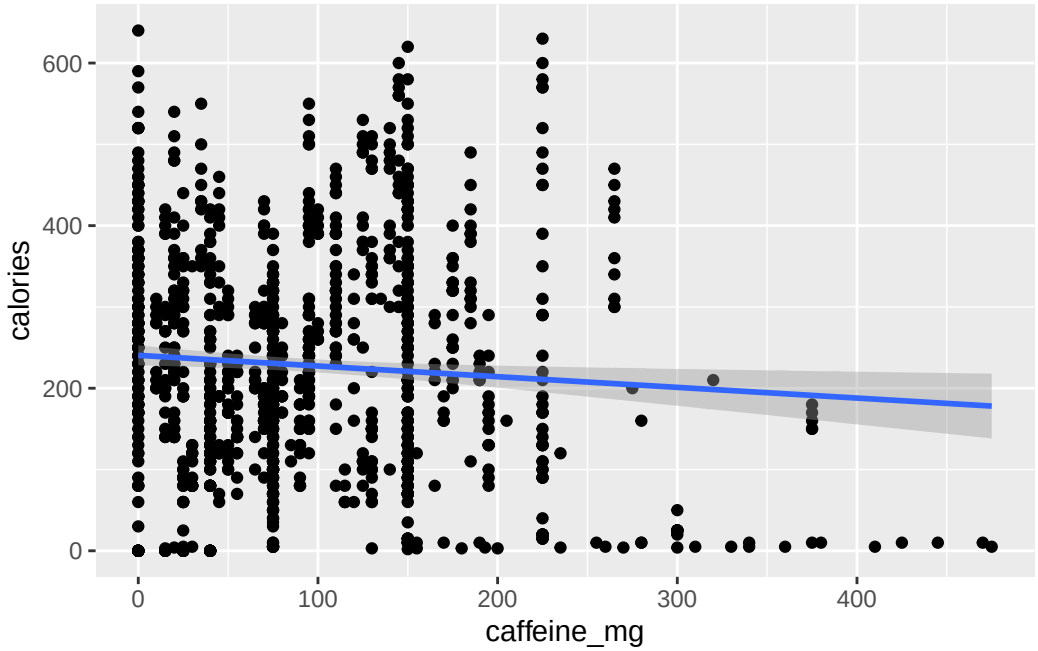



Figura 10: Dispersión entre cafeína y calorías, estimación lineal por OLS e intarvalo de confianza de 95%

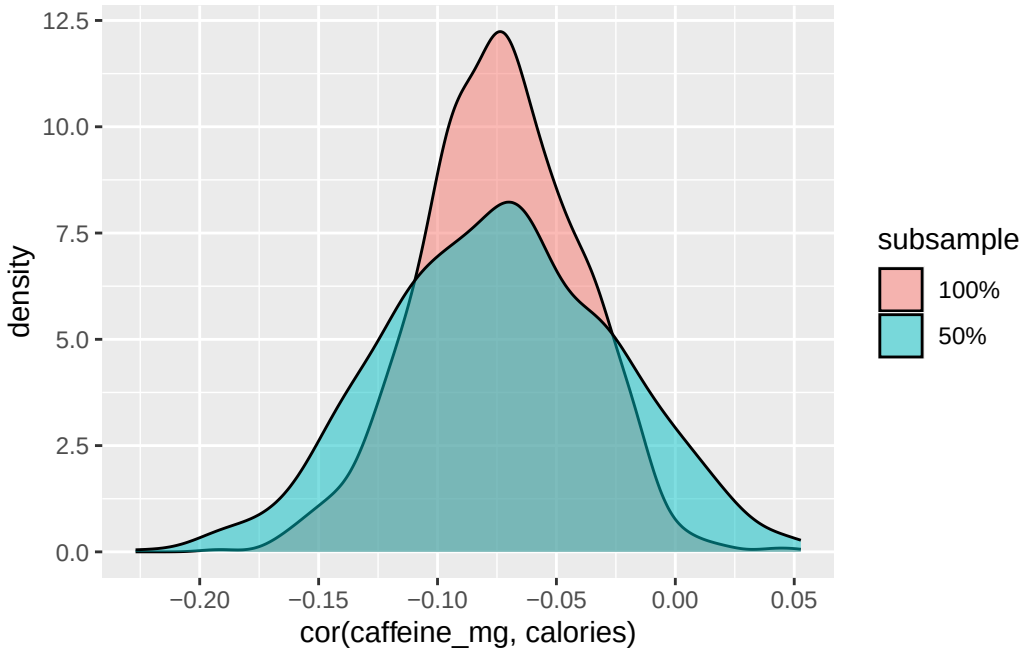



Figura 11: Densidad de las correlaciones entre cafeína y calorías por tamaño de la submuestra por Bootstrap

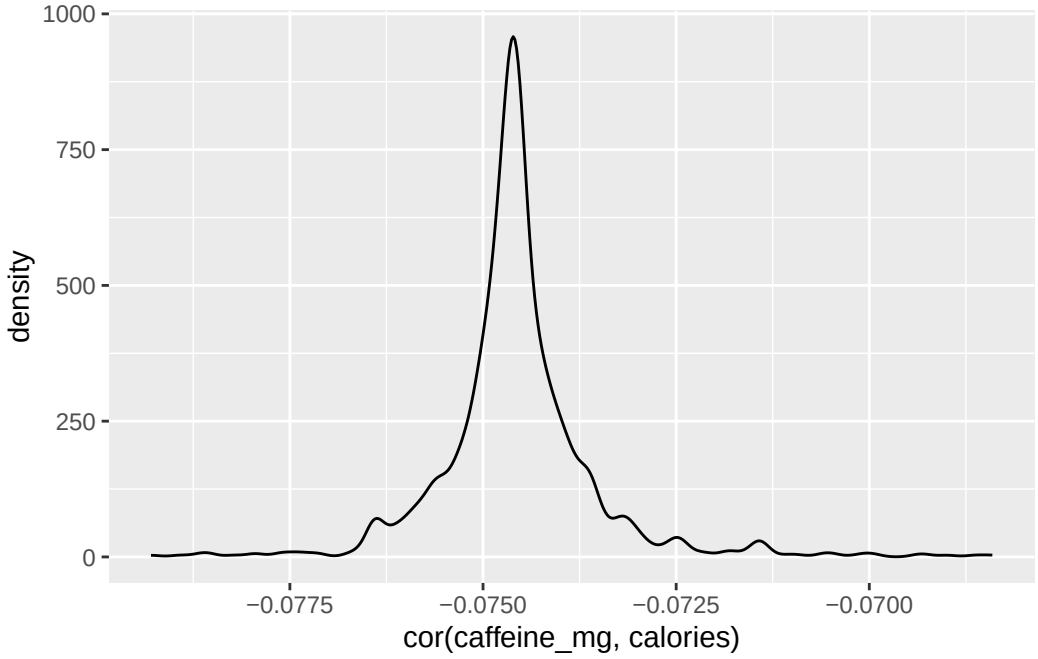



Figura 12: Densidad de estimaciones de la correlación entre cafeína y calorías por Jackknife