Modelos estadísticos simples y su variación

Recordemos

- ¿Qué es un modelo?
  - **una representación** con un **propósito particular**
- ¿Qué es un modelo estadístico?
  - Descripción de un **proceso probabilístico** con **parámetros desconocidos** que deben ser **estimados** en base a **suposiciones** y un conjunto de **datos observados** que puede haber **originado**
- ¿Para qué se usan?
  - **describir** o resumir datos
  - clasificar objetos o **predecir** resultados
  - **anticipar los resultados** de intervenciones (en ocasiones)

Modelo estadístico

- En general, un modelo tiene la forma[1]:

$$y_i = \text{(modelo)} + \text{error}$$

  - y_i es el i-ésimo valor observado de la variable respuesta Y (o variable de salida o **variable dependiente**)
  - (modelo) es el resultado de una **función determinista** basada en un **conjunto de parámetros**, y
  - error es la diferencia **aleatoria** entre estos valores
  - esta ecuación a veces se encuentra como:

$$y_i = \text{(modelo)} + \varepsilon_i$$

$$y_i = f(x_i) + \varepsilon_i$$

Modelo estadístico

- En general, un modelo tiene la forma:

$$y_i = \text{(modelo)} + \text{error}$$

  - aquí "error" **no se refiere** que se ha cometido una "equivocación"
  - sino a la **variación natural** que existe entre los **valores observados** y los **valores pronosticados** por el modelo
  - para no confundir (¿?), a veces se les llama *variación no sistemática*, *variación aleatoria* o **residuos** (e incluso, *residuales*)

Modelo estadístico

- En general, un modelo tiene la forma:

$$y_i = \textbf{(modelo)} + \textbf{error}$$

  - luego, este "error" está relacionado con la **calidad del modelo**
    - entre **menor** el error, **mejor** el modelo
    - y al contrario, errores más grandes son señales de un modelo fallido, que no describe bien los datos, no ayuda a predecirlos bien, no ayuda a clasificarlos bien
    - por ejemplo, ¿qué modelo es mejor para representar la estatura media del equipo?:
      - Y = 100 cm
      - Y = 170 cm

Modelos estadísticos simples

- En general, un modelo tiene la forma:

$$y_i = (\text{modelo}) + \text{error}$$

  - pero ya hemos trabajado con modelos estadísticos simples
  - por ejemplo, si comenzamos a **lanzar un dado**
    - ¿qué proporción de veces obtendremos un número par?
    - ¿qué cantidad de puntos veré en promedio?

Modelos estadísticos simples

- En general, un modelo tiene la forma:

$$y_i = \textbf{(modelo)} + \textbf{error}$$

  - pero ya hemos trabajado con modelos estadísticos simples
  - por ejemplo, si comenzamos a **lanzar un dado**
    - ¿qué proporción de veces obtendremos un número par?

la mitad de las veces

    - ¿qué cantidad de puntos veré en promedio?

3,5

Modelos estadísticos simples

- En general, un modelo tiene la forma:

$$y_i = \textbf{(modelo) + error}$$

  - pero ya hemos trabajado con modelos estadísticos simples
  - por ejemplo, si comenzamos a **lanzar un dado**
    - ¿qué **proporción** de veces obtendremos un número par?

la mitad de las veces

    - ¿qué cantidad de puntos veré en **promedio**?

3,5

Modelos estadísticos simples

- El **promedio** (o **media**) y la **proporción** son entonces modelos estadísticos simples que ya hemos utilizado

$$y_i = \textbf{(media)} + \textbf{error}$$

$$y_i = \textbf{(proporción)} + \textbf{error}$$

- ¿Qué tan buen modelos son estas medidas?
  - tomemos un dado y comencemos a lanzarlo...
  - o **simulemos** en R...

Modelos estadísticos simples

- Por ejemplo, tiremos un dado 6 veces:

```
set.seed(111)
dado <- 1:6
n <- 6
salida <- sample(dado, size = n, replace = TRUE)
```

  - ¿qué esperamos ver en la variable `salida`?
    - que 3 veces sean números pares de puntos ($\bar{p}$ = 0,5), y
    - 3,5 puntos en promedio ($\bar{x}$ = 3,5)
  - veamos:

```
> sum(salida %% 2 == 0) / n
[1] 0.3333333
> mean(salida)
[1] 3.666667
```

  - primera impresión: parece haber **desviaciones** importantes
    - la proporción observada fue más bien 1/3, no 1/2 (error: 1/6)
    - la media estuvo más cerca (error = 1/6)

Modelos estadísticos simples

- ¿Y si repetimos el experimento?

```
set.seed(XXX)
dado <- 1:6
n <- 6
salida <- sample(dado, size = n, replace = TRUE)
print(sum(salida %% 2 == 0) / n)
print(mean(salida))
```

  - veamos que obtiene cada equipo
    - y guardemos los valores en variables

```
props <- c(                                )


proms <- c(                                )
```

Modelos estadísticos simples

- Ahora visualicemos estos datos
  - primero, las proporciones

```
mp <- 0.5
m <- length(props)
dlp <- data.frame(
  Experimento = 1:m,
  Proporción = props,
  Ymin = sapply(props, function(x) min(x, mp)),
  Ymax = sapply(props, function(x) max(x, mp))
)

cols <- c("#6D9EC1", "#D55E00")
p1 <- ggplot(dlp, aes(x = Experimento, y = Proporción))
p1 <- p1 + geom_point(size = 3, colour = cols[1])
p1 <- p1 + coord_cartesian(xlim = c(1, m), ylim = c(0.1, 0.9))
p1 <- p1 + scale_y_continuous(breaks = seq(0.1, 0.9, 0.2))
p1 <- p1 + scale_x_discrete(name = "", breaks = NULL)
p1 <- p1 + geom_hline(aes(yintercept = mp), colour = cols[1])
p1 <- p1 + geom_linerange(
  aes(ymin = Ymin, ymax = Ymax),
  color = cols[2], linetype = "dashed"
)
p1 <- p1 + theme_pubr()
```

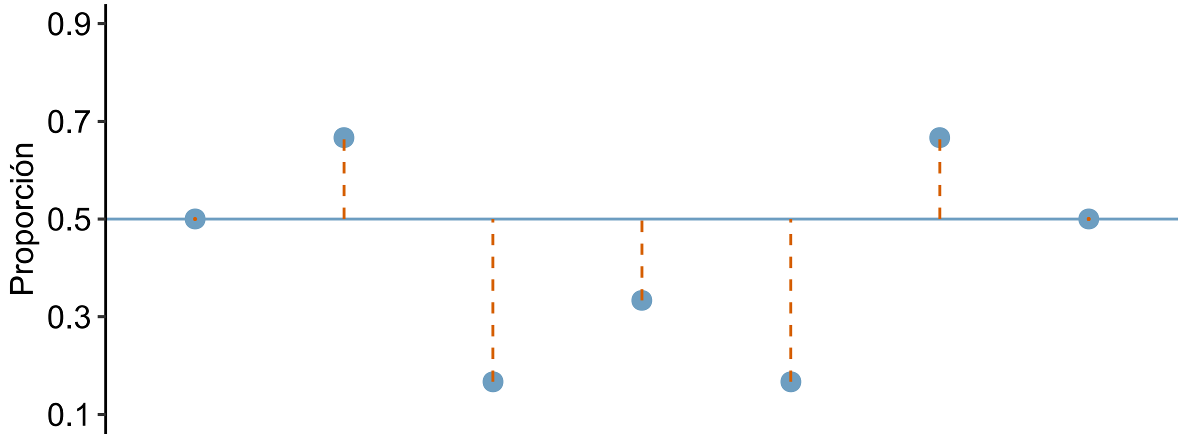
Modelos estadísticos simples

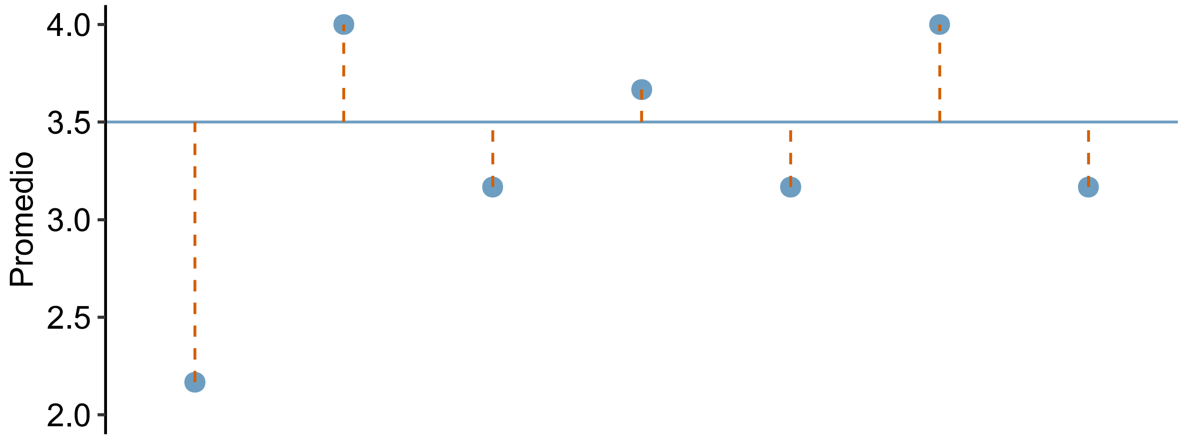
- Ahora visualicemos estos datos
  - luego, los promedios

```
mm <- 3.5
m <- length(proms)
dlm <- data.frame(
  Experimento = 1:m,
  Promedio = proms,
  Ymin = sapply(proms, function(x) min(x, mm)),
  Ymax = sapply(proms, function(x) max(x, mm))
)

cols <- c("#6D9EC1", "#D55E00")
p2 <- ggplot(dlm, aes(x = Experimento, y = Promedio))
p2 <- p2 + geom_point(size = 3, colour = cols[1])
p2 <- p2 + coord_cartesian(xlim = c(1, m), ylim = c(2, 4))
p2 <- p2 + scale_y_continuous(breaks = seq(2, 4, 0.5))
p2 <- p2 + scale_x_discrete(name = "", breaks = NULL)
p2 <- p2 + geom_hline(aes(yintercept = mm), colour = cols[1])
p2 <- p2 + geom_linerange(
  aes(ymin = Ymin, ymax = Ymax),
  color = cols[2], linetype = "dashed"
)
p2 <- p2 + theme_pubr()
```

Desviaciones

- Ahora visualicemos estos datos
  - Resultan gráficos como estos





Desviaciones

- ¿Y? ¿son modelos buenos, malos, más o menos?
  - calculemos las desviaciones

```
> props - mp

> sum(props - mp)


> proms - mm

> sum(proms - mm)
```

Desviaciones

- ¿Y? ¿son modelos buenos, malos, más o menos?
  - calculemos las desviaciones, que en mi caso son:

```
> props - mp
[1]  0.000000  0.166667 -0.333333 -0.166667 -0.333333  0.166667  0.000000
> sum(props - mp)
[1] -0.5

> proms - mm
[1] -1.333333  0.500000 -0.333333  0.166667 -0.333333  0.500000 -0.333333
> sum(proms - mm)
[1] -1.166667
```

  - vemos que hay valores **positivos** y **negativos**
    - que **pueden anularse** fácilmente

Medidas de variabilidad

- ¿Qué hacemos entonces?
  - aunque no es la única alternativa, lo más usado es el **cuadrado de las desviaciones**
  - la **suma de los cuadrados de las desviaciones** se usa como **medida de la variabilidad** en los datos
    - se suele denotar **SS** del inglés *sum of squared deviations*

```
> (props - mp)^2
[1] 0.000000 0.0277778 0.111111 0.0277778 0.111111 0.0277778 0.000000
> sum((props - mp)^2)
[1] 0.305556

> (proms - mm)^2
[1] 1.777778 0.250000 0.111111 0.0277778 0.111111 0.250000 0.111111
> sum((proms - mm)^2)
[1] 2.638889
```

    - pero SS tiene un inconveniente: **crece** con más datos

Medidas de variabilidad

- ¿Qué hacemos entonces?
  - normalizar al **cuadrado de las desviaciones promedio**
    - se suele denotar **MS** del inglés *mean SS*
    - dividiendo SS por los **grados de libertad**
      - denotado gdl (o gl) o dof (o **df**) del inglés *degrees of freedom*
      - modelos que fijan **un solo valor** (el promedio, la proporción) para un conjunto de ***m*** valores, tienen **(*m* – 1)** grados de libertad

```
> sum((props - mp)^2) / (m - 1)
[1] 0.050926

> sum((proms - mm)^2) / (m - 1)
[1] 0.439815
```

    - pero ahora, la medida **no está en la misma escala** que la variable original

Medidas de variabilidad

- ¿Qué hacemos entonces?
  - por lo que se suele usar la **raíz cuadrada** del MS
    - se suele denotar **RMSD** o **RMSE** por sus siglas en inglés

```
> sqrt(sum((props - mp)^2) / (m - 1))
[1] 0.225668

> sqrt(sum((proms - mm)^2) / (m - 1))
[1] 0.663186
```

    - así, para los datos del ejemplo, **el modelo cometió en promedio**
      - **23% de error** al predecir la proporción de números pares obtenidos al lanzar un dado 7 veces
      - **0,66 puntos** de error al predecir la media de puntos obtenidos al lanzar un dado 7 veces
    - si este es o no un **nivel de error aceptable** dependerá del contexto de los datos

Medidas de variabilidad

- ¿Qué hacemos entonces?
  - por lo que se suele usar la **raíz cuadrada** del MS
    - se suele denotar **RMSD** o **RMSE** por sus siglas en inglés

```
> sqrt(sum((props - mp)^2) / (m - 1))
[1] 0.225668

> sqrt(sum((proms - mm)^2) / (m - 1))
[1] 0.663186
```

    - así, para los datos del ejemplo, **el modelo cometió en promedio**
      - **23% de error** al predecir la proporción de números pares obtenidos al lanzar un dado 7 veces
      - **0,66 puntos** de error al predecir la media de puntos obtenidos al lanzar un dado 7 veces
    - si este es o no un **nivel de error aceptable** dependerá del contexto de los datos
  - Notemos: MS y RMSD son la **varianza** y la **desviación estándar** en distribuciones normales

En resumen

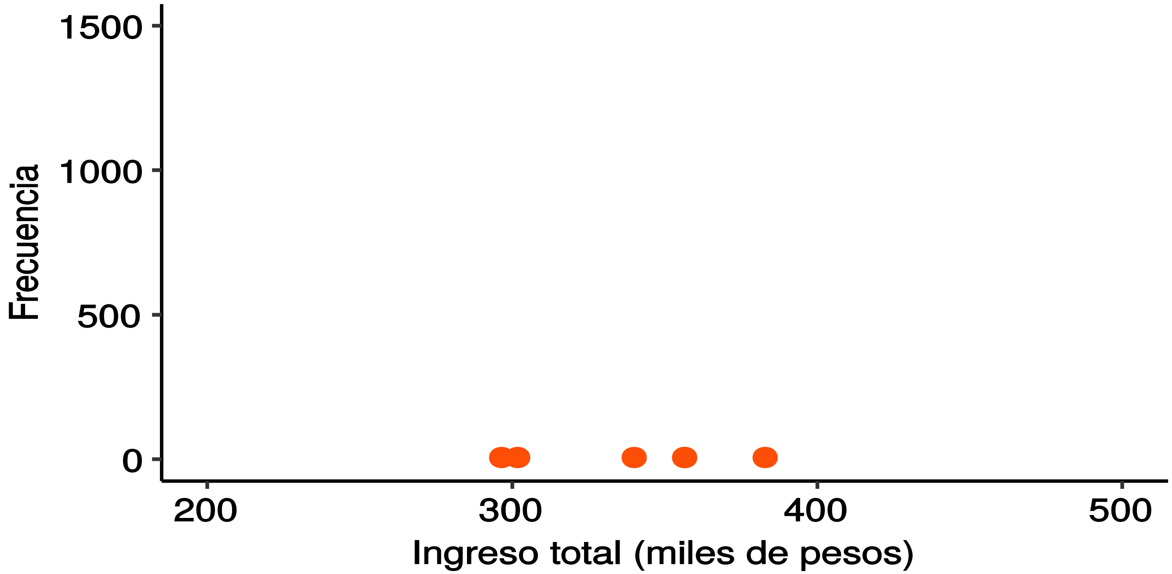
- La ecuación más importante del semestre:

$$y_i = (\text{modelo}) + \text{error}$$

  - en general, **menor error = mejor modelo**
- La **media** y la **proporción** son modelos muy simples
  - la media nos resume una v.a. **numérica**
  - la proporción nos resume una v.a. **categórica**
- Cuando no se considera a toda la población
  - hay **variaciones naturales** debido al **muestreo**
  - este es el "error" de la ecuación, no es "equivocación"
  - se puede medir por medio de la **varianza** (MS) o su raíz

Desviaciones

- Pero esta variabilidad natural ¿es un problema?
  - la hacer estimaciones puntuales de la media poblacional



  - ¿solo incertidumbre?
    - hasta mediados del siglo XVII se pensaba que sí
    - desde entonces, grandes matemáticos (e.g. Pascal, Fermat, Mèré, Bernoulli, Laplace, etc.) han encontrado que no todo está perdido

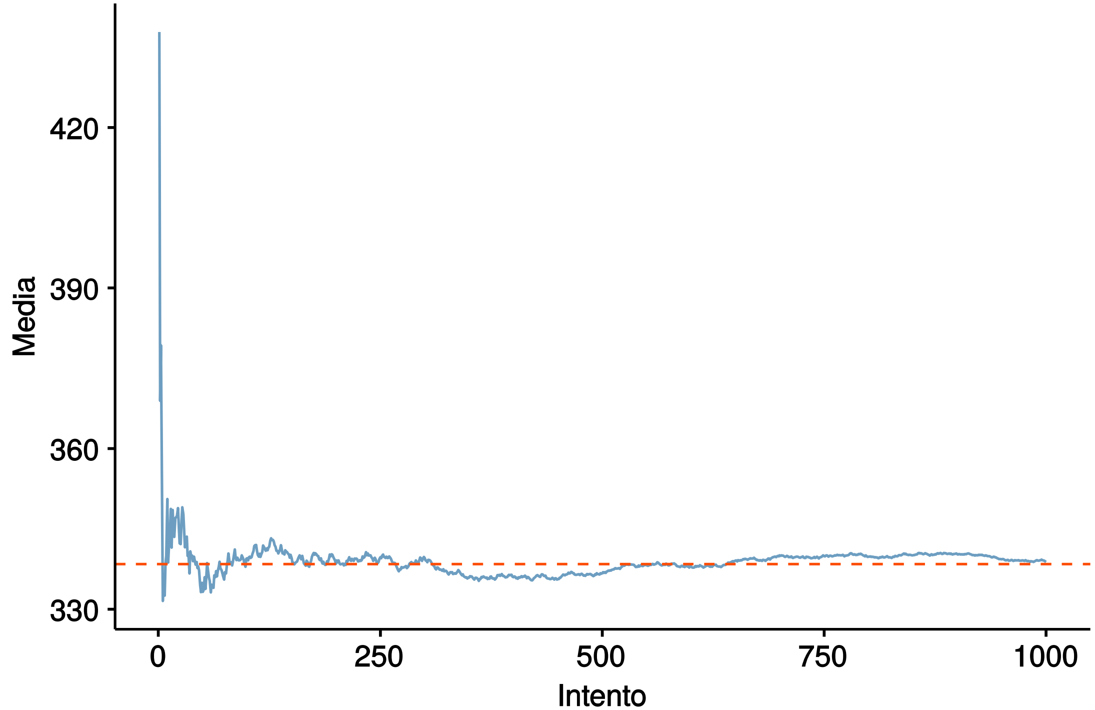
Media acumulada

- A la larga...

```
tamaño.muestra <- 1000

set.seed(semilla)
muestra <- sample(población, tamaño.muestra)

media.acum <- cumsum(muestra) / seq(along = muestra)
media.población <- mean(población)

m1 <- data.frame(
  Intento = 1:tamaño.muestra,
  Media = media.acum
)

p1 <- ggline(
  data = m1, x = "Intento", "Media",
  plot_type = "l",
  color = "#6D9EC1"
)
p1 <- p1 + geom_hline(
  aes(yintercept = media.población),
  color = "#FC4E07", linetype = 2
)
```

Media acumulada

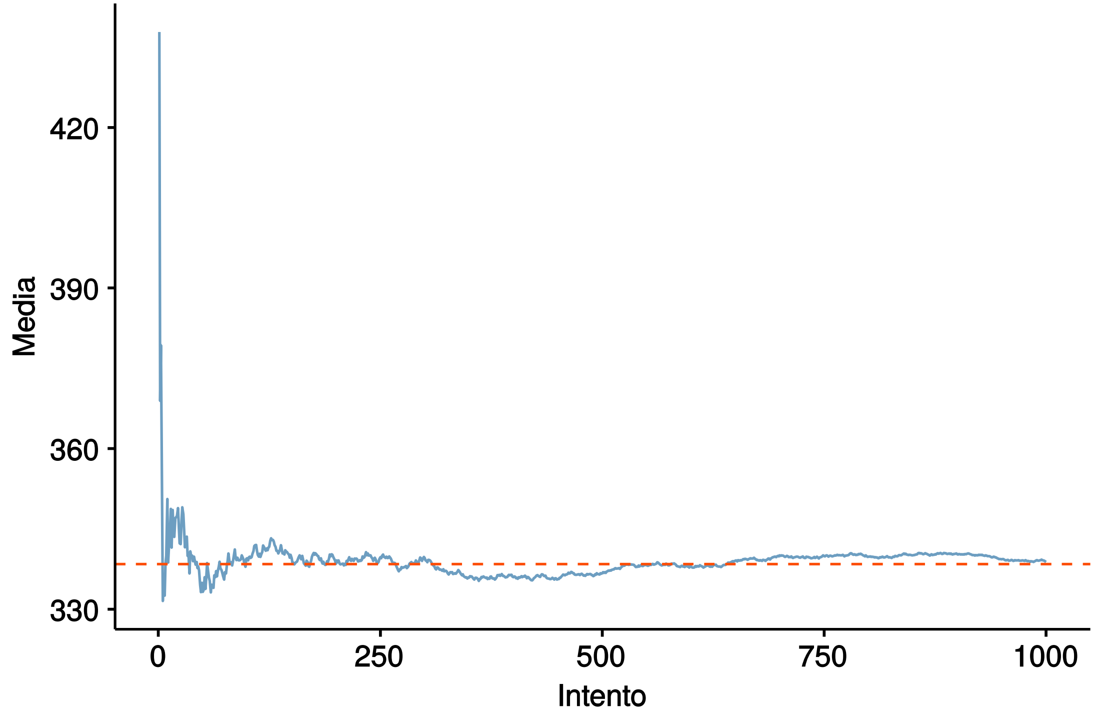
- A la larga...



  - se converge a la “verdadera” media de la población
  - ¿cómo se llama esto?

Media acumulada

- A la larga...

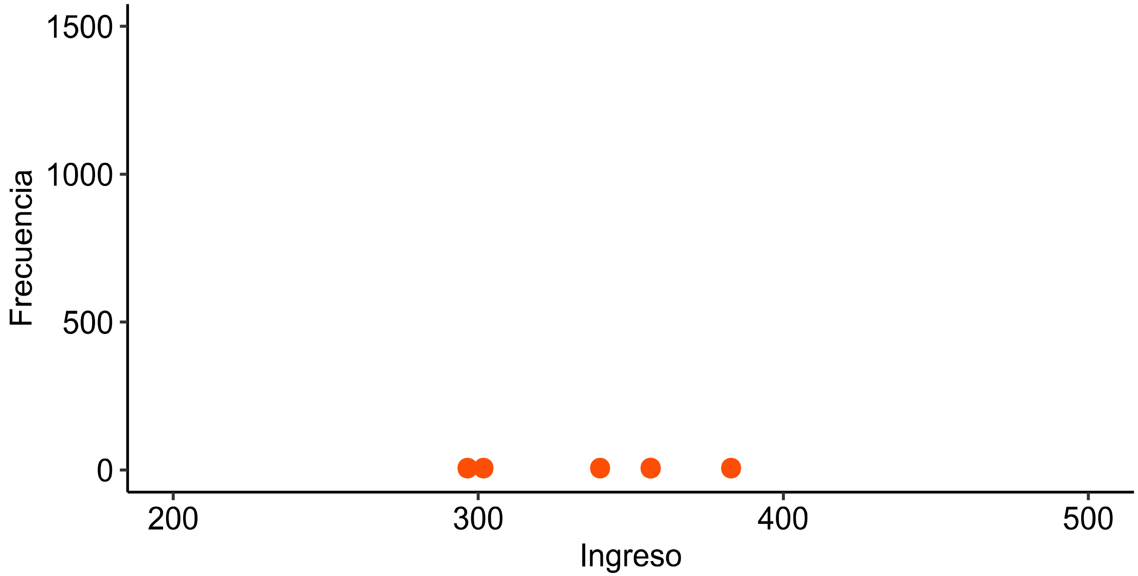


  - se converge a la "verdadera" media de la población
  - ¿cómo se llama esto?

Ley de los grandes números

Estimaciones de la media

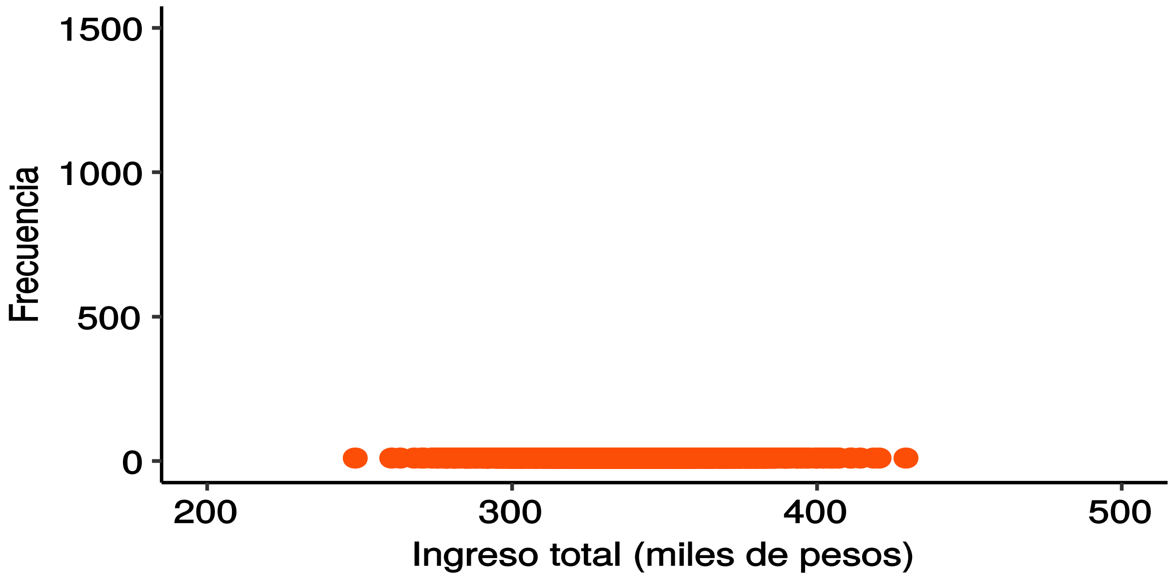
- Volvamos al gráfico de estimaciones puntuales
  - que muestra medias estimadas con diferentes muestras



  - repitamos lo mismo pero con 10.000 muestras

Estimaciones de la media

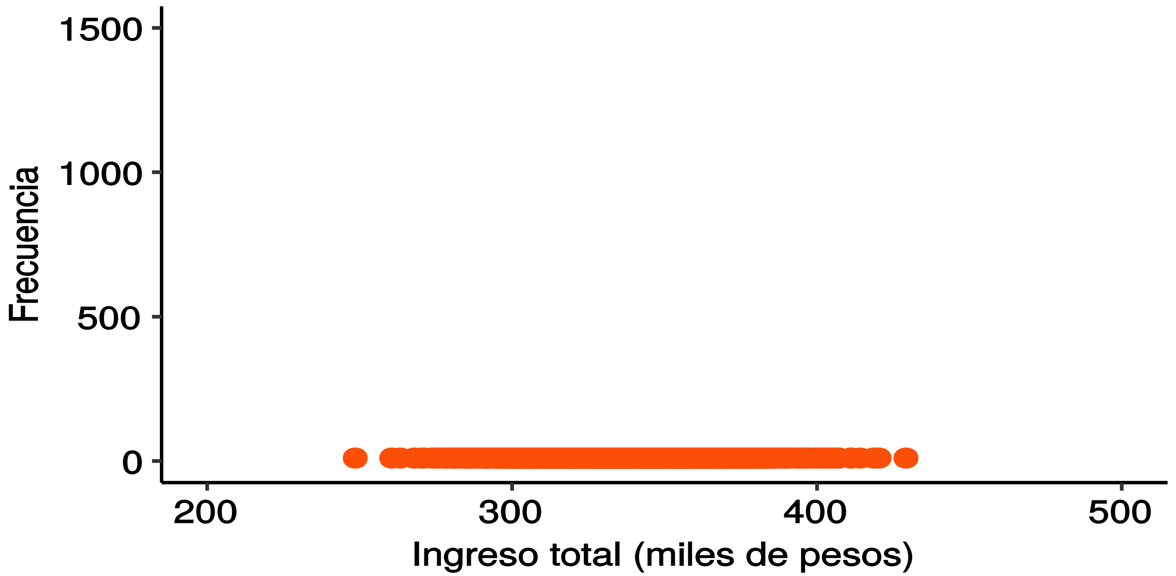
- Volvamos al gráfico de estimaciones puntuales
  - que muestra medias estimadas con diferentes muestras



  - ¿vemos algo?

Estimaciones de la media

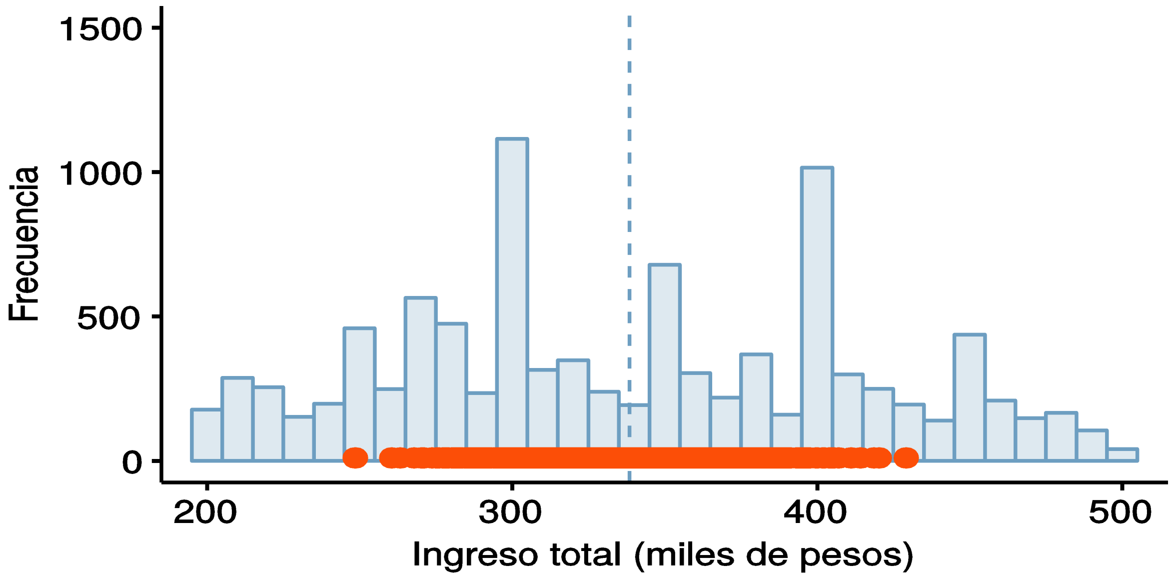
- Volvamos al gráfico de estimaciones puntuales
  - que muestra medias estimadas con diferentes muestras



- ¿vemos algo?
  - al menos **no** se van a a los **extremos**
  - ¿veamos la población? (somos dioses)

Estimaciones de la media

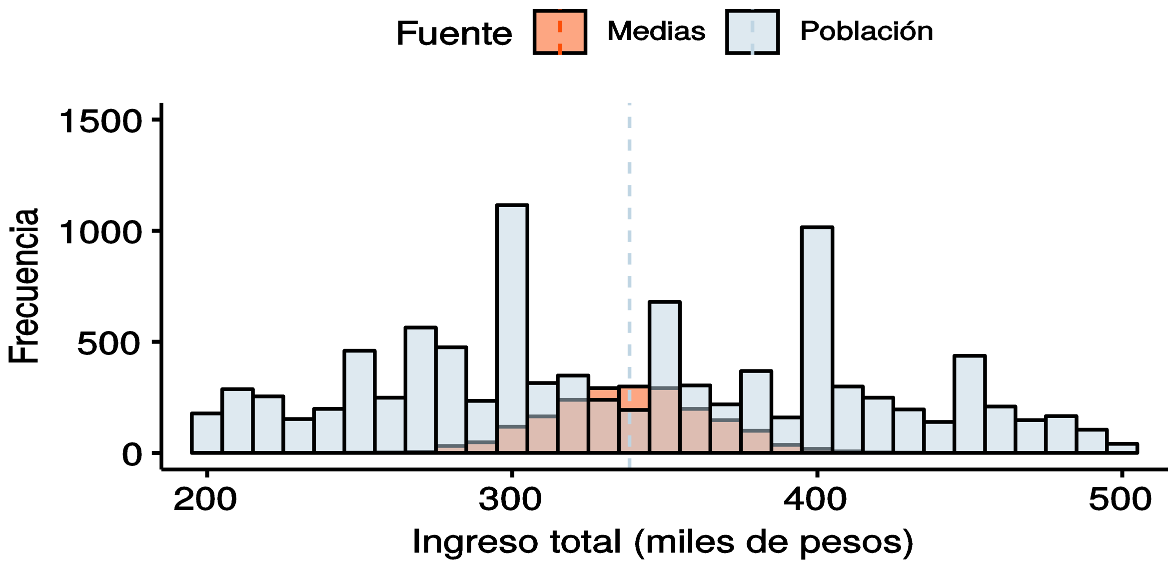
- Volvamos al gráfico de estimaciones puntuales
  - que muestra medias estimadas con diferentes muestras



- ¿vemos algo?
  - al menos **no** se van a a los **extremos**
  - ¿veamos la población (somos dioses)?
  - aquí podría ser importante **las frecuencias**

Estimaciones de la media

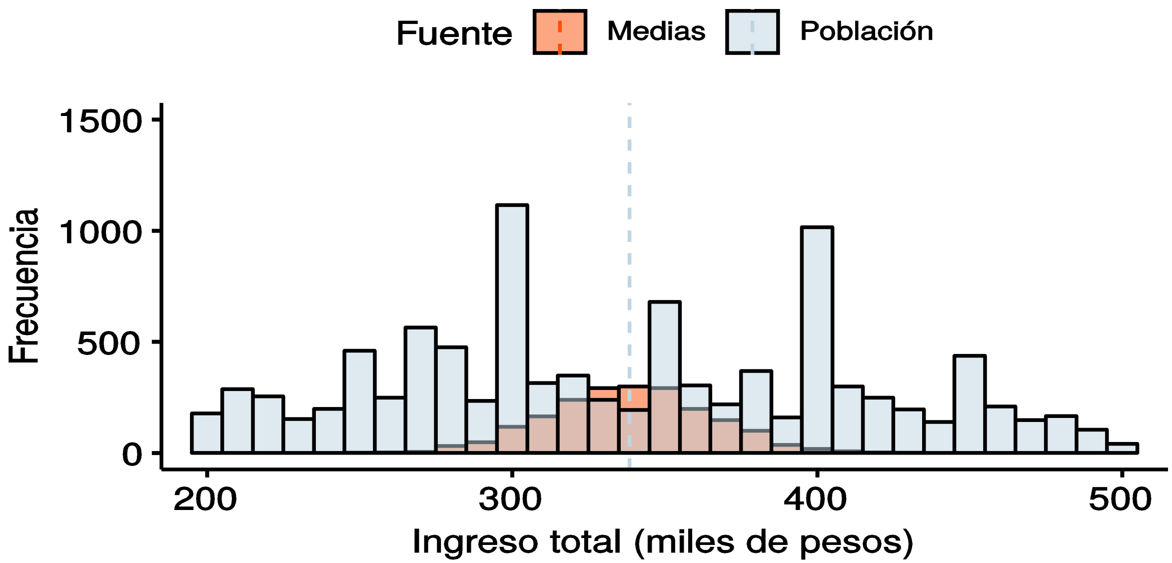
- Volvamos al gráfico de estimaciones puntuales
  - que muestra medias estimadas con diferentes muestras



  - ¿qué significa esto?

Estimaciones de la media

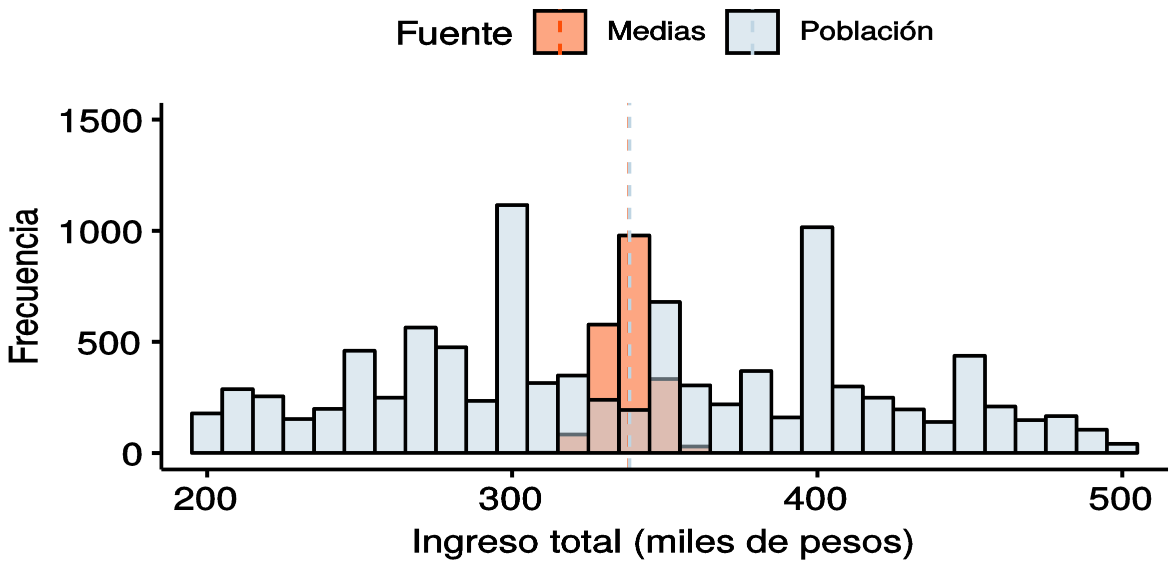
- Volvamos al gráfico de estimaciones puntuales
  - que muestra medias estimadas con diferentes muestras



  - ¿qué significa esto?
    - parece que las **medias de las muestras** se **aglutinan** alrededor de la **media de la población**
    - ¡no todo es oscuridad!

Estimaciones de la media

- Volvamos al gráfico de estimaciones puntuales
  - ahora medias estimadas con **muestras más grandes**



  - ¿qué paso?
    - las medias de las muestras se **aglutinan más cerca** de la media de la población

Referencias

[1] A. Field, J. Miles, Z. Field (2012). Discovering statistics using R. Sage publications.

Estas ideas también se encuentran en:

- David M. Diez, Christopher D. Barr, Mine Çetinkaya-Rundel (2015). OpenIntro Statistics; 3rd Edition. Disponible en www.openintro.org
- Rudolf J. Freund, William J. Wilson, Donna L. Mohr (2010). Statistical Methods; 3rd Edition, Academic Press
- Jay L. Devore (2011). Probability and Statistics for Engineering and the Sciences; 8th Edition, Duxbury Press
- David A. Freedman (2009). Statistical Models: Theory and Practice. Revised Edition. Cambridge University Press