Introducción a los Modelos Lineales: fundamentos teóricos y prácticos

Javier Fernández-López, Profesor Ayudante Doctor
Unidad de Matemática Aplicada
Departamento de Biodiversidad, Ecología y Evolución, UCM

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

DOKTOREGO ESKOLA
ESCUELA DE DOCTORADO
DOCTORAL SCHOOL

Programas de doctorado:
- Agrobiología Ambiental
- Calidad y Seguridad Alimentaria
- Biodiversidad, Funcionamiento y Gestión de Ecosistemas

Sobre mi…

Javier Fernández-López
javfer05@ucm.es

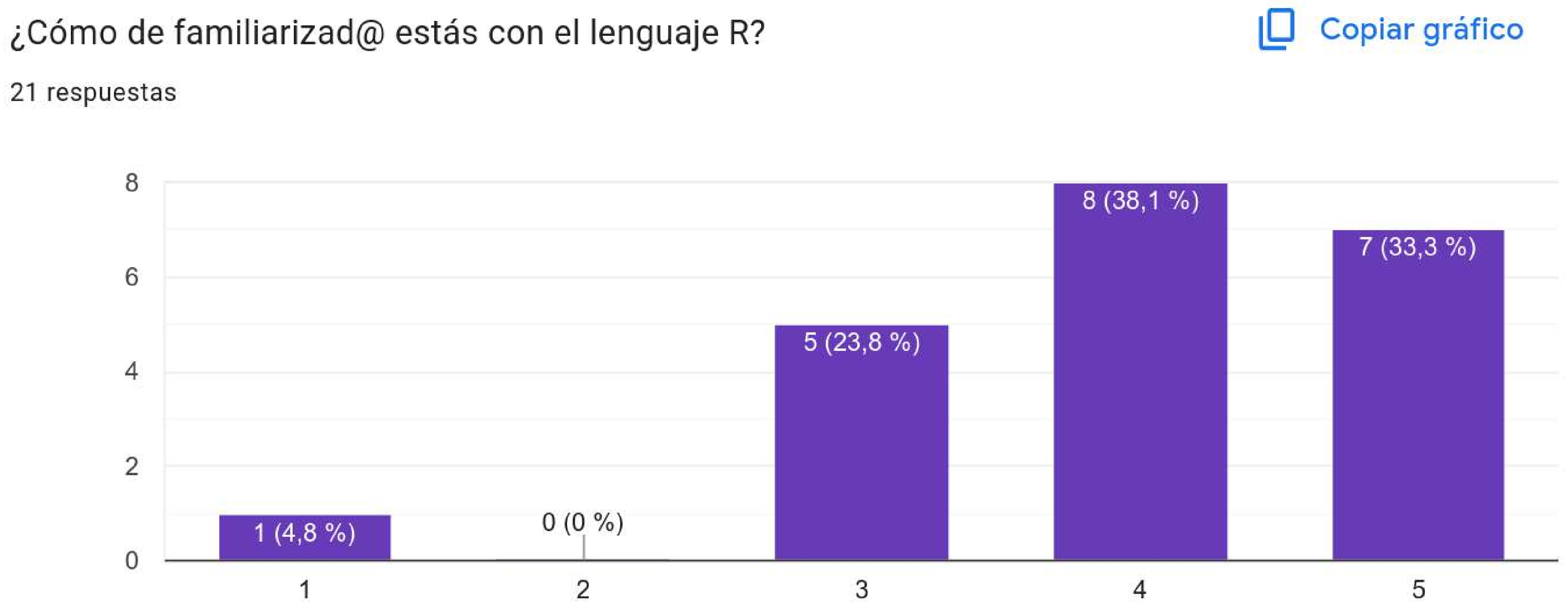
REAL JARDÍN BOTÁNICO

irec

UMASS BOSTON

CENTRE D'ECOLOGIE FONCTIONNELLE & EVOLUTIVE

Sobre vosotrxs...

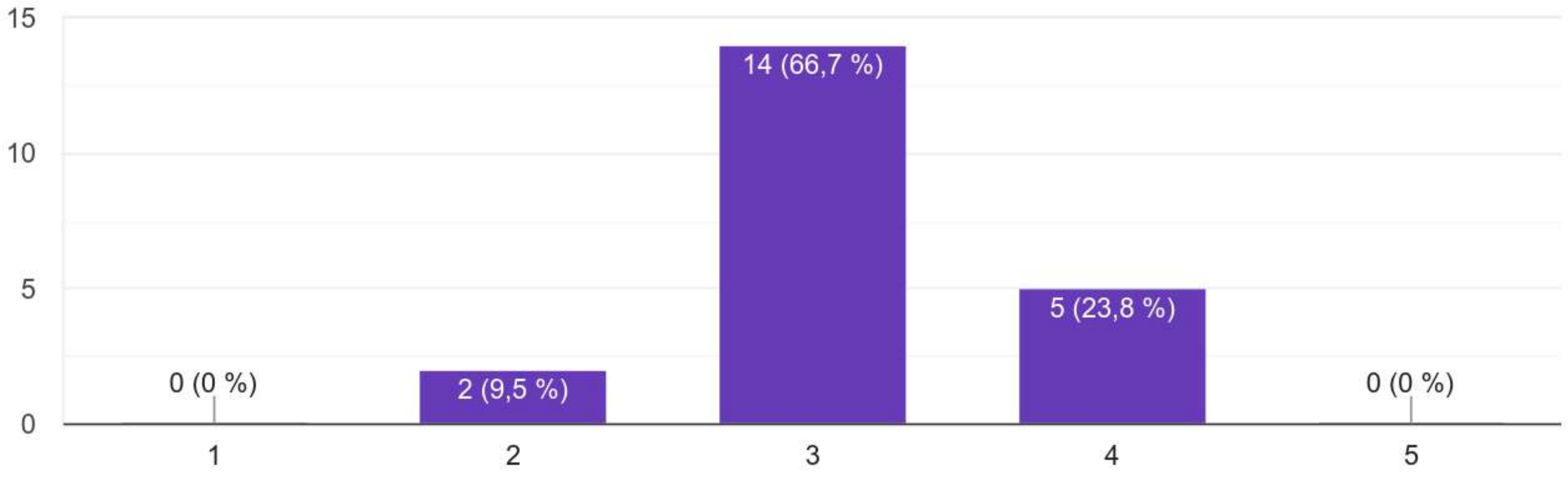


Sobre vosotrxs...

¿Cómo de familiarizad@ estás con los modelos lineales (regresiones, GLM, etc.)?

Copiar gráfico

21 respuestas



Sobre vosotrxs...

Grupo muy diverso (tres programas de doctorado):
50% BFGE
20% CSA
23.3% ABA
6.7% otros.

Muchas disciplinas diferentes: ganadería, agricultura, ecología, microbiología, calidad alimentaria, nutrición, botánica, zoología…

Reto, pero también oportunidad!

Sobre el curso...

- Módulo 1 - Introducción a la modelización

  (5 horas, 20 y 27 de octubre)

Sobre el curso...

- Módulo 1 - Introducción a la modelización

  (5 horas, 20 y 27 de octubre)

- Módulo 2 - Modelos lineales generalizados

  (5 horas, 3 y *10?* de noviembre)

Sobre el curso...

- Módulo 1 - Introducción a la modelización

  (5 horas, 20 y 27 de octubre)

- Módulo 2 - Modelos lineales generalizados

  (5 horas, 3 y *10?* de noviembre)

- Módulo 3 - Modelos mixtos y otras extensiones

  (5 horas, *17?* y 24 de noviembre)

Sobre el curso...

- Dinámica. Teoría y práctica (y Paint)

Sobre el curso...

- Dinámica. Teoría y práctica (y Paint)

- Presentaciones, material y grabación:

  https://jabiologo.github.io/web/tutorials/estadisticaUPV.html

Sobre el curso...

- Dinámica. Teoría y práctica (y Paint)
- Presentaciones, material y grabación:

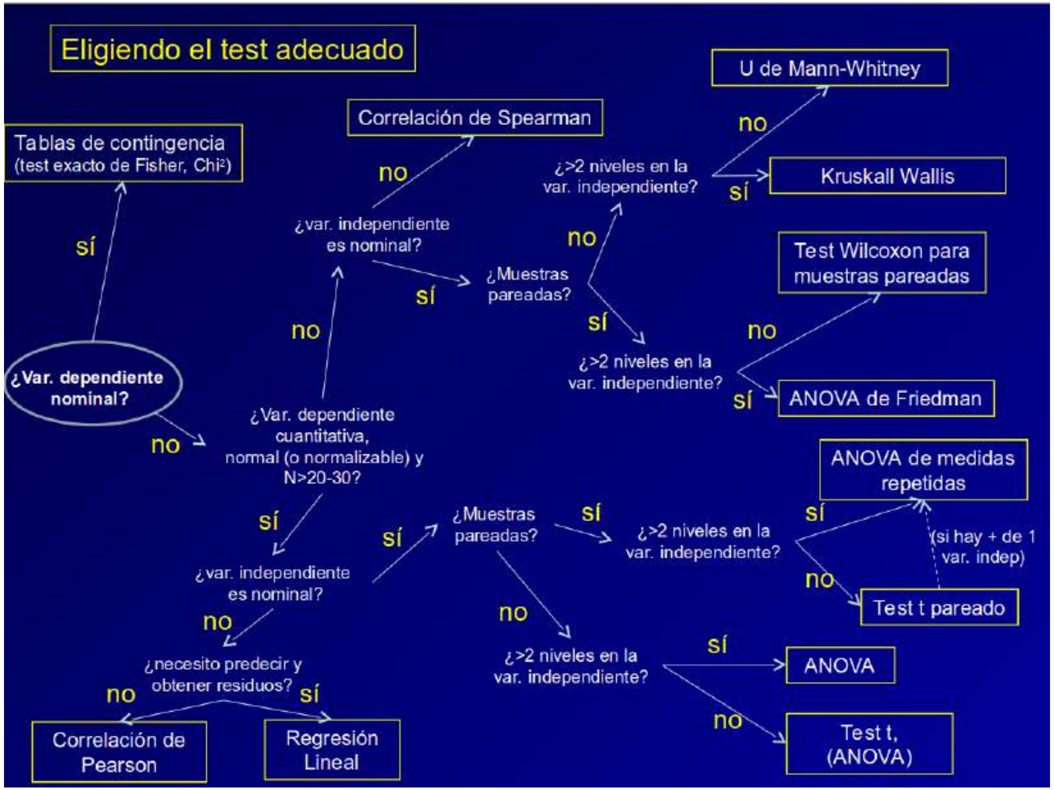
  https://jabiologo.github.io/web/tutorials/estadisticaUPV.html
- Dudas y preguntas: estadística (interrumpir?) y software (esperar?)

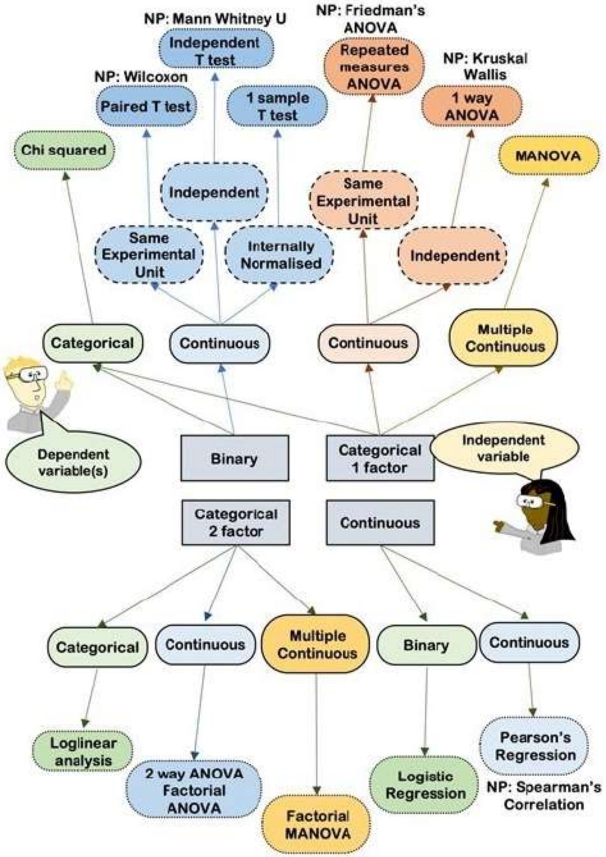
Sobre el curso...

- Dinámica. Teoría y práctica (y Paint)
- Presentaciones, material y grabación:

  https://jabiologo.github.io/web/tutorials/estadisticaUPV.html
- Dudas y preguntas: estadística (interrumpir?) y software (esperar?)
- Filosofía docente

Sobre el curso...





Bloque 1: Introducción a la modelización

- El método científico y la estadística
- Diseños experimentales VS datos observacionales
- Concepto de modelo
- Las variables y las distribuciones de probabilidad
- Las simulaciones como un laboratorio
- Introducción al modelo lineal

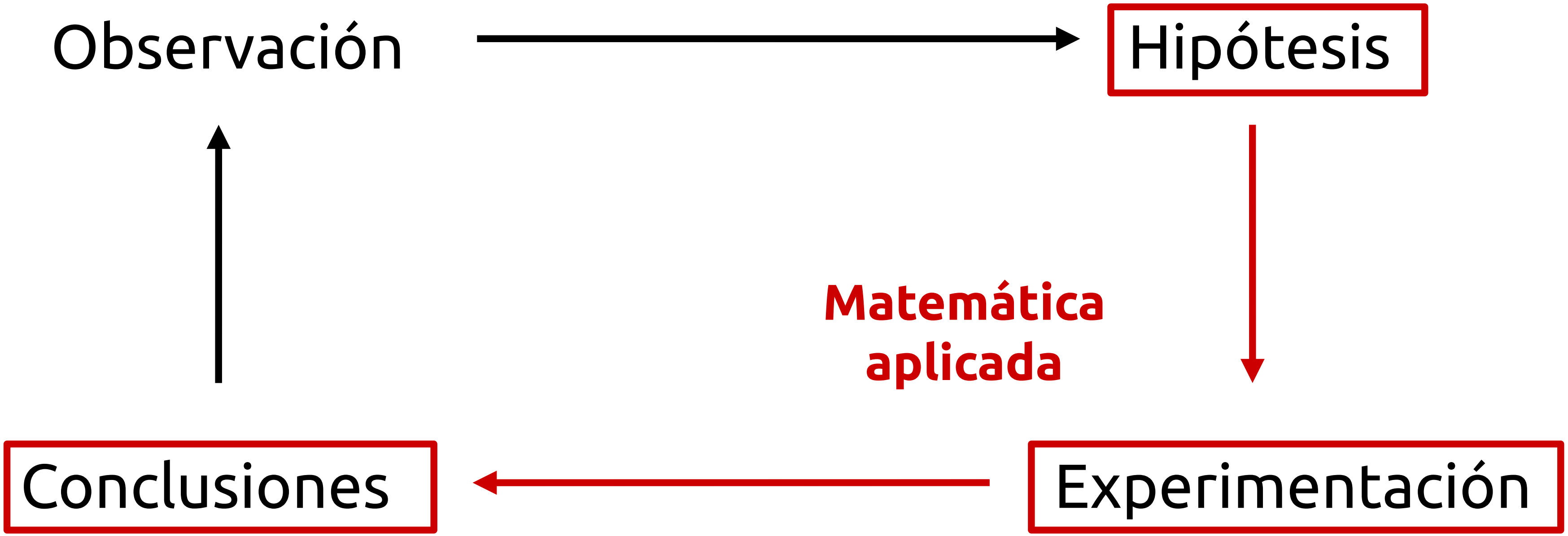
Bloque 1: Introducción a la modelización

- El método científico y la estadística
- Diseños experimentales VS datos observacionales
- Concepto de modelo
- Las variables y las distribuciones de probabilidad
- Las simulaciones como un laboratorio
- Introducción al modelo lineal

El método científico

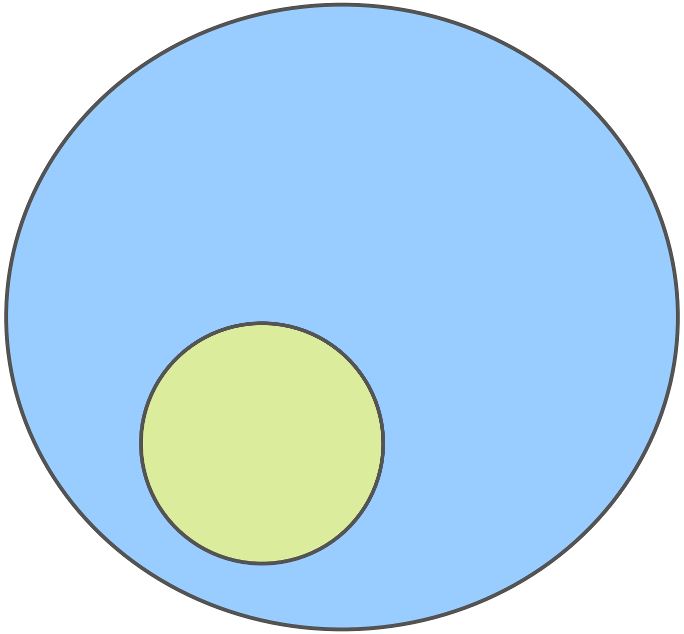
$$
\begin{tikzcd}
\text{Observación} \arrow[r] & \text{Hipótesis} \arrow[d] \\
\text{Conclusiones} \arrow[u] & \text{Experimentación} \arrow[l]
\end{tikzcd}
$$

El método científico



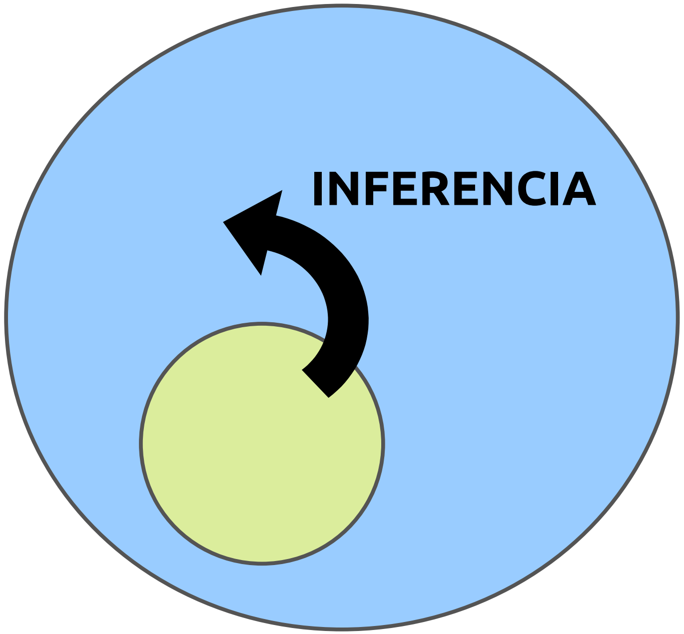
- Población: Conjunto **total de individuos u observaciones** sobre los que queremos sacar conclusiones.

- Muestra: **Subconjunto representativo** de la población, seleccionado para su estudio.



- Población: Conjunto **total de individuos u observaciones** sobre los que queremos sacar conclusiones.

- Muestra: **Subconjunto representativo** de la población, seleccionado para su estudio.



Inferencia estadística

El proceso de usar los datos de una muestra para estimar o tomar decisiones sobre una población, reconociendo la incertidumbre del muestreo.

DISEÑO EXPERIMENTAL VS DATOS OBSERVACIONALES

	Diseño experimental	Estudio observacional
Control	El investigador **manipula** una o más variables (tratamientos) y **controla** las condiciones.	El investigador **no manipula** nada; solo **observa y registra** lo que ocurre naturalmente.
Ejemplo		
Fuentes de variabilidad		

DISEÑO EXPERIMENTAL VS DATOS OBSERVACIONALES

	Diseño experimental	Estudio observacional
Control	El investigador **manipula** una o más variables (tratamientos) y **controla** las condiciones.	El investigador **no manipula** nada; solo **observa y registra** lo que ocurre naturalmente.
Ejemplo	Ensayo clínico, experimento de laboratorio, cultivo en condiciones controladas.	Ecología de campo, estudios sociales, encuestas, registros médicos.
Fuentes de variabilidad		

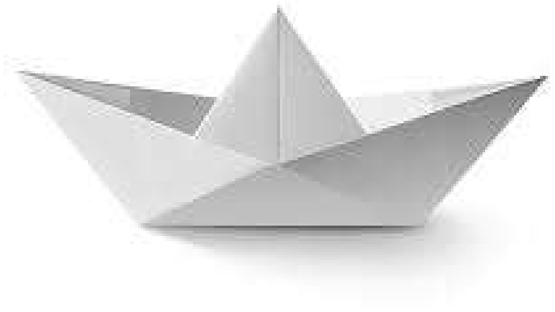
DISEÑO EXPERIMENTAL VS DATOS OBSERVACIONALES

	Diseño experimental	Estudio observacional
Control	El investigador **manipula** una o más variables (tratamientos) y **controla** las condiciones.	El investigador **no manipula** nada; solo **observa y registra** lo que ocurre naturalmente.
Ejemplo	Ensayo clínico, experimento de laboratorio, cultivo en condiciones controladas.	Ecología de campo, estudios sociales, encuestas, registros médicos.
Fuentes de variabilidad	Pueden **controlarse o aislarse** (aleatorización, réplicas, bloqueos).	Difíciles de controlar: el entorno, el comportamiento, la historia previa...

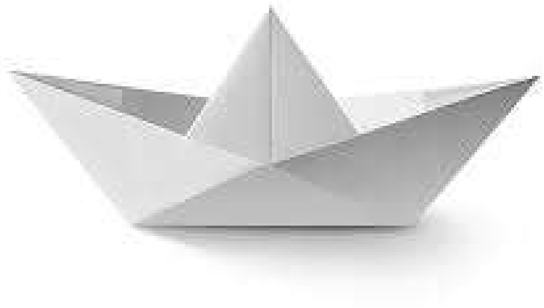
En experimentos, la estadística ayuda a diseñar los tratamientos, minimizar sesgos, y evaluar diferencias entre grupos. En estudios observacionales permite controlar confusores, modelar la variabilidad natural y extraer patrones sin control directo.

¿Qué es un modelo?

¿Qué es un modelo?

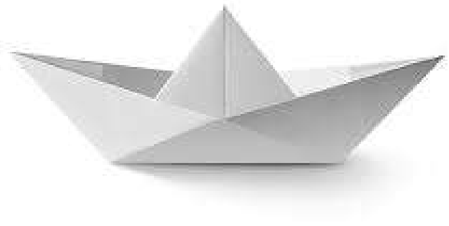


¿Qué es un modelo?



¿Qué es un modelo?

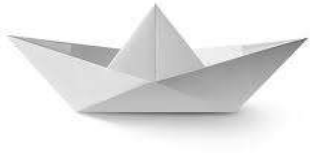
Un modelo es una **representación simplificada** de la realidad, elaborada con el fin de **explicar y/o predecir** ciertos aspectos de la misma



¿Qué es un modelo?

Un **modelo matemático** es la traducción de una hipótesis sobre la realidad a lenguaje matemático, de modo que pueda ponerse a prueba y contrastarse con datos.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

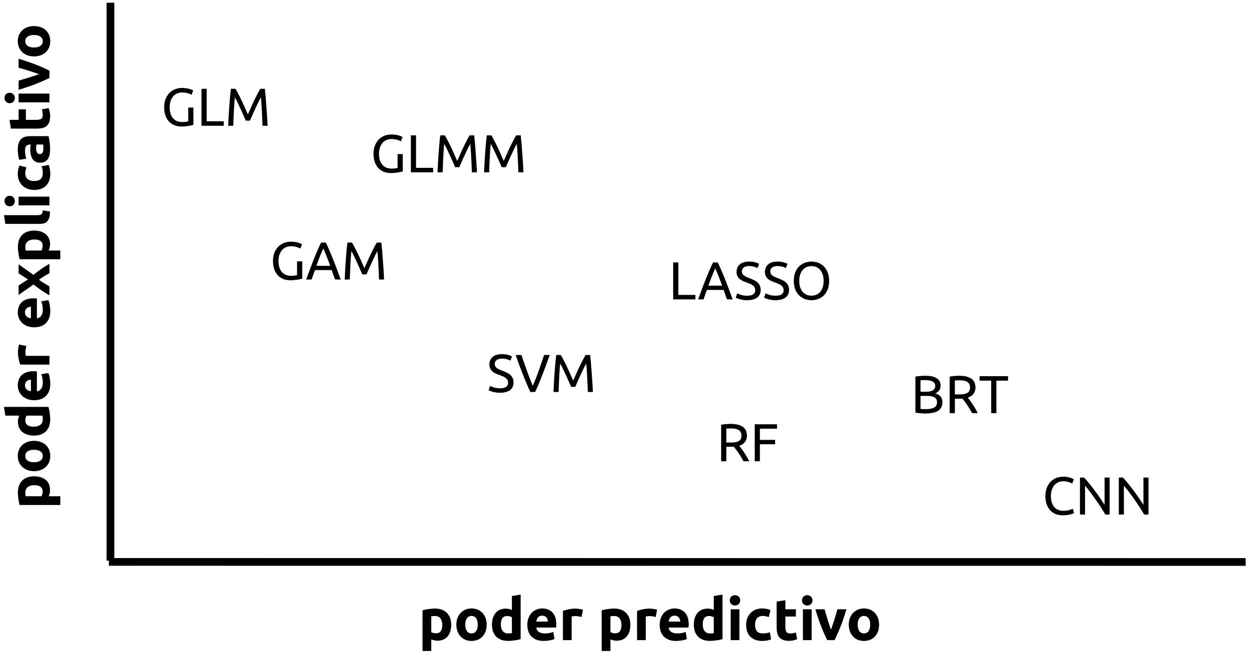


Sobre explicaciones y predicciones..

Un modelo es una **representación simplificada** de la realidad, elaborada con el fin de **explicar** **y/o** **predecir** ciertos aspectos de la misma

Sobre explicaciones y predicciones..

Un modelo es una **representación simplificada** de la realidad, elaborada con el fin de **explicar y/o predecir** ciertos aspectos de la misma



Tipos de variables (clasificación clásica)

Tipos de variables (clasificación clásica)

- Cuantitativas (se expresa mediante números, se puede operar)

Continuas (decimales)

Discretas (enteros)

Tipos de variables (clasificación clásica)

- Cuantitativas (se expresa mediante números, se puede operar)

Continuas (decimales)

Discretas (enteros)

- Semicuantitativas u ordinales (variable cualitativa que se puede ordenar)

Tipos de variables (clasificación clásica)

- Cuantitativas (se expresa mediante números, se puede operar)

Continuas (decimales)

Discretas (enteros)

- Semicuantitativas u ordinales (variable cualitativa que se puede ordenar)

- Cualitativas o nominales (describe un atributo o categoría)

Tipos de variables (clasificación clásica)

- Cuantitativas (se expresa mediante números, se puede operar)

Continuas (decimales)

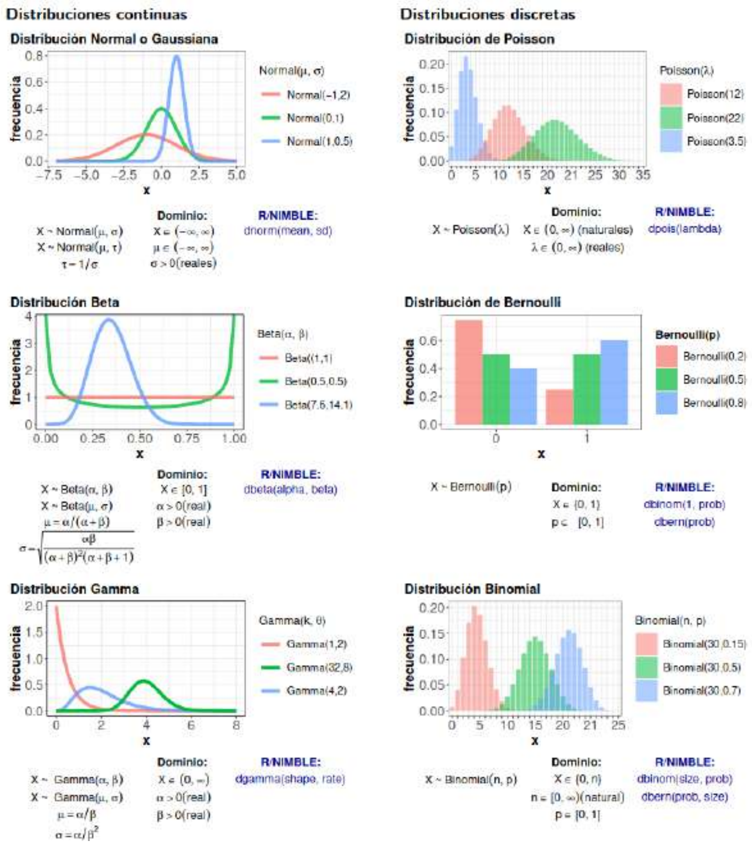
Discretas (enteros)

- Semicuantitativas u ordinales (variable cualitativa que se puede ordenar)

- Cualitativas o nominales (describe un atributo o categoría)

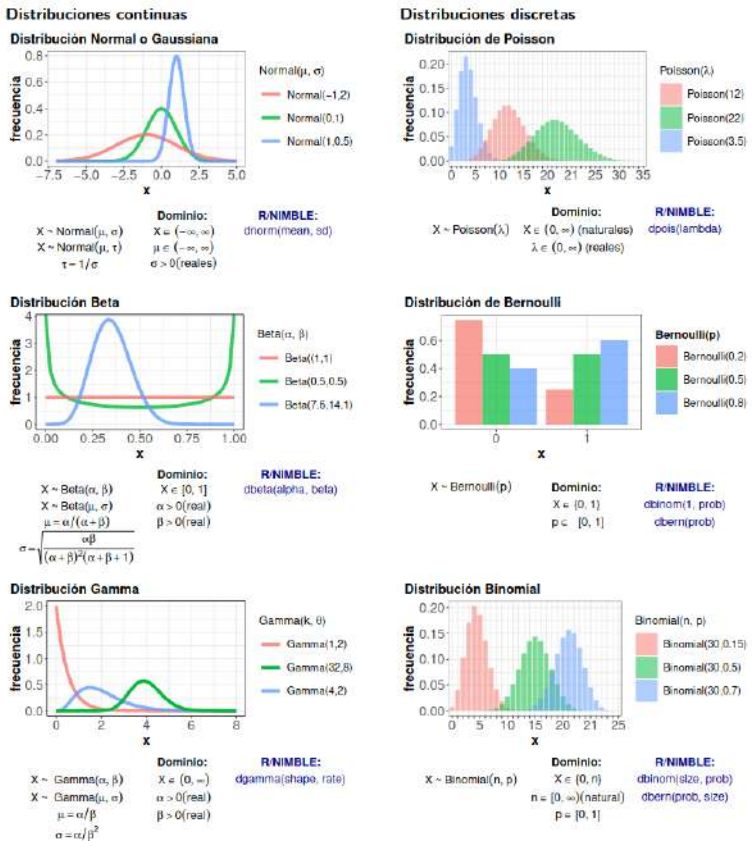
¿Con qué variables trabajáis? ¿de qué tipo son?

Distribuciones de probabilidad



- Van a ser nuestros modelos, nuestra abstracción matemática, la representación simplificada de la realidad!

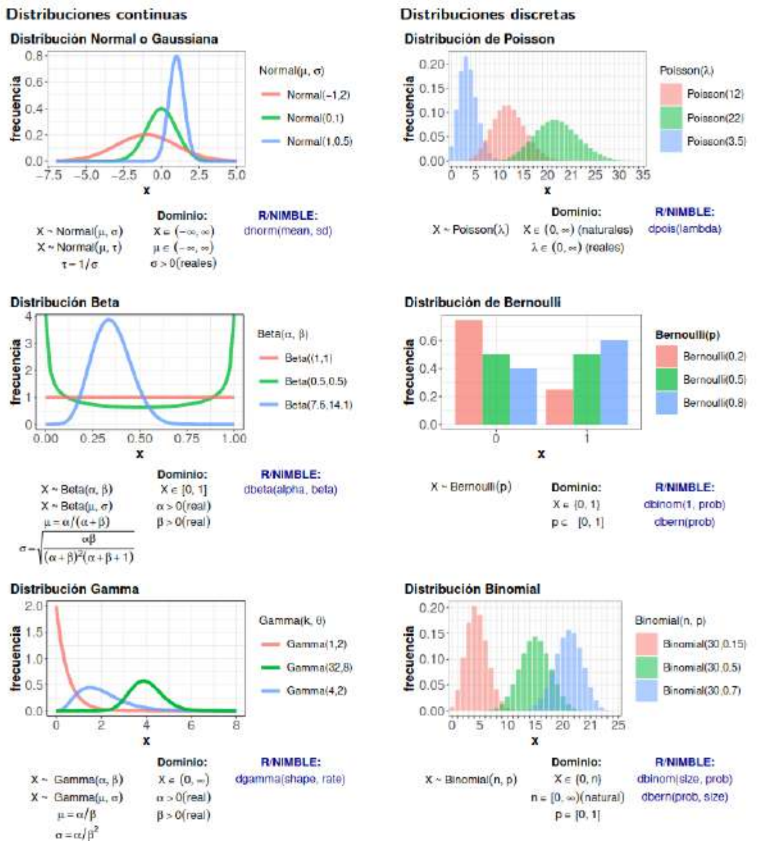
Distribuciones de probabilidad



- Van a ser nuestros modelos, nuestra abstracción matemática, la representación simplificada de la realidad!

- Cada una tiene unas características diferentes que hacen que sean más aptas para uno u otro tipos de datos.

Distribuciones de probabilidad

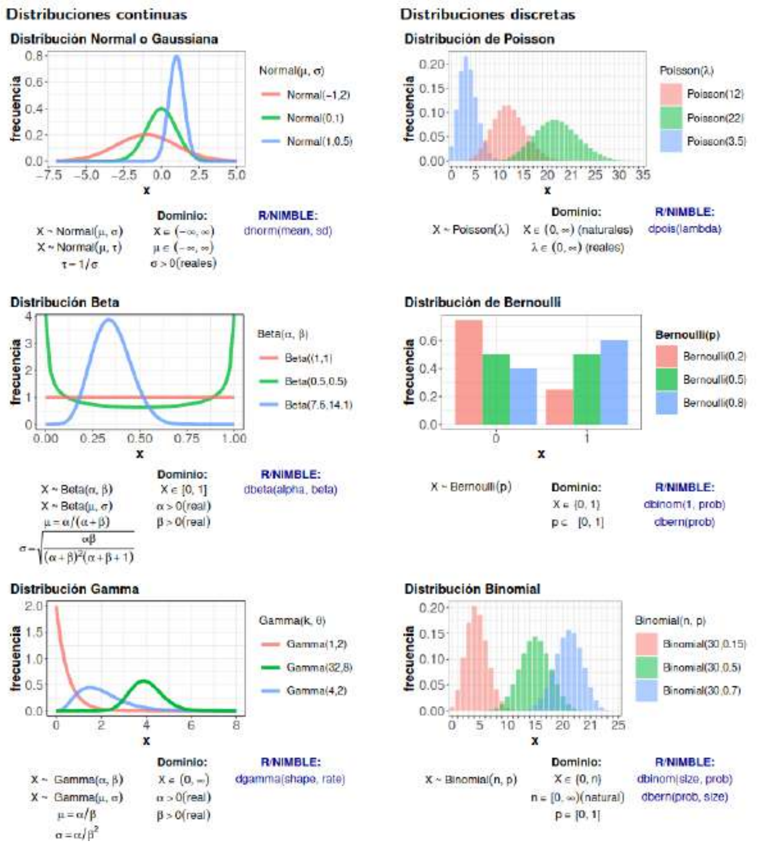


- Van a ser nuestros modelos, nuestra abstracción matemática, la representación simplificada de la realidad!

- Cada una tiene unas características diferentes que hacen que sean más aptas para uno u otro tipos de datos.

- Es necesario conocer las más importantes y sus principales características.

Distribuciones de probabilidad



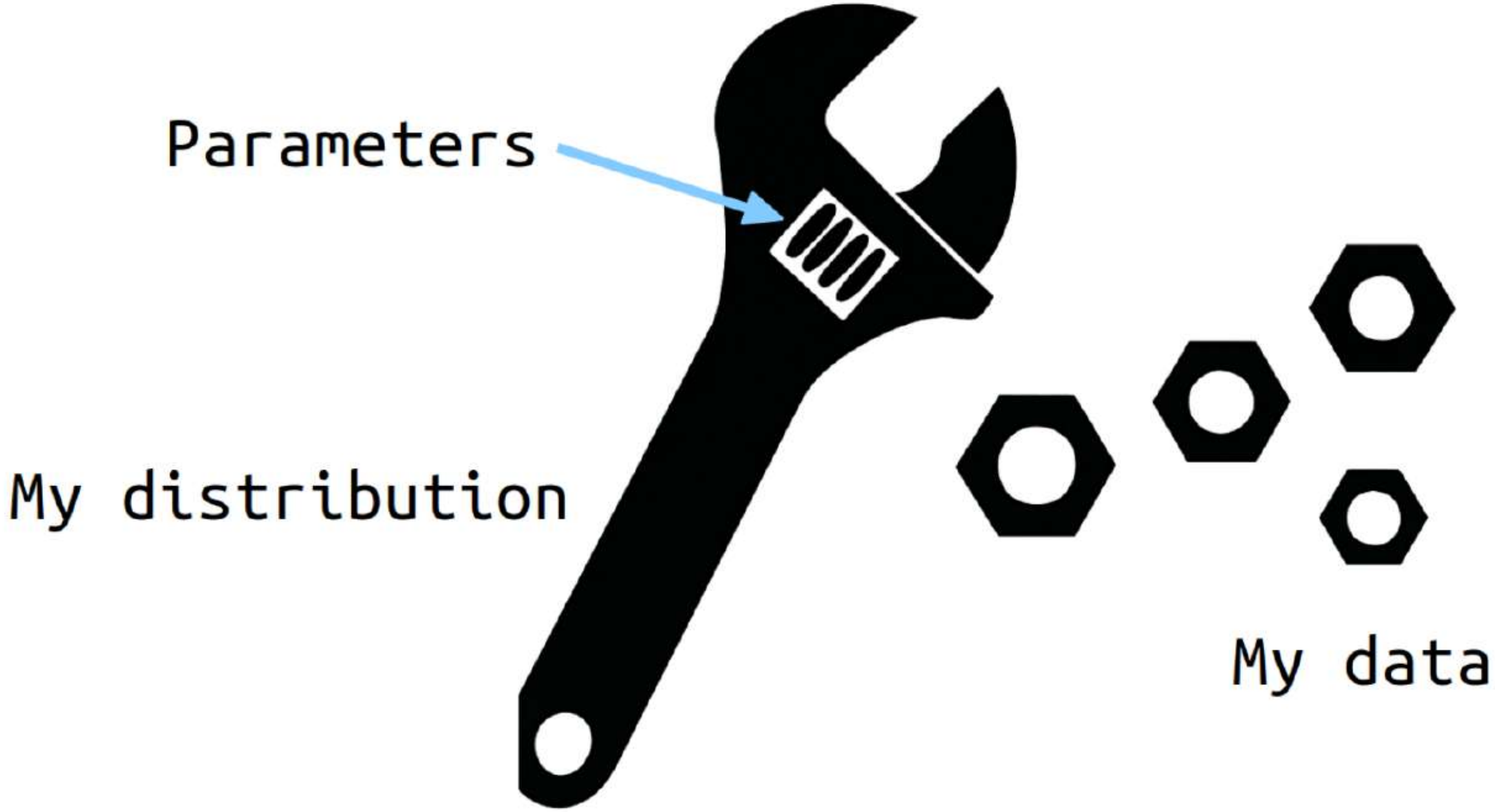
- Van a ser nuestros modelos, nuestra abstracción matemática, la representación simplificada de la realidad!

- Cada una tiene unas características diferentes que hacen que sean más aptas para uno u otro tipos de datos.

- Es necesario conocer las más importantes y sus principales características.

- Hay muchísimas, pero me gustaría que te aprendieses sólo tres (+1): la **Normal**, la de **Poisson** y la **Binomial** (+Uniforme).

Distribuciones de probabilidad



Distribuciones de probabilidad: notación

$$X \sim Name(parameter_1, parameter_2)$$

Distribuciones de probabilidad: notación

$$\boxed{X} \sim Name(parameter_1, parameter_2)$$

Mi variable

Distribuciones de probabilidad: notación

"se distribuye como"

$$\boxed{X} \sim Name(parameter_1, parameter_2)$$

Mi variable

Distribuciones de probabilidad: notación

"se distribuye como"

$$\boxed{X} \sim \boxed{Name}(parameter_1, parameter_2)$$

Mi variable

Nombre de la distribución

Distribuciones de probabilidad: notación

“se distribuye como”

$$\boxed{X} \sim \boxed{Name}(\underline{parameter_1, parameter_2})$$

Mi variable

Nombre de la distribución

Parámetros de la distribución

Distribuciones de probabilidad: notación

OJO: la misma distribución puede parametrizarse de diferentes formas

"se distribuye como"

$$\boxed{X} \sim \boxed{Name}(\underline{parameter_1, parameter_2})$$

Mi variable

Nombre de la distribución

Parámetros de la distribución

Distribuciones de probabilidad: notación

Temperatura media anual ~ Normal(media = 15, desviación estándar = 5)

Distribuciones de probabilidad: notación

Temperatura media anual ~ Normal(media = 15, desviación estándar = 5)

Temperatura media anual ~ Normal(μ = 15, σ = 5)

Temperatura media anual ~ N(15, 5)

Distribuciones de probabilidad: notación

Temperatura media anual ~ Normal(media = 15, desviación estándar = 5)

Temperatura media anual ~ Normal(μ = 15, σ = 5)

Temperatura media anual ~ N(15, 5)

Standard deviation

$$X \sim \mathcal{N}(\mu, \sigma)$$

Distribuciones de probabilidad: notación

Temperatura media anual ~ Normal(media = 15, desviación estándar = 5)

Temperatura media anual ~ Normal(μ = 15, σ = 5)

Temperatura media anual ~ N(15, 5)

Standard deviation

$$X \sim \mathcal{N}(\mu, \sigma)$$

Variance

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Distribuciones de probabilidad: notación

Temperatura media anual ~ Normal(media = 15, desviación estándar = 5)

Temperatura media anual ~ Normal(μ = 15, σ = 5)

Temperatura media anual ~ N(15, 5)

Standard deviation

$$X \sim \mathcal{N}(\mu, \sigma)$$

Variance

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Precision

$$X \sim \mathcal{N}(\mu, \tau)$$

$$\tau = \frac{1}{\sigma^2}$$

Distribuciones de probabilidad: notación

Temperatura media anual ~ Normal(media = 15, desviación estándar = 5)

Temperatura media anual ~ Normal(μ = 15, σ = 5)

Temperatura media anual ~ N(15, 5)

Standard deviation

$$\boxed{X \sim \mathcal{N}(\mu, \sigma)}$$

Variance

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Precision

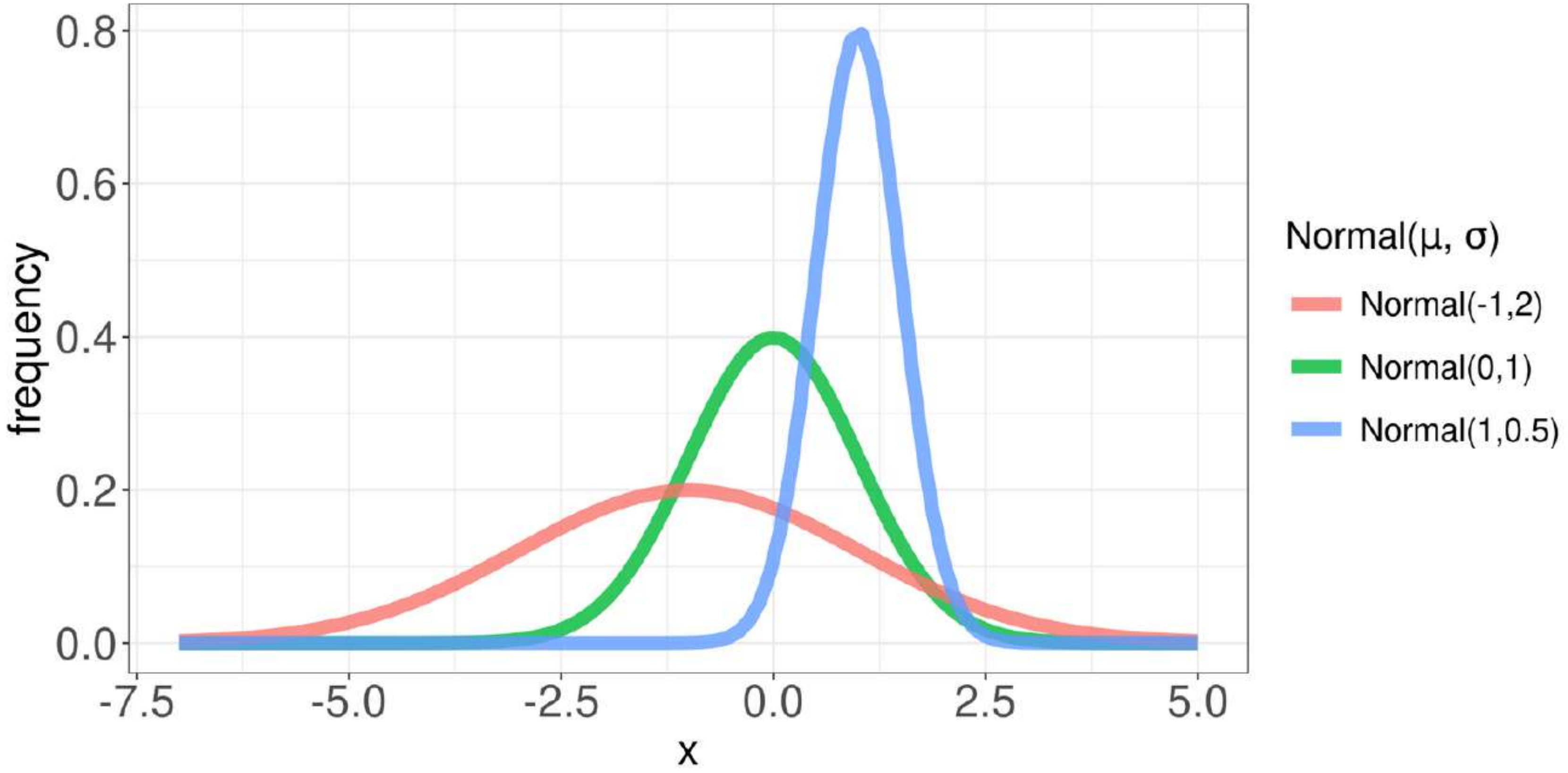
$$X \sim \mathcal{N}(\mu, \tau)$$

$$\tau = \frac{1}{\sigma^2}$$

Continuas

$$X \sim Normal(\mu, \sigma)$$

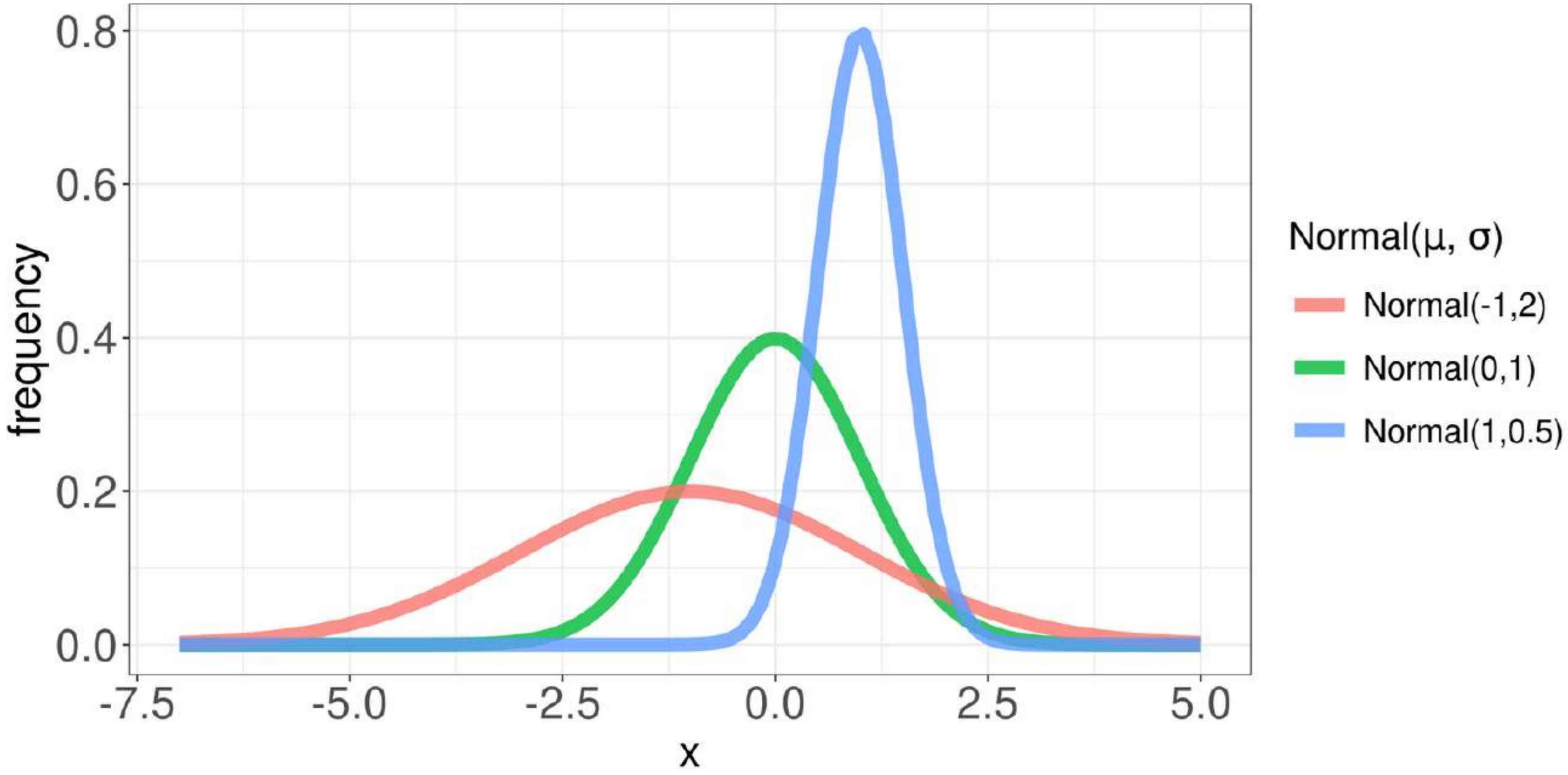
Support: $X \in (-\infty, \infty); \mu \in (-\infty, \infty); \sigma > 0$ (real)



Continuas

$$X \sim Normal(\mu, \sigma)$$

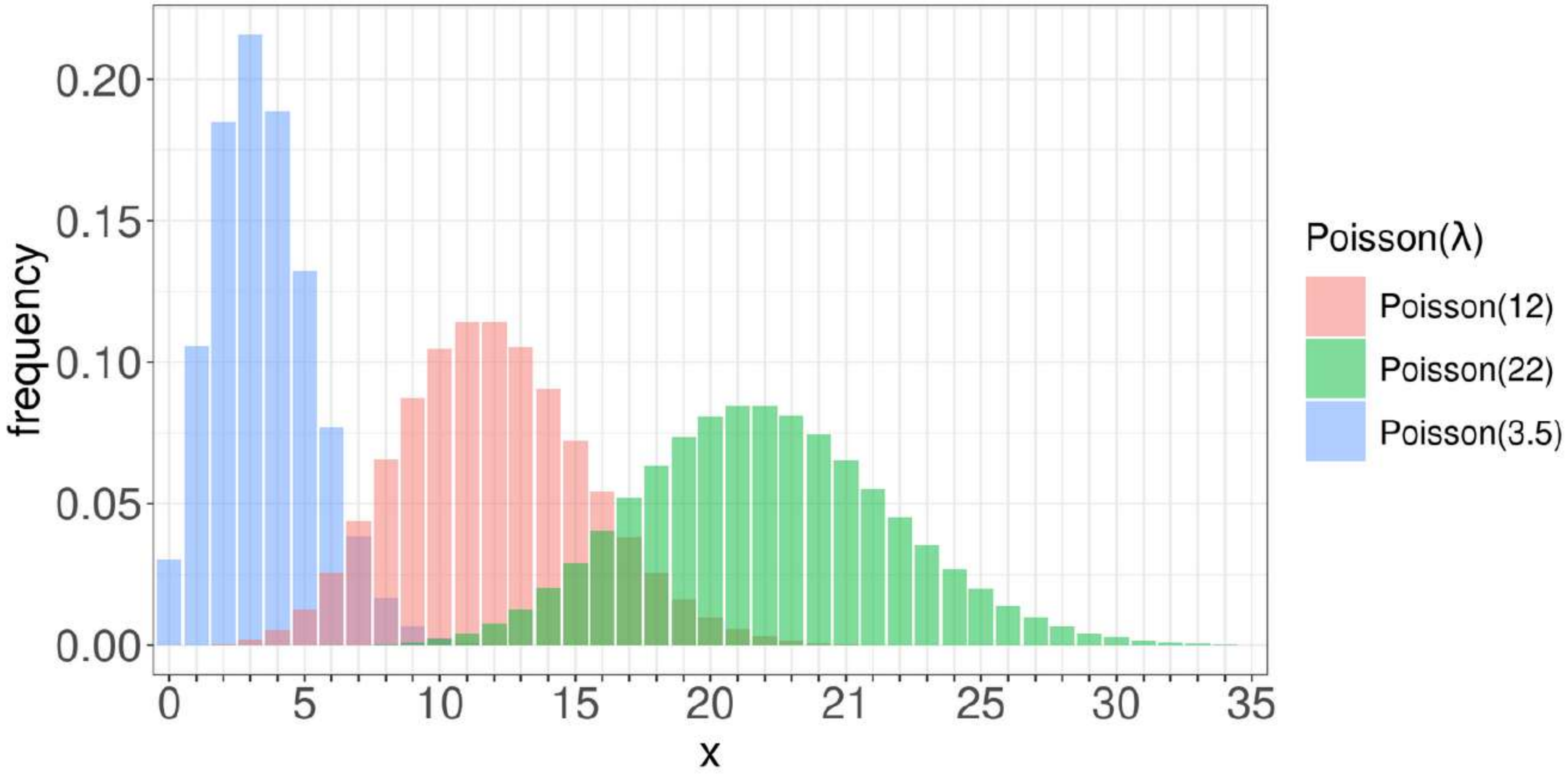
Support: $X \in (-\infty, \infty); \mu \in (-\infty, \infty); \sigma > 0$ (real)



$$X \sim Poisson(\lambda)$$

Conteos

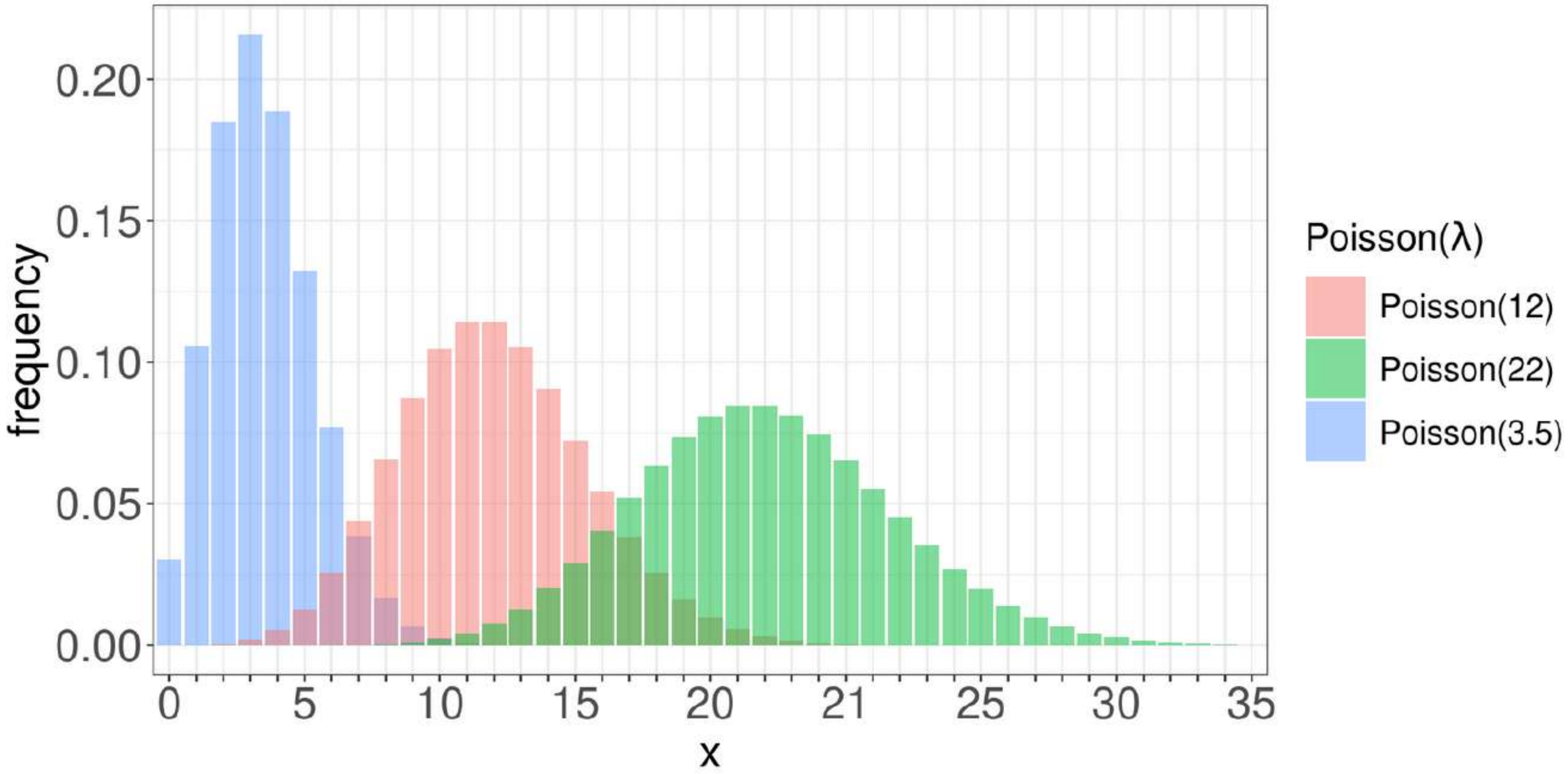
Support: $X \in (0, \infty)$ (natural); $\lambda \in (0, \infty)$ (real)



$$X \sim Poisson(\lambda)$$

Conteos

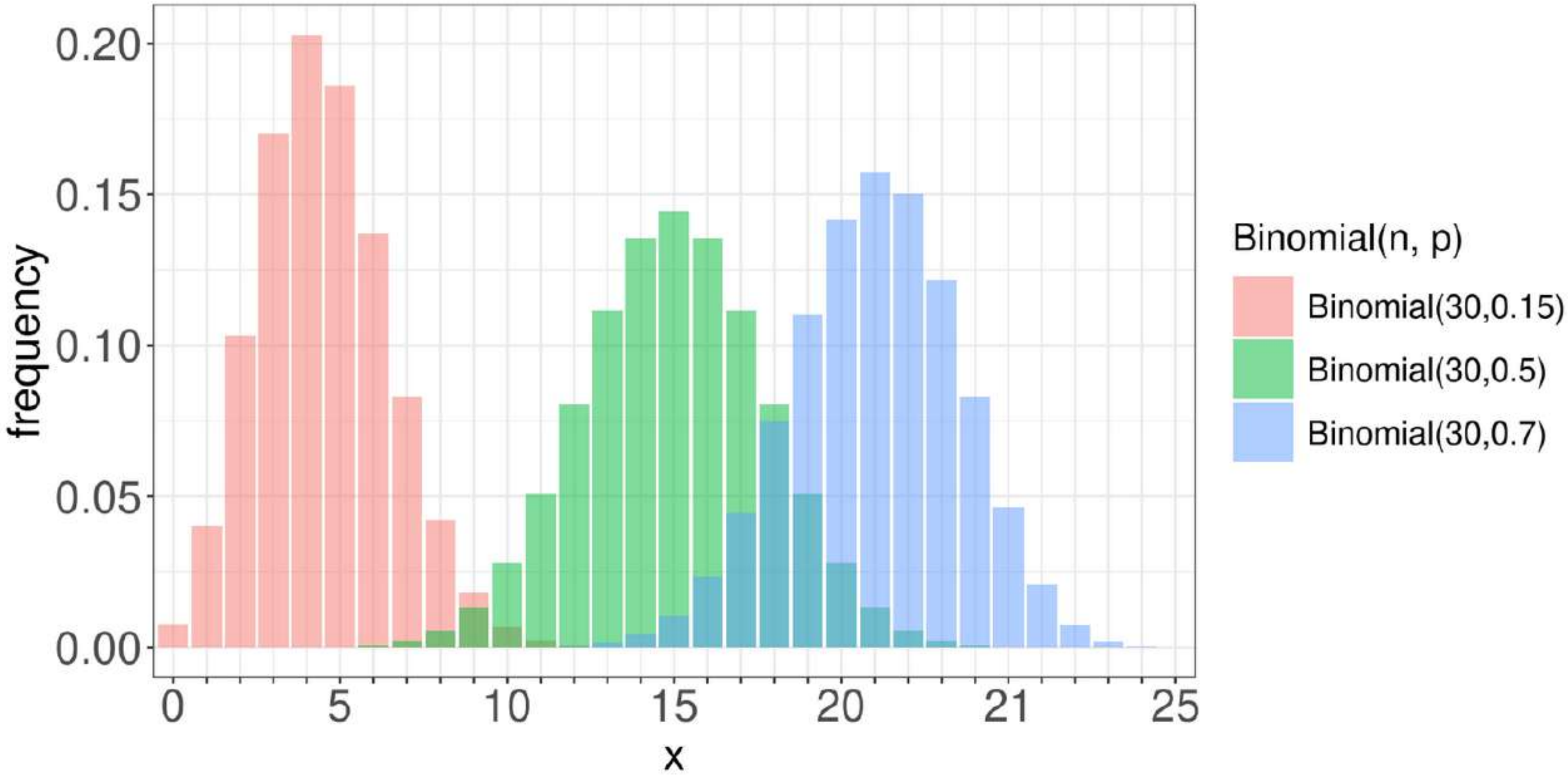
Support: $X \in (0, \infty)$ (natural); $\lambda \in (0, \infty)$ (real)



$$X \sim Binomial(n, p)$$

Presencia/ausencia

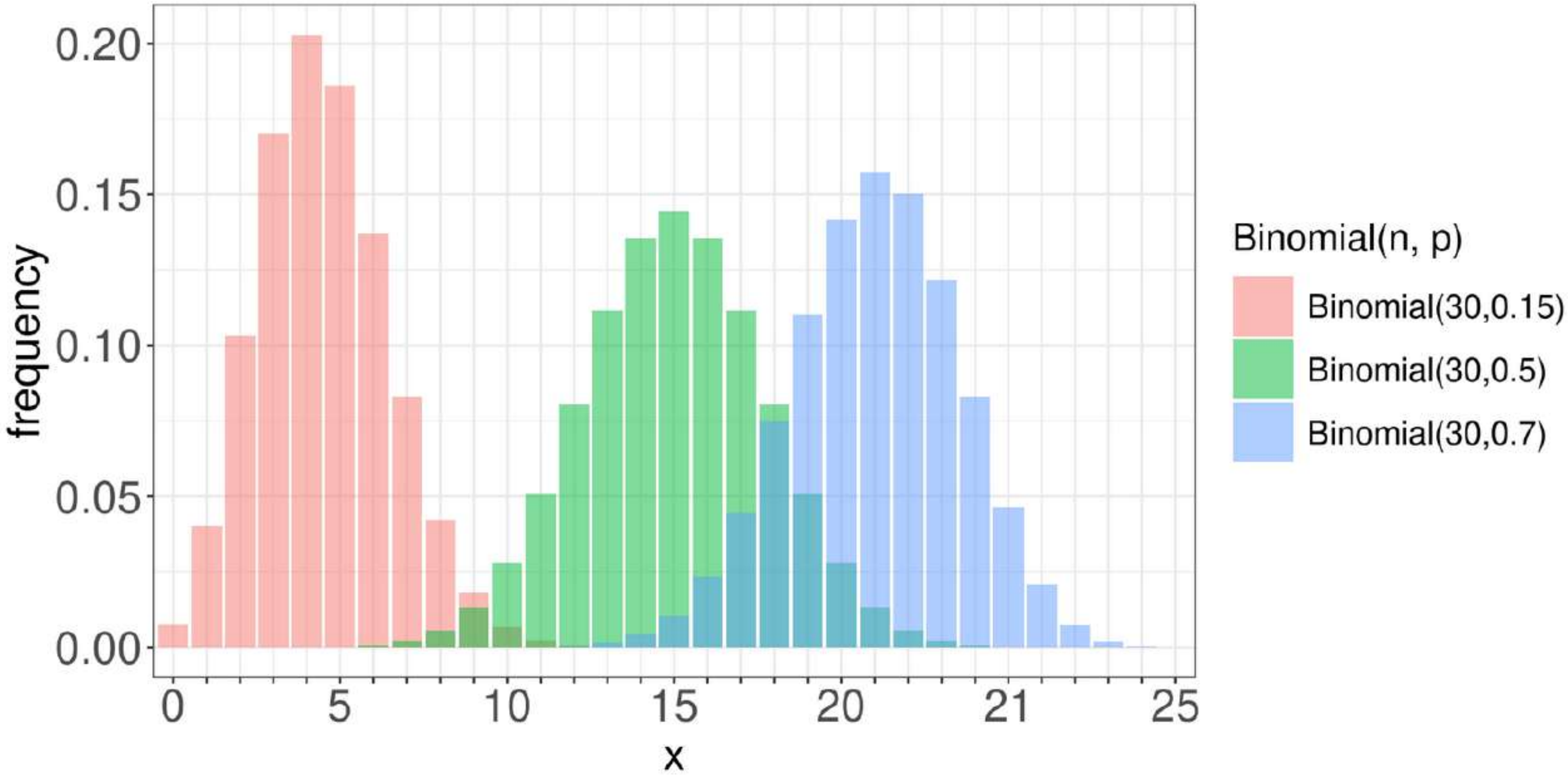
Support: $X = \{0, 1\}$;$n \in (0, \infty)$ (natural) $p \in (0, 1)$



$$X \sim Binomial(n, p)$$

Presencia/ausencia

Support: $X = \{0, 1\}$;$n \in (0, \infty)$ (natural) $p \in (0, 1)$

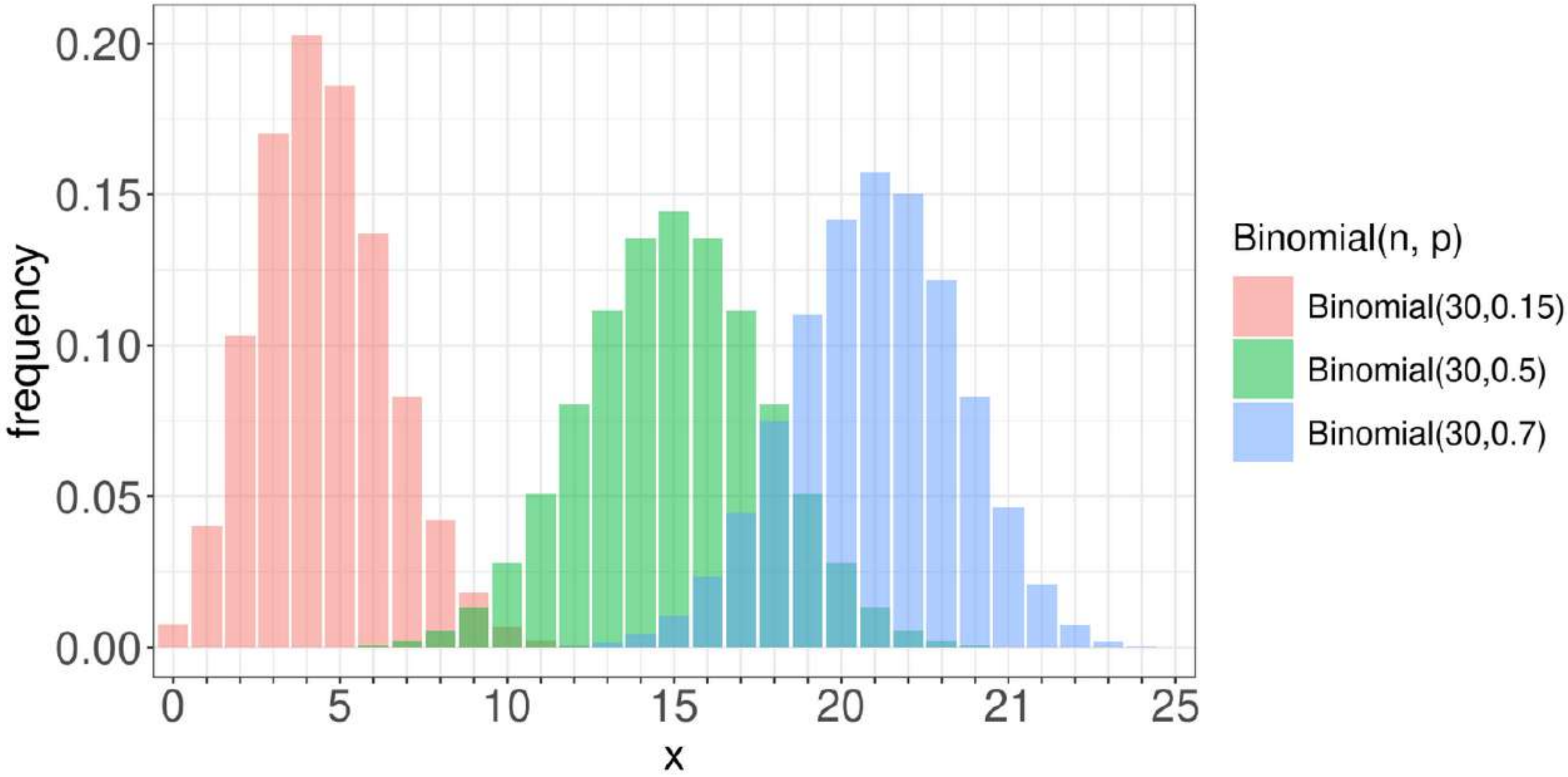


$$X \sim Binomial(n,p)$$

Presencia/ausencia

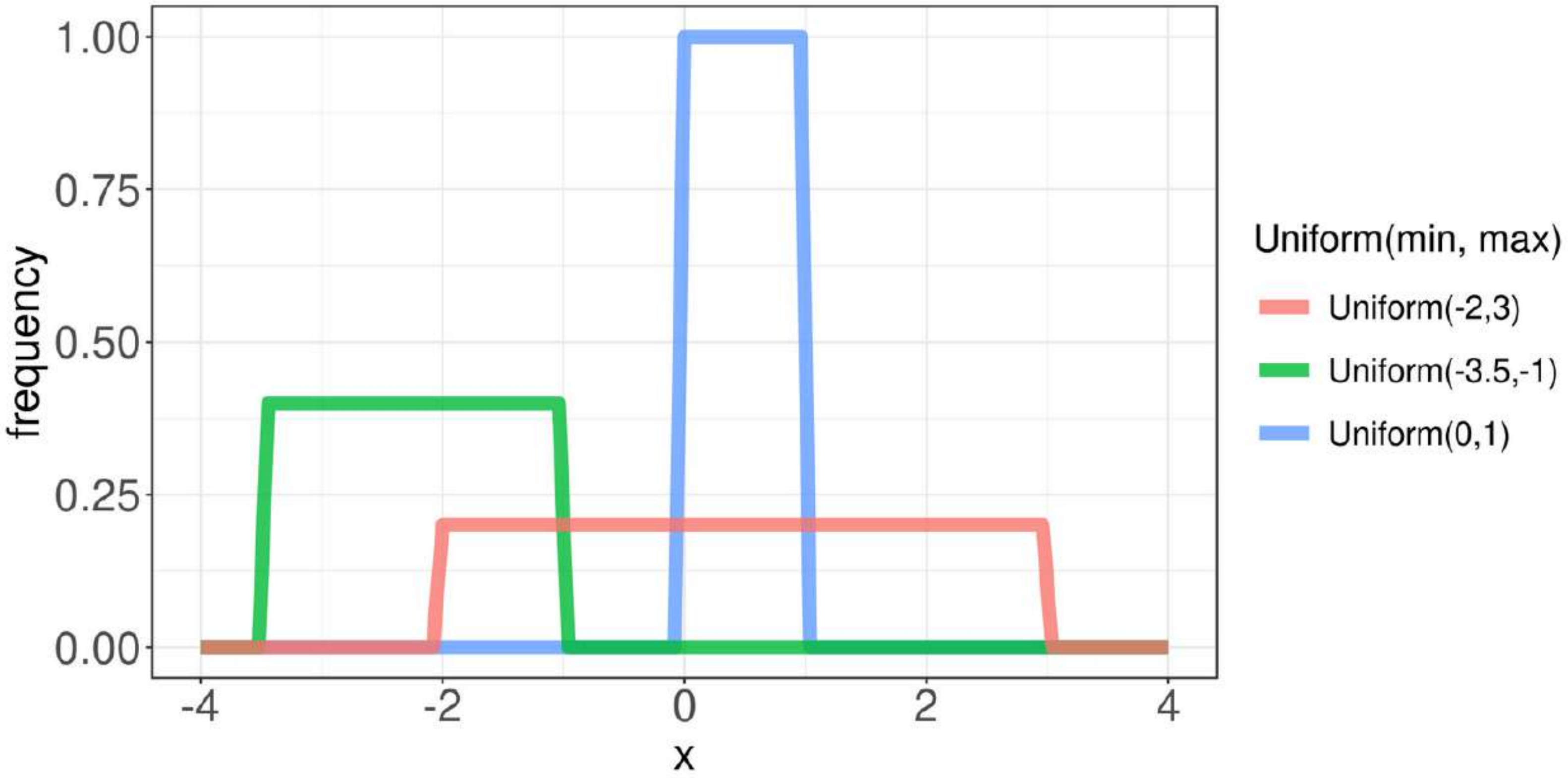
Support: $X = \{0,1\}$;$n \in (0,\infty)$ (natural) $p \in (0,1)$

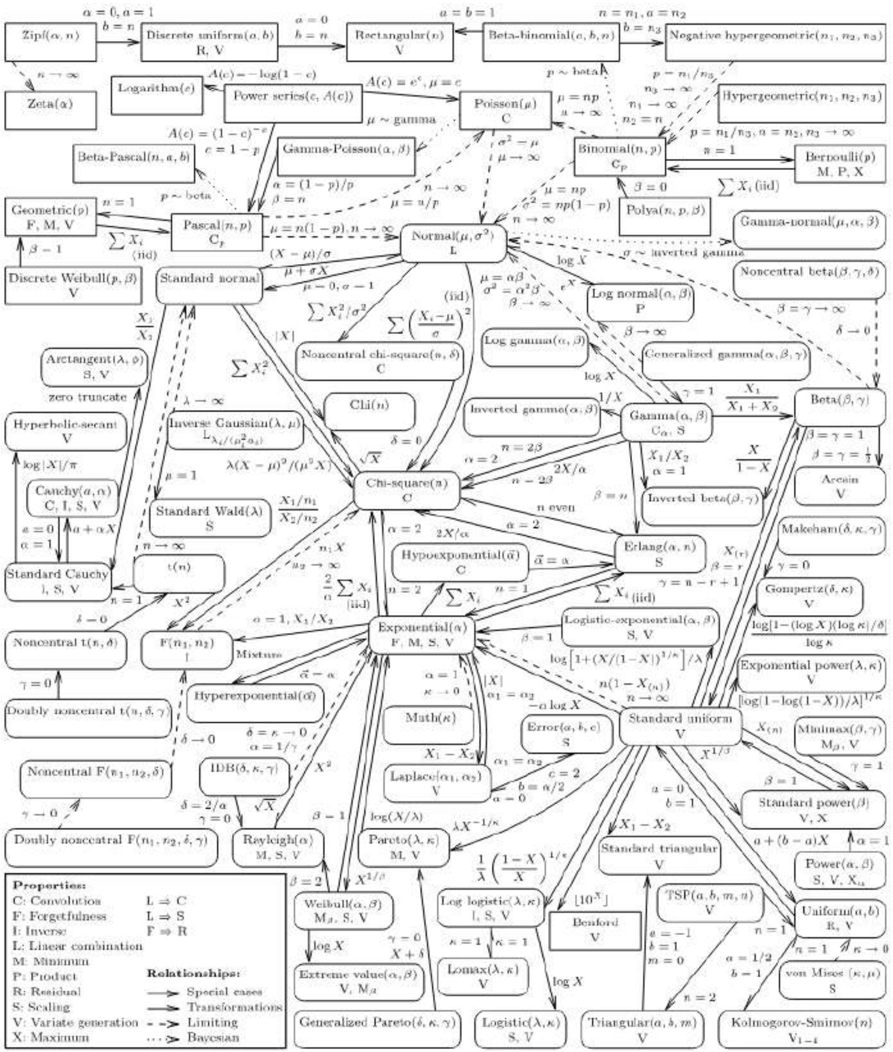
Bernoulli = Binomial(1,p)



$$X \sim Uniform(min, max)$$

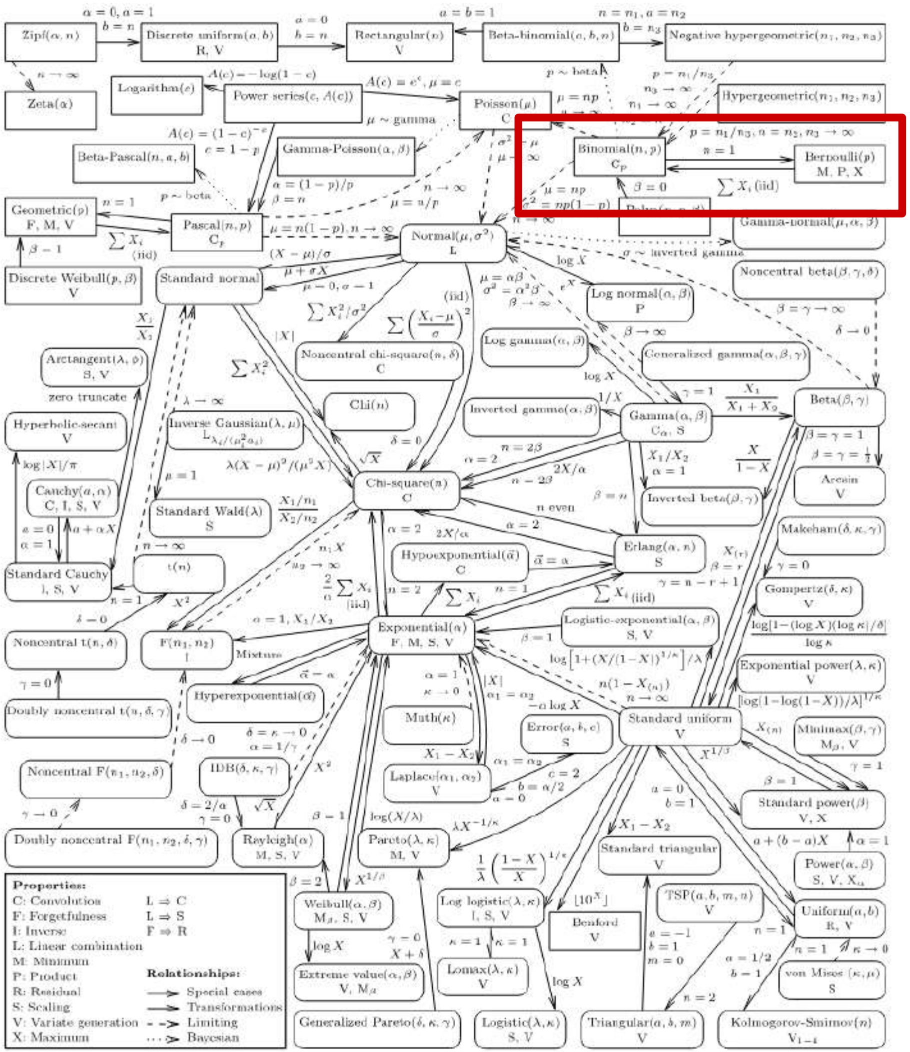
Support: $X \in (-\infty, \infty); min \in (-\infty, \infty)$ (real); $max \in (-\infty, \infty)$ (real)



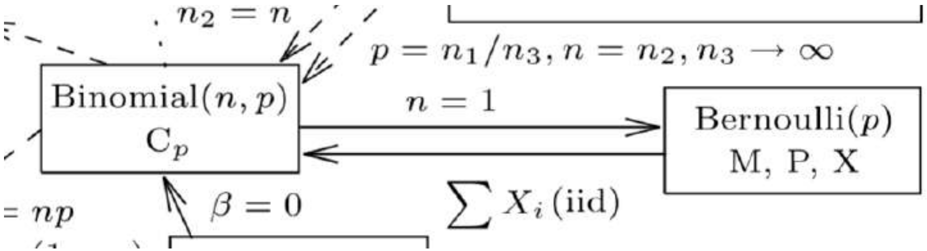


Relación entre distribuciones de probabilidad univariadas

Leemis & McQueston, American Satistician (2008)

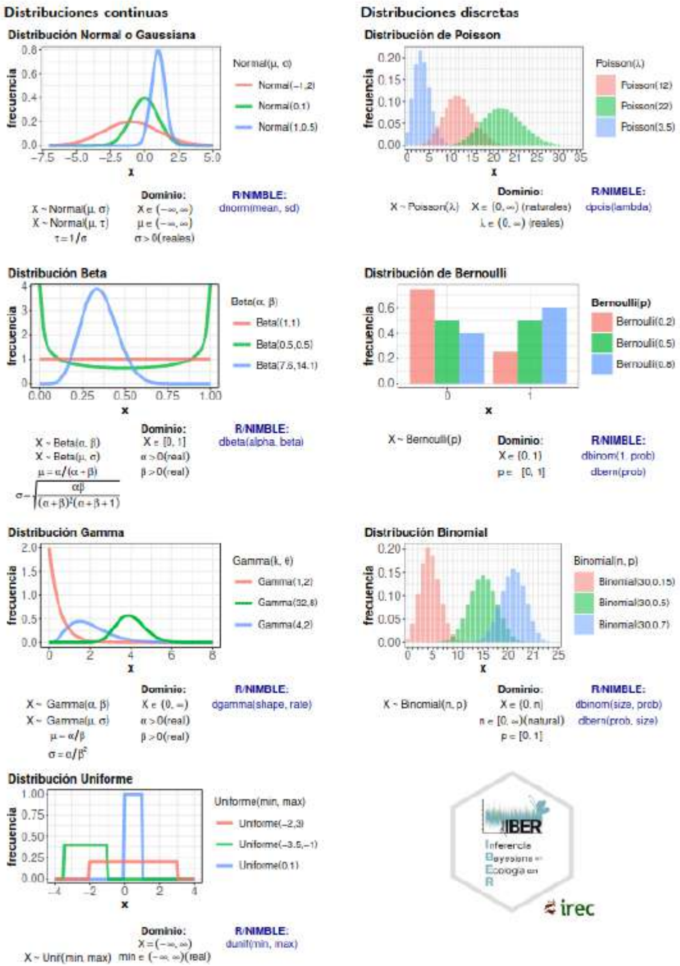


Relación entre distribuciones de probabilidad univariadas



Leemis & McQueston, American Satistician (2008)

VARIABLES Y DISTRIBUCIONES DE PROBABILIDAD



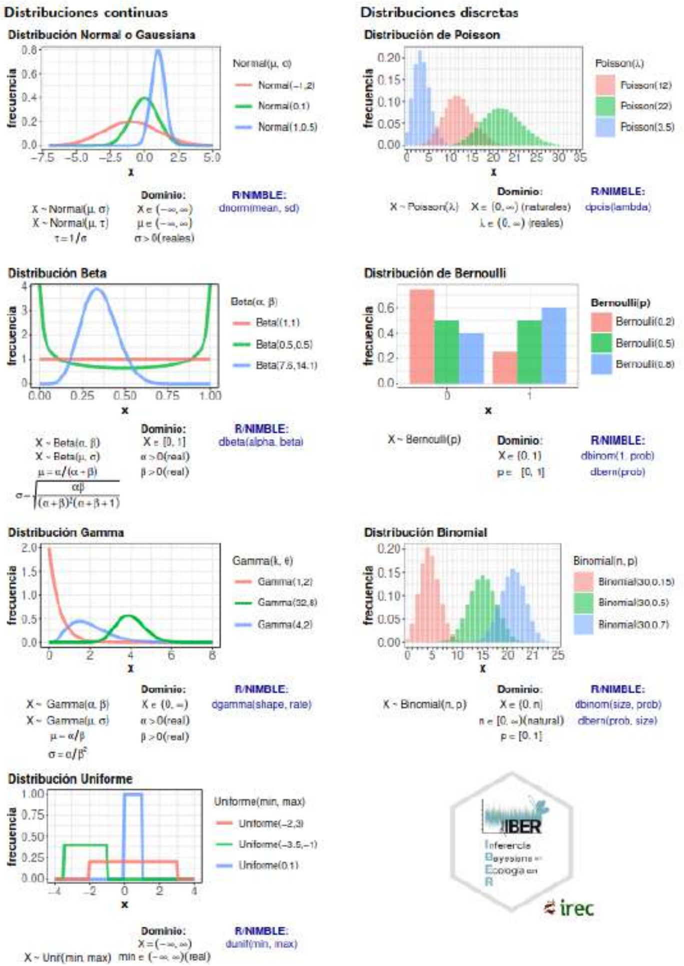
Principales funciones vínculo para GLM

Las funciones vínculo (*link functions*) sirven para enlazar los parámetros de las distribuciones de probabilidad con la ecuación general del modelo lineal, convirtiendo el modelo lineal *general* en *generalizado*. Los parámetros de algunas distribuciones no pueden tomar cualquier valor, por lo que se utilizan las funciones vínculo para adaptar los resultados de la ecuación del modelo lineal a los requerimientos del parámetro. En teoría (matemáticamente) es posible emplear cualquier función vínculo con cualquier distribución siempre que nuestros datos lo permitan. Sin embargo, en la práctica se suelen utilizar unas pocas funciones con cada una de las distribuciones de probabilidad. En la tabla se muestran las más comunes para las distribuciones estudiadas. Nótese que algunas distribuciones no se muestran con su parametrización más habitual.

Distribución	Links	Fórmula	Inversa
Gaussiana(μ, σ)	**Identidad**	$\mu = \beta X$	$\mu = \beta X$
Beta(μ, σ)	**Logit**, probit, cloglog	$log(\frac{\mu}{1-\mu}) = \beta X$	$\mu = \frac{e^{(\beta X)}}{1+e^{(\beta X)}}$
Gamma(μ, σ)	**Log**	$log(\mu) = \beta X$	$\mu = e^{(\beta X)}$
Poisson(λ)	**Log**	$log(\lambda) = \beta X$	$\lambda = e^{(\beta X)}$
Bernoulli(p)	**Logit**, probit, cloglog	$log(\frac{p}{1-p}) = \beta X$	$p = \frac{e^{(\beta X)}}{1+e^{(\beta X)}}$
Binomial(n,p)	**Logit**, probit, cloglog	$log(\frac{p}{1-p}) = \beta X$	$p = \frac{e^{(\beta X)}}{1+e^{(\beta X)}}$

https://jabiologo.github.io/web/tutorials/prob_dist_links_merged.pdf

VARIABLES Y DISTRIBUCIONES DE PROBABILIDAD



¿Con qué distribuciones de probabilidad habéis trabajado?

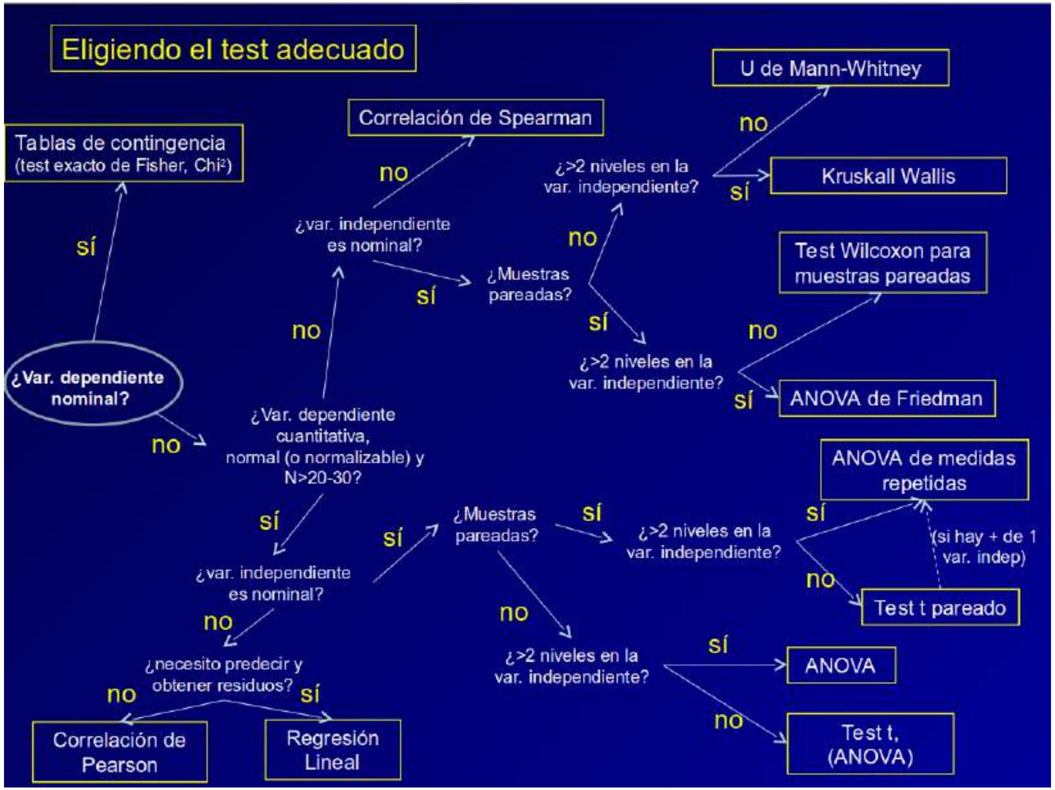
¿Qué otras distribuciones de probabilidad conocéis?

Matt Bognar website:

https://homepage.divms.uiowa.edu/~mbognar/

https://jabiologo.github.io/web/tutorials/prob_dist_links_merged.pdf

Las simulaciones en el proceso de aprendizaje



Las simulaciones en el proceso de aprendizaje

REALIDAD

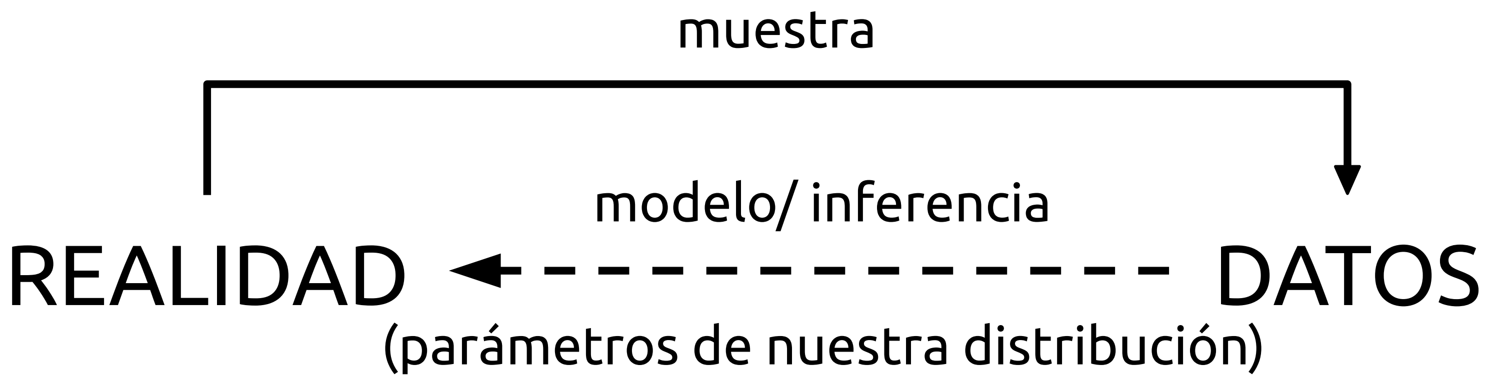
Las simulaciones en el proceso de aprendizaje

muestra

REALIDAD → DATOS

LA SIMULACIÓN COMO UN LABORATORIO

Las simulaciones en el proceso de aprendizaje



Las simulaciones en el proceso de aprendizaje

muestra

modelo/ inferencia

REALIDAD ← - - - - - - ? - - - - - - - DATOS

(parámetros de nuestra distribución)

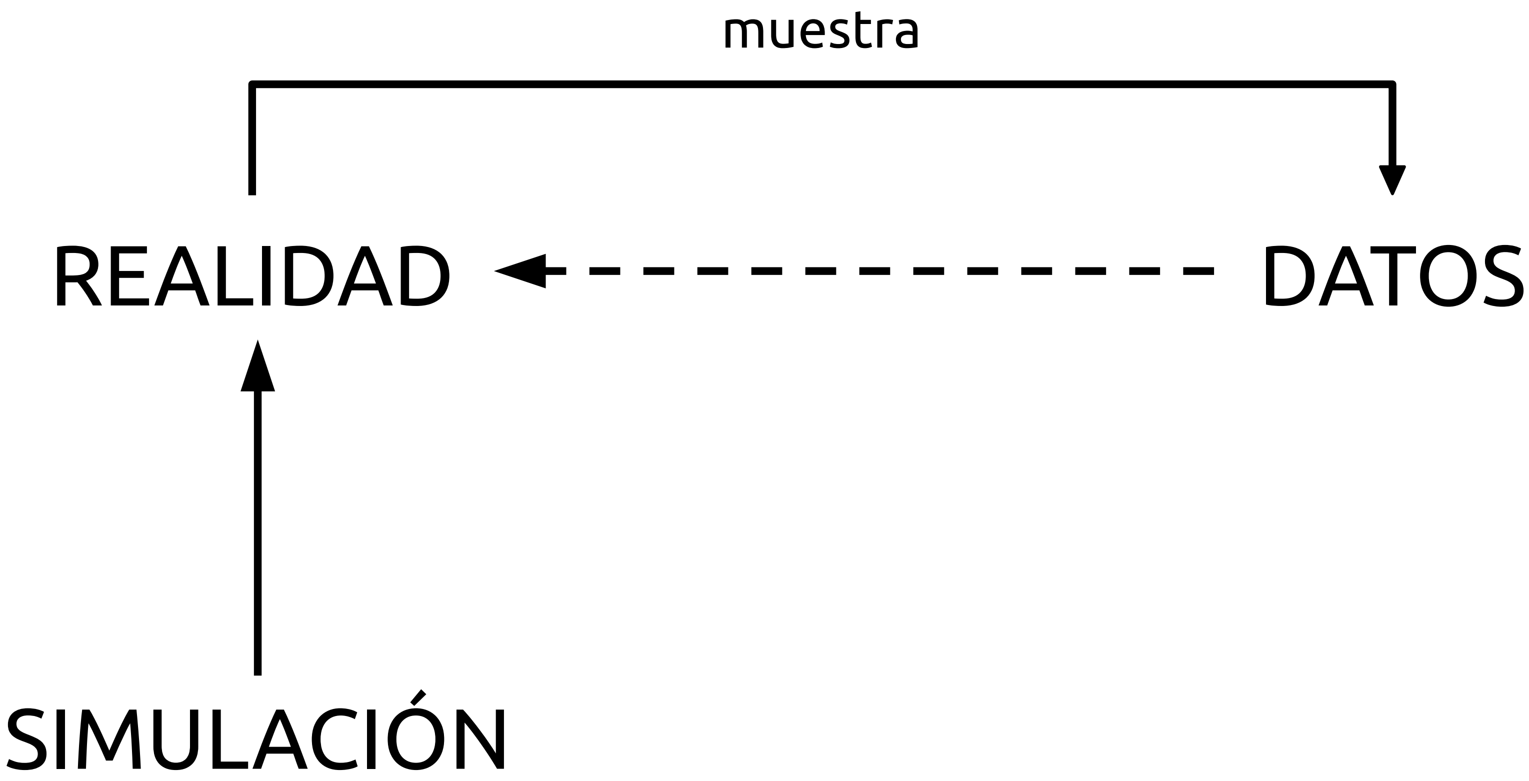
Las simulaciones en el proceso de aprendizaje

REALIDAD

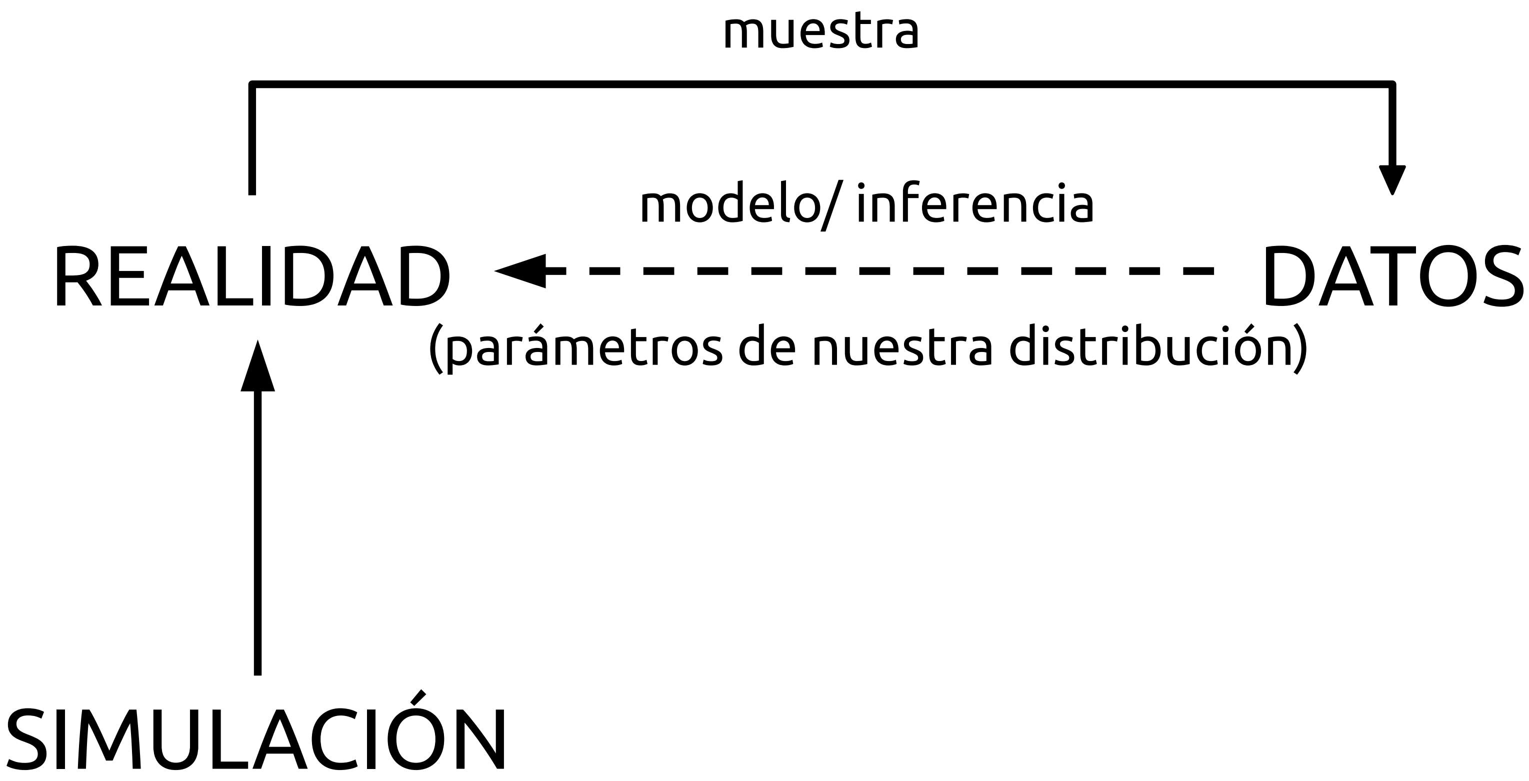
SIMULACIÓN

Las simulaciones en el proceso de aprendizaje

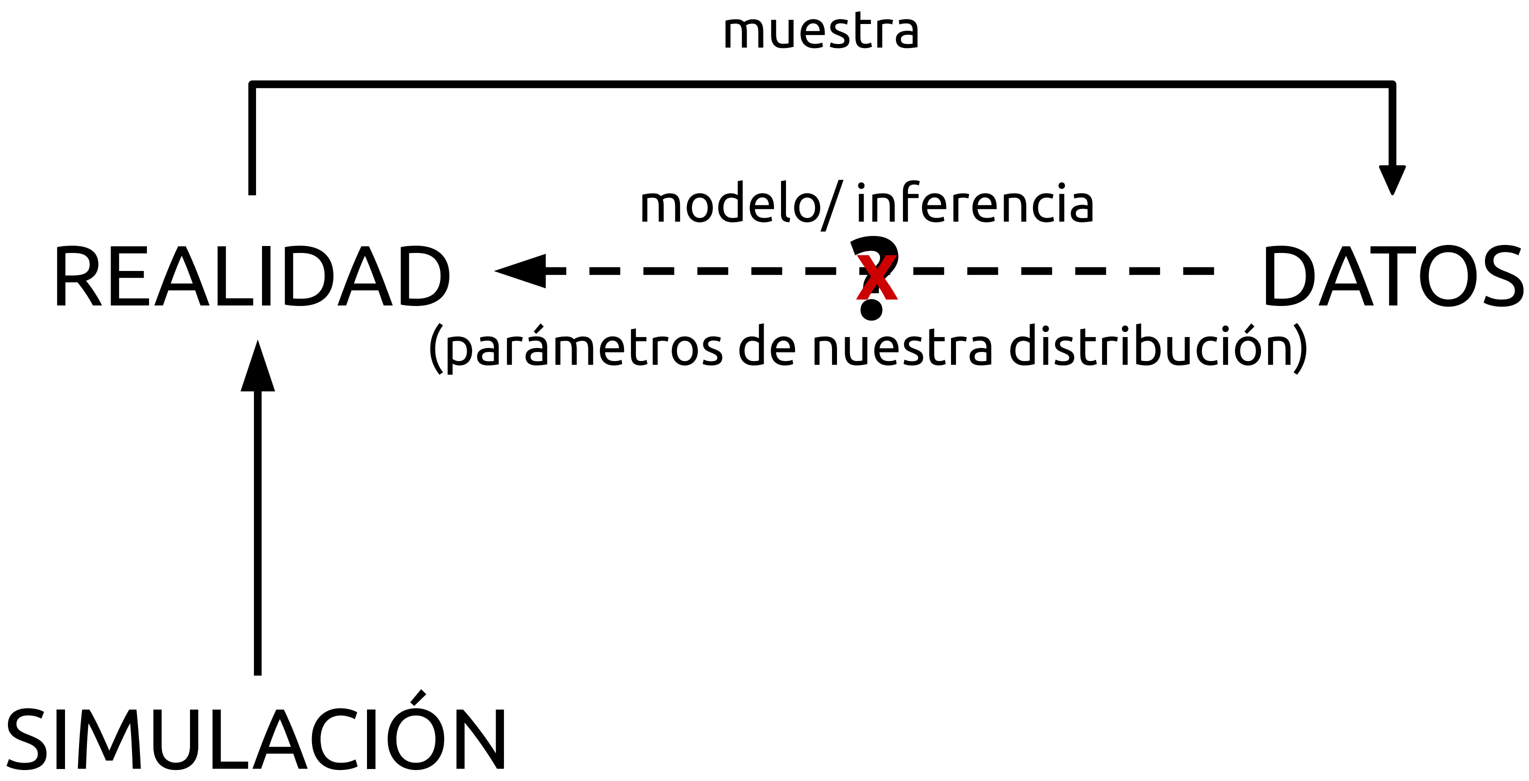
Las simulaciones en el proceso de aprendizaje



Las simulaciones en el proceso de aprendizaje



Las simulaciones en el proceso de aprendizaje



Las simulaciones en el proceso de aprendizaje

muestra

modelo/ inferencia

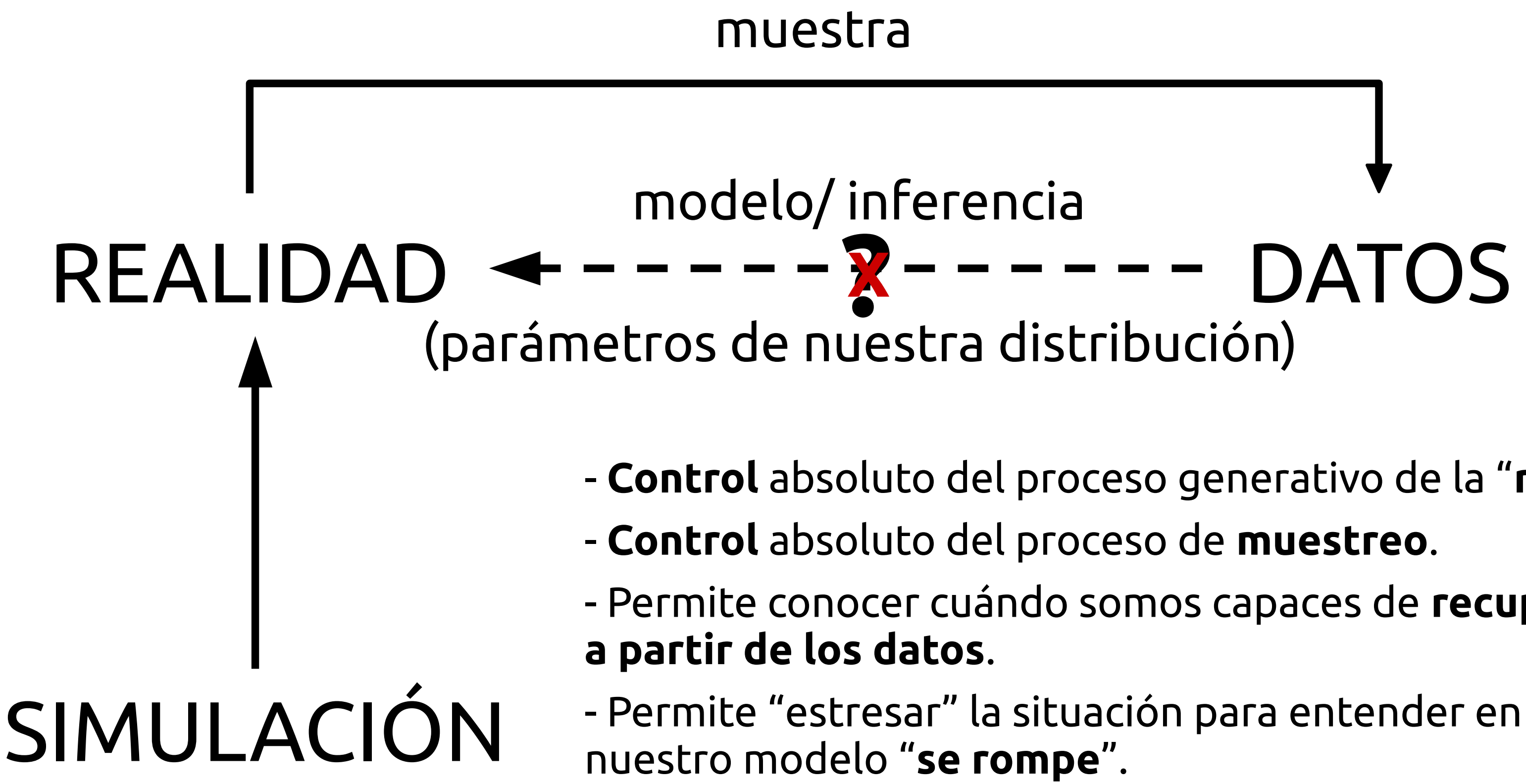
REALIDAD ← - - - - - - ? - - - - - - - DATOS

(parámetros de nuestra distribución)

SIMULACIÓN

- **Control** absoluto del proceso generativo de la “**realidad**”.
- **Control** absoluto del proceso de **muestreo**.

Las simulaciones en el proceso de aprendizaje



- **Control** absoluto del proceso generativo de la “**realidad**”.
- **Control** absoluto del proceso de **muestreo**.
- Permite conocer cuándo somos capaces de **recuperar la realidad a partir de los datos**.
- Permite “estresar” la situación para entender en qué condiciones nuestro modelo “**se rompe**”.

Las simulaciones en el proceso de aprendizaje

muestra

modelo/ inferencia

REALIDAD ← - - - - - ? - - - - - - DATOS

(parámetros de nuestra distribución)

SIMULACIÓN

- **Control** absoluto del proceso generativo de la "**realidad**".
- **Control** absoluto del proceso de **muestreo**.
- Permite conocer cuándo somos capaces de **recuperar la realidad a partir de los datos**.
- Permite "estresar" la situación para entender en qué condiciones nuestro modelo "**se rompe**".
- **Reto**: ser consciente del grado de semejanza entre simulación y realidad (ser honestos, cura de humildad).

Las simulaciones en el proceso de aprendizaje

muestra

modelo/ inferencia

REALIDAD ←- - - - - - ? - - - - - - DATOS

(parámetros de nuestra distribución)

SIMULACIÓN

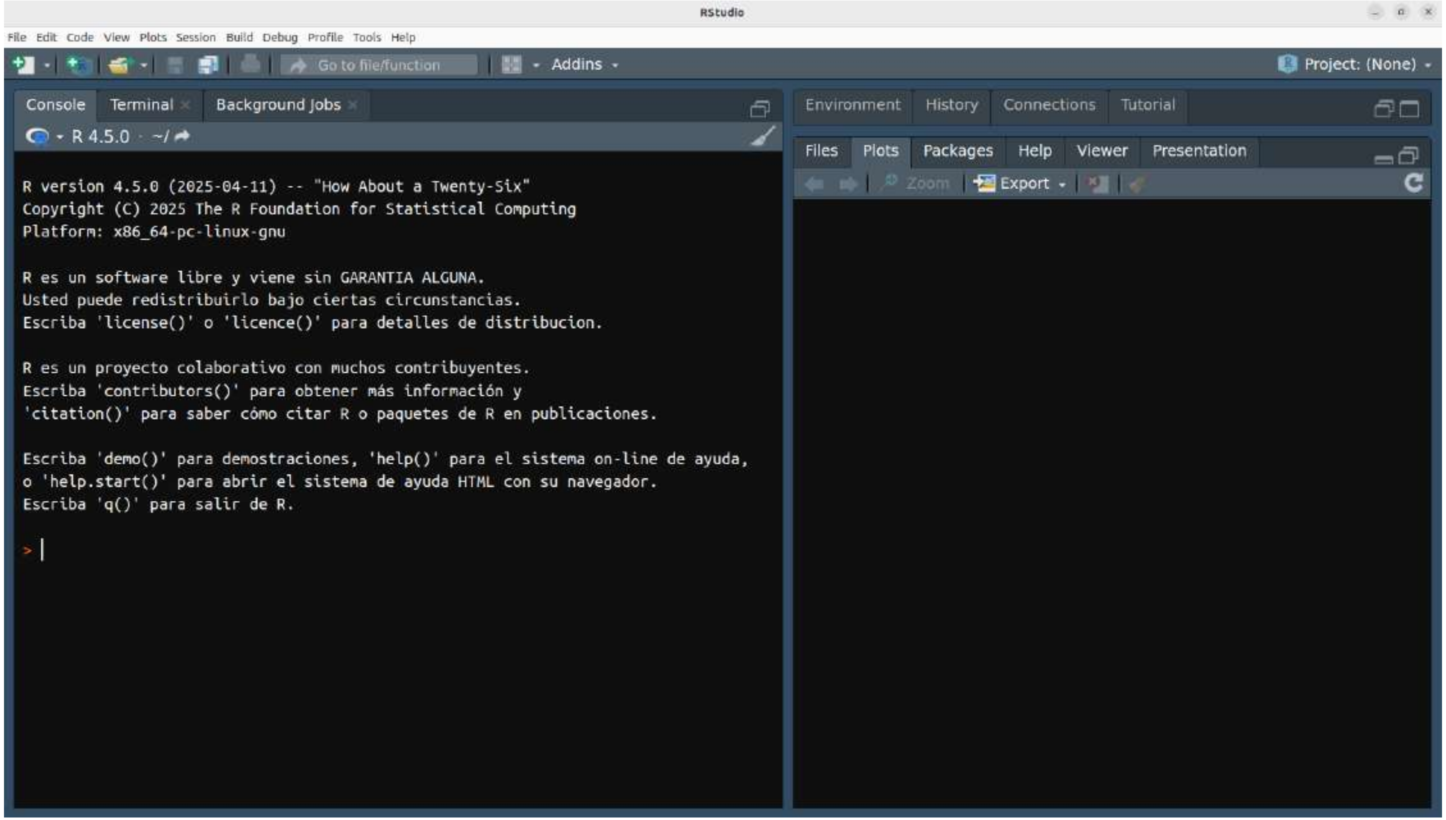
"If you can simulate it, you understand it."

— **Andrew Gelman**

"To understand a model, you should be able to simulate data from it."

— **George Box**

LA SIMULACIÓN COMO UN LABORATORIO



Recap...

REALIDAD

Recap...

i	*peso*	*latitud*
1	27	40
2	35	51
3	42	59
...	...	...
n	50	65

REALIDAD $\xrightarrow{\text{muestra}}$ DATOS

Recap...

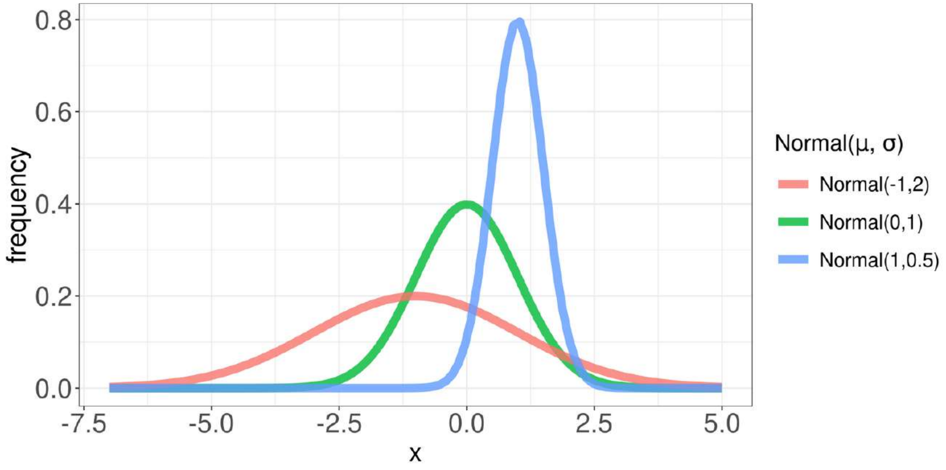
i	*peso*	*latitud*
1	27	40
2	35	51
3	42	59
...	...	...
n	50	65

muestra

modelo/ inferencia

REALIDAD ←- - - - - - - - - - - - - DATOS

(parámetros de nuestra distribución)



Recap...

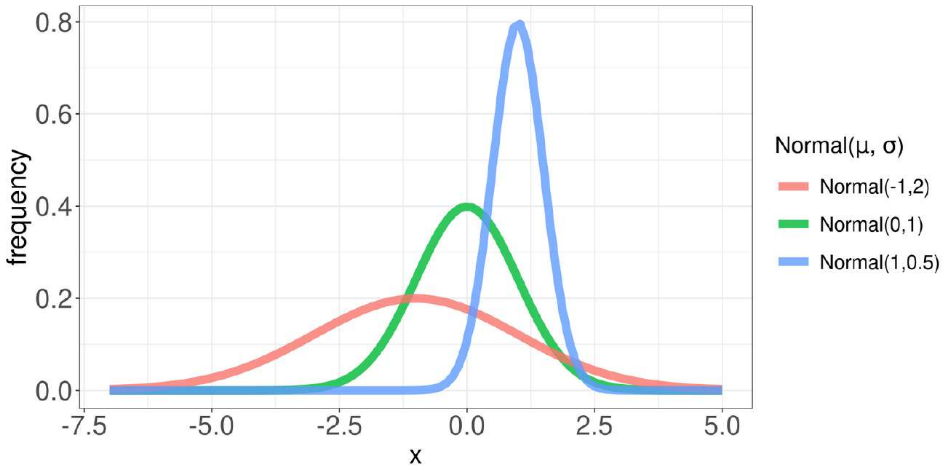
i	*peso*	*latitud*
1	27	40
2	35	51
3	42	59
...	...	...
n	50	65

muestra

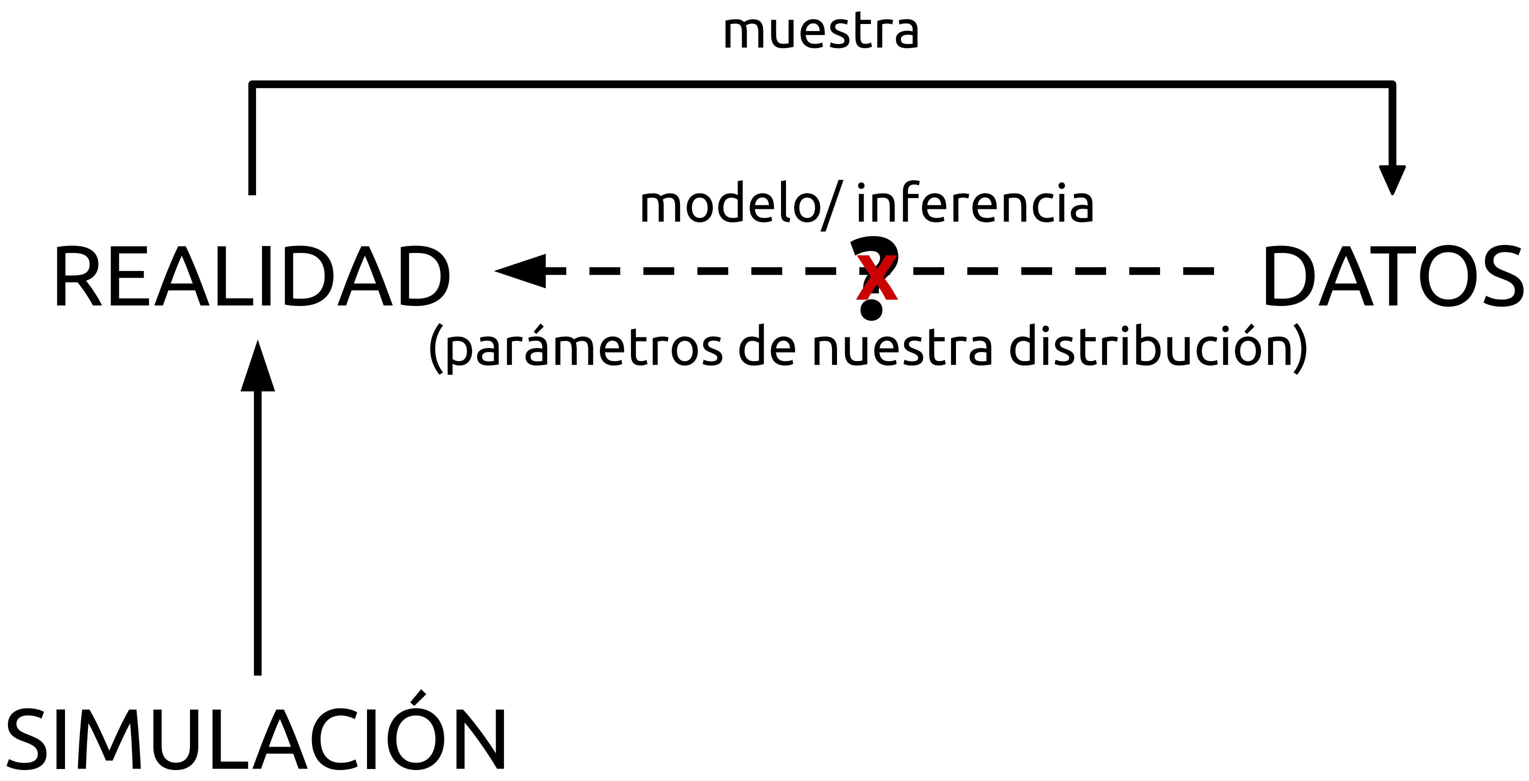
modelo/ inferencia

REALIDAD ← - - - - - - ? - - - - - - - DATOS

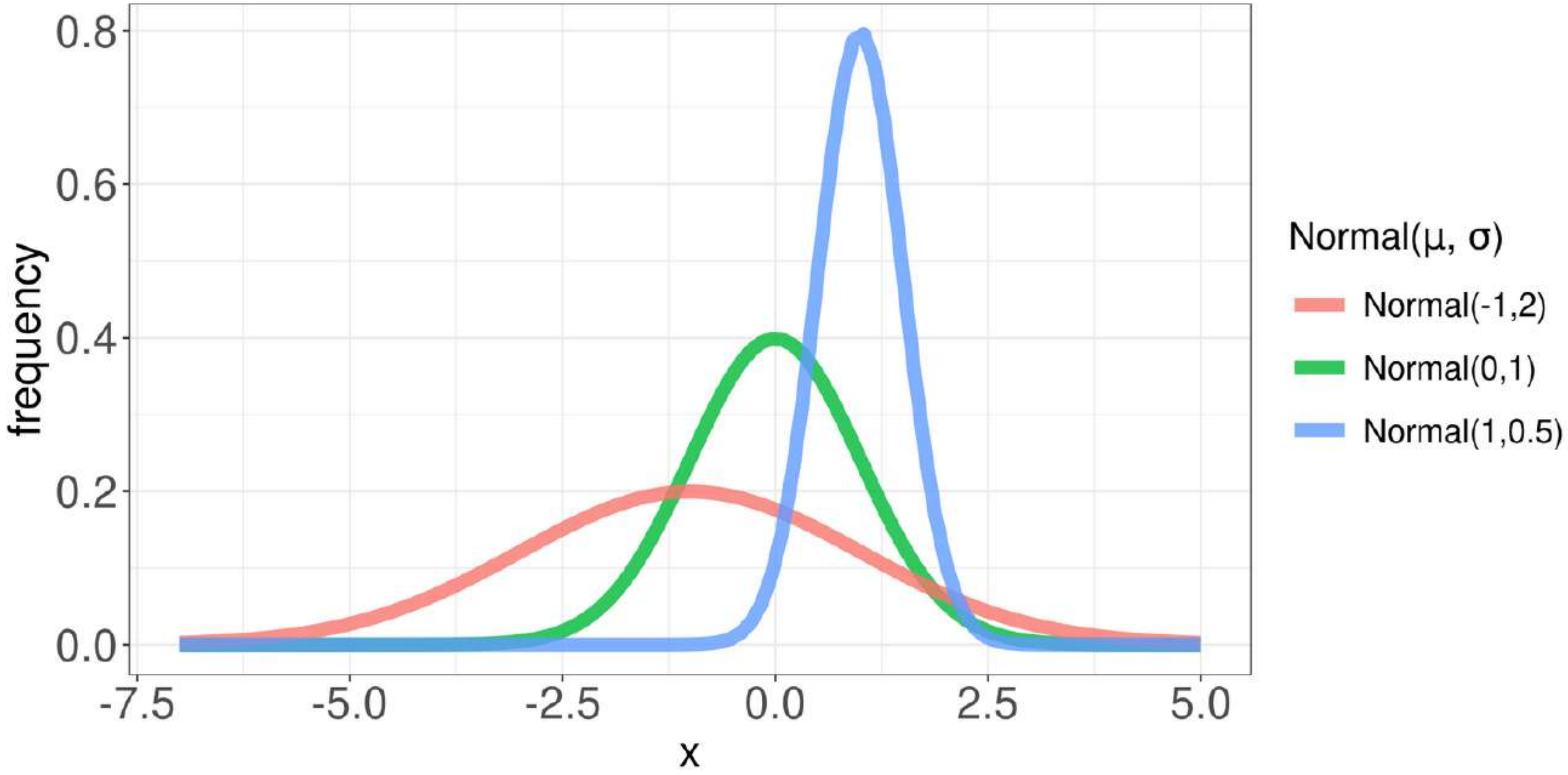
(parámetros de nuestra distribución)



Recap...


muestra
modelo/ inferencia
REALIDAD
?
x
DATOS
(parámetros de nuestra distribución)
SIMULACIÓN

Nuestros modelos...



El modelo lineal (o cuando "la media se mueve")

i	*peso*
1	27
2	22
3	31
...	...
n	28

$$weight(kg) \sim Normal(mean = 27, sd = 3)$$

El modelo lineal (o cuando "la media se mueve")

i	*peso*
1	27
2	22
3	31
...	...
n	28

$$weight(kg) \sim Normal(mean = 27, sd = 3)$$

$$Y \sim Normal(\mu, \sigma)$$

$$Y \sim Normal(27, 3)$$

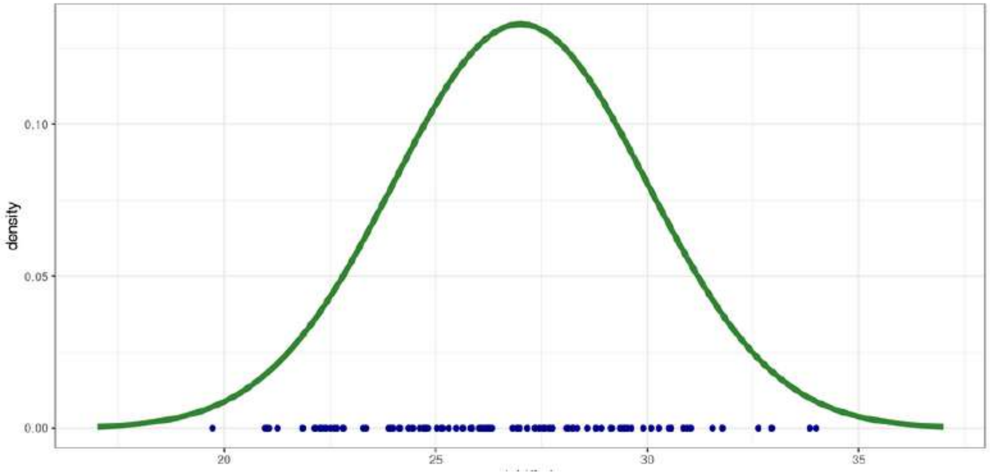
El modelo lineal (o cuando “la media se mueve”)

i	$peso$
1	27
2	22
3	31
...	...
n	28

$$weight(kg) \sim Normal(mean = 27, sd = 3)$$

$$Y \sim Normal(\mu, \sigma)$$

$$Y \sim Normal(27, 3)$$



El modelo lineal (o cuando "la media se mueve")

i	*peso*	*latitud*
1	27	40
2	35	51
3	42	59
...	...	...
n	50	65

$$weight_i \sim Normal(\mu_i, \sigma)$$

El modelo lineal (o cuando "la media se mueve")

i	*peso*	*latitud*
1	27	40
2	35	51
3	42	59
...	...	...
n	50	65

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 X_i$$

El modelo lineal (o cuando “la media se mueve”)

i	*peso*	*latitud*
1	27	40
2	35	51
3	42	59
...	...	...
n	50	65

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 X_i$$

$a + bX_i$

$\alpha + \beta X_i$

El modelo lineal (o cuando "la media se mueve")

i	*peso*	*latitud*
1	27	40
2	35	51
3	42	59
...	...	...
n	50	65

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 X_i$$

$$\mu_i = \beta_0 + \beta_1 Latitude_i$$

El modelo lineal (o cuando “la media se mueve”)

i	$peso$	$latitud$
1	27	40
2	35	51
3	42	59
...	...	...
n	50	65

Variable respuesta o dependiente

$$\boxed{weight_i} \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 X_i$$

$$\mu_i = \beta_0 + \beta_1 Latitude_i$$

El modelo lineal (o cuando "la media se mueve")

i	*peso*	*latitud*
1	27	40
2	35	51
3	42	59
...	...	...
n	50	65

Variable respuesta o dependiente

$$\boxed{weight_i} \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 X_i$$

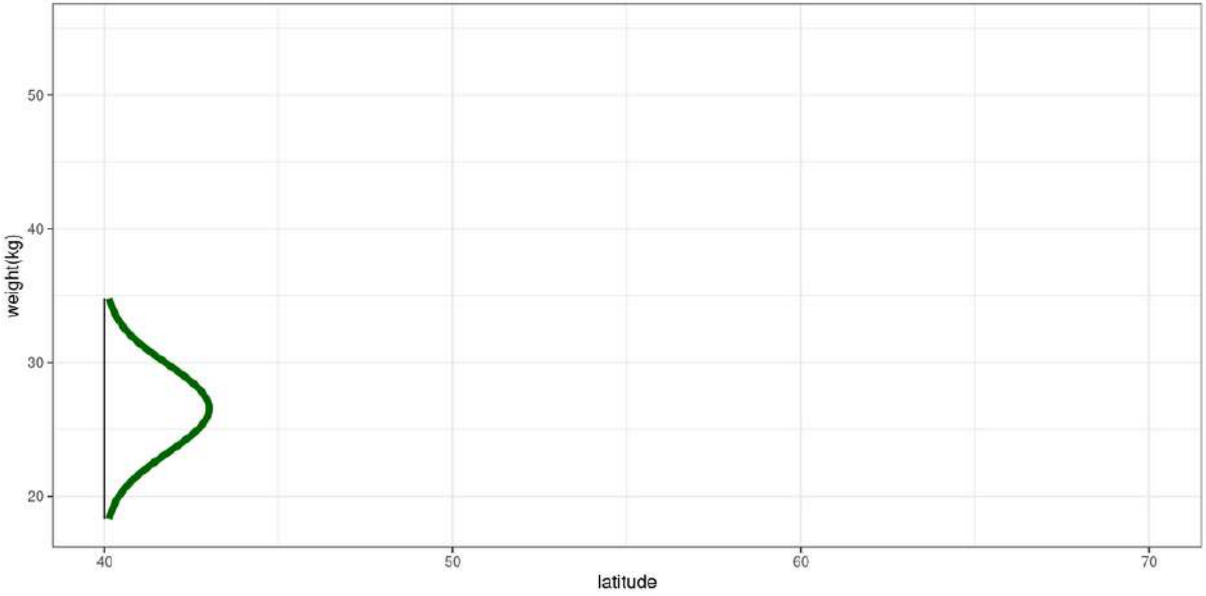
$$\mu_i = \beta_0 + \beta_1 \boxed{Latitude_i}$$

Variable o covariable
predictora o independiente

El modelo lineal (o cuando "la media se mueve")

$$weight_i \sim Normal(\mu_i, \sigma)$$

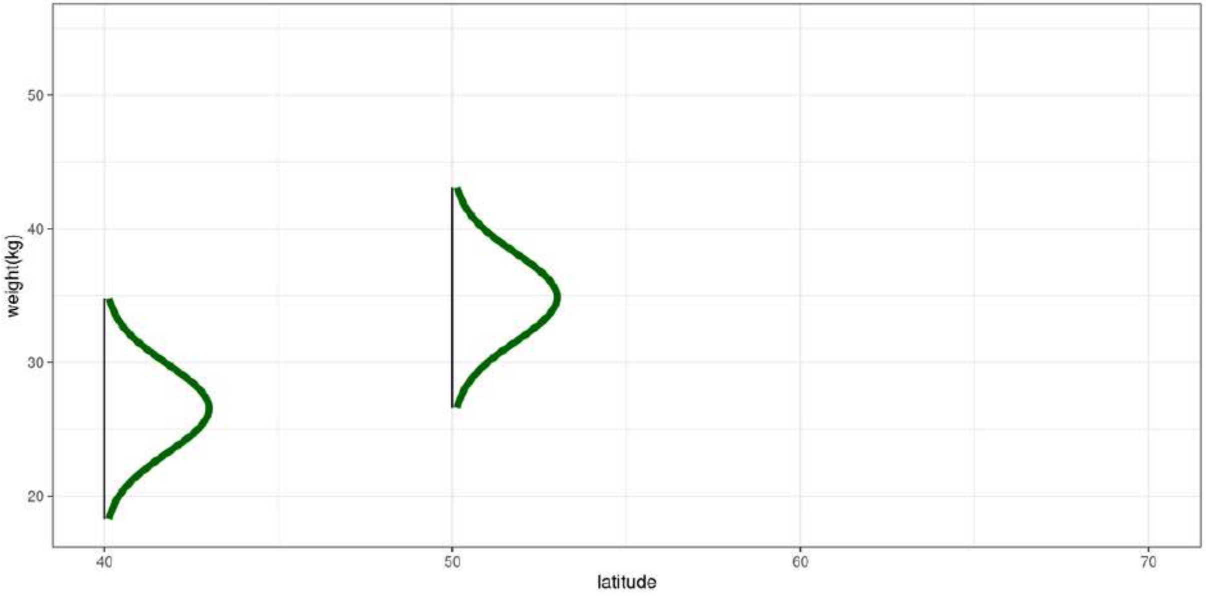
$$\mu_i = \beta_0 + \beta_1 Latitude_i$$



El modelo lineal (o cuando “la media se mueve”)

$$weight_i \sim Normal(\mu_i, \sigma)$$

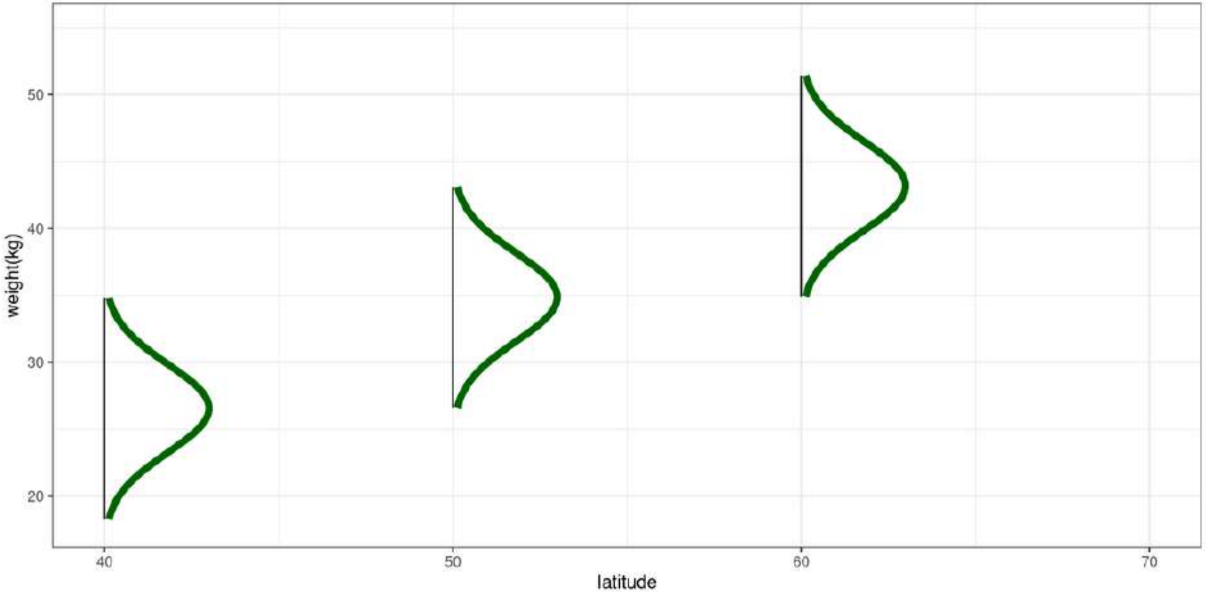
$$\mu_i = \beta_0 + \beta_1 Latitude_i$$



El modelo lineal (o cuando "la media se mueve")

$$weight_i \sim Normal(\mu_i, \sigma)$$

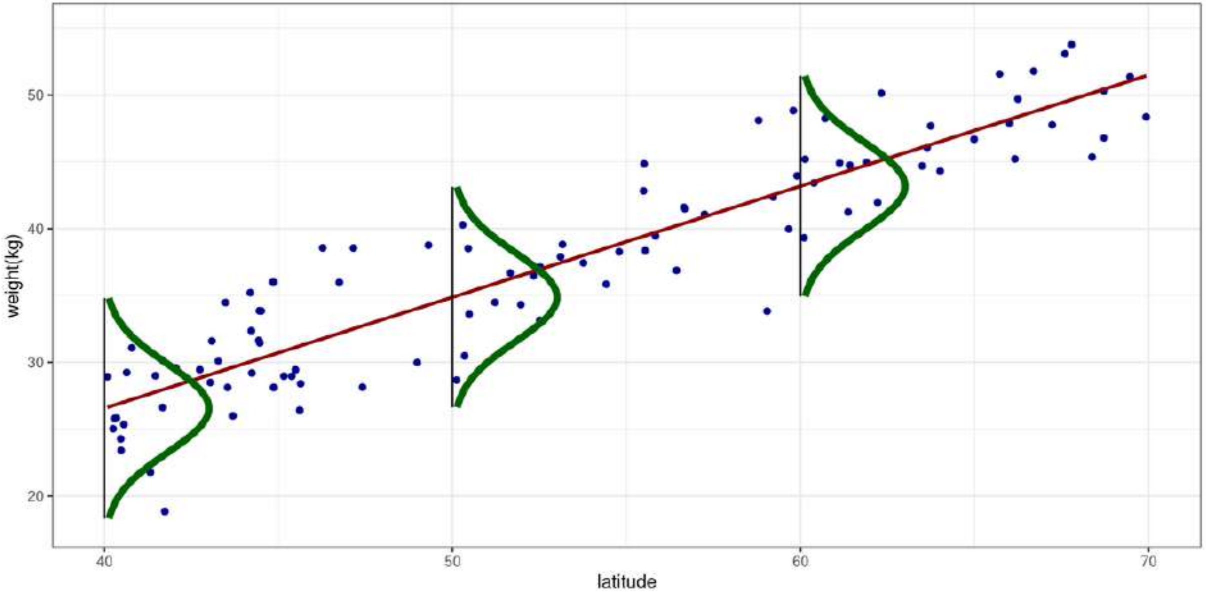
$$\mu_i = \beta_0 + \beta_1 Latitude_i$$



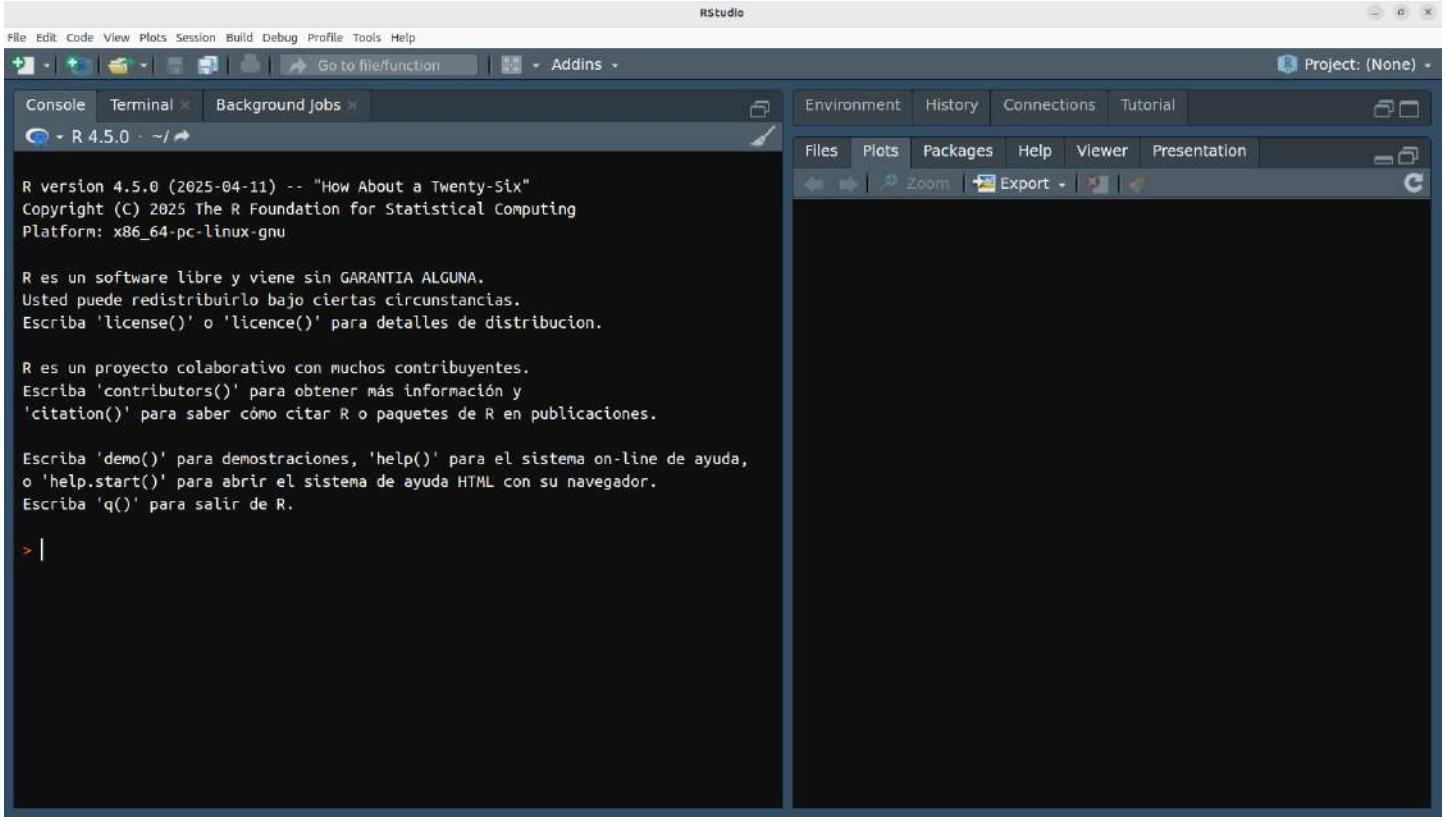
El modelo lineal (o cuando "la media se mueve")

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 Latitude_i$$



El modelo lineal (o cuando “la media se mueve”)



El modelo lineal

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 Latitude_i$$

El modelo lineal

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \boxed{\beta_0} + \beta_1 Latitude_i$$

- **Intercepto (β_0)**: Valor esperado de la variable respuesta cuando el predictor vale cero. En muchas ocasiones se interpreta como un valor de referencia matemático, no literal.

El modelo lineal

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \boxed{\beta_1} Latitude_i$$

- **Intercepto (β_0)**: Valor esperado de la variable respuesta cuando el predictor vale cero. En muchas ocasiones se interpreta como un valor de referencia matemático, no literal.

- ***Estimates* o coeficientes (β_1)**: tasa de cambio esperada en la variable respuesta por una unidad de aumento en el predictor. Cada aumento de 1 grado de latitud se asocia con un incremento promedio de 0.8 kg en el peso esperado. Es la pendiente de la recta ($\Delta y/\Delta x$).

El modelo lineal: asunciones

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 Latitude_i$$

- No hay heterogeneidad no explicada (no faltan predictores).

El modelo lineal: asunciones

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 Latitude_i$$

- No hay heterogeneidad no explicada (no faltan predictores).

- Conocemos la relación entre el predictor y la respuesta (y es lineal en los parámetros).

El modelo lineal: asunciones

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 Latitude_i$$

- No hay heterogeneidad no explicada (no faltan predictores).

- Conocemos la relación entre el predictor y la respuesta (y es lineal en los parámetros).

- Conocemos la distribución de los datos: Gaussiana o Normal.

El modelo lineal: asunciones

$$weight_i \sim Normal(\mu_i, \sigma)$$

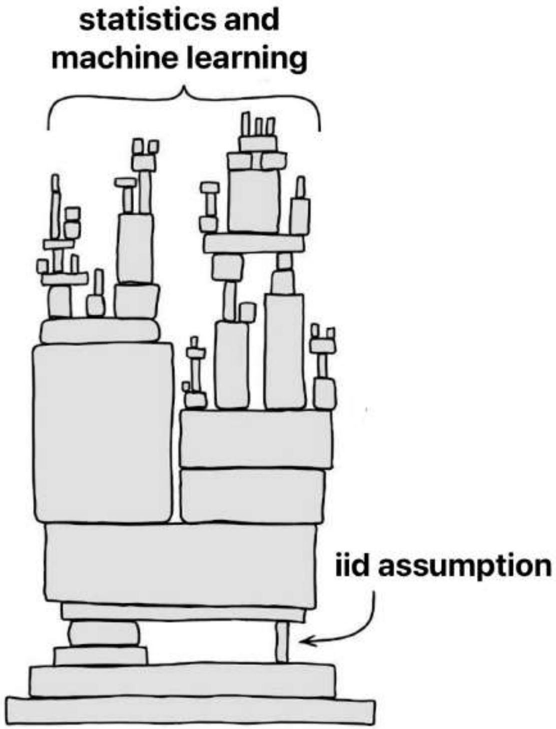
$$\mu_i = \beta_0 + \beta_1 Latitude_i$$

- No hay heterogeneidad no explicada (no faltan predictores).

- Conocemos la relación entre el predictor y la respuesta (y es lineal en los parámetros).

- Conocemos la distribución de los datos: Gaussiana o Normal.

- Los errores son *iid*: "*independent and identically distributed*", en nuestro caso con una distribución normal con media en 0.



El modelo lineal: asunciones

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 Latitude_i$$

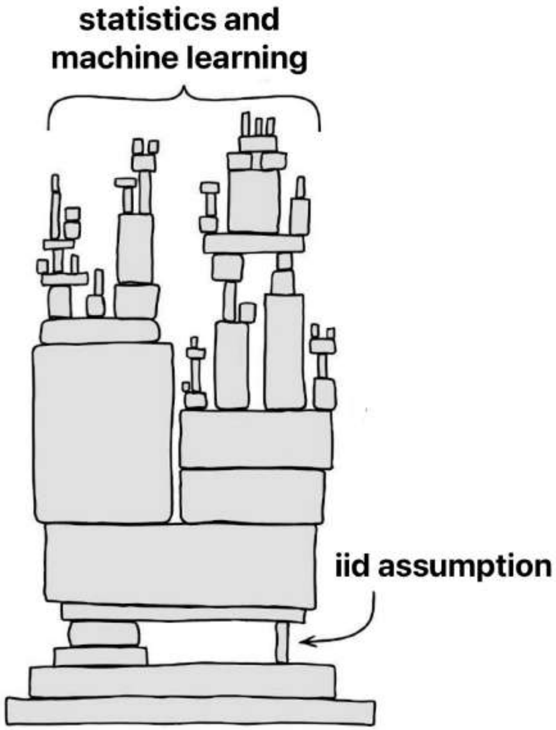
- No hay heterogeneidad no explicada (no faltan predictores).

- Conocemos la relación entre el predictor y la respuesta (y es lineal en los parámetros).

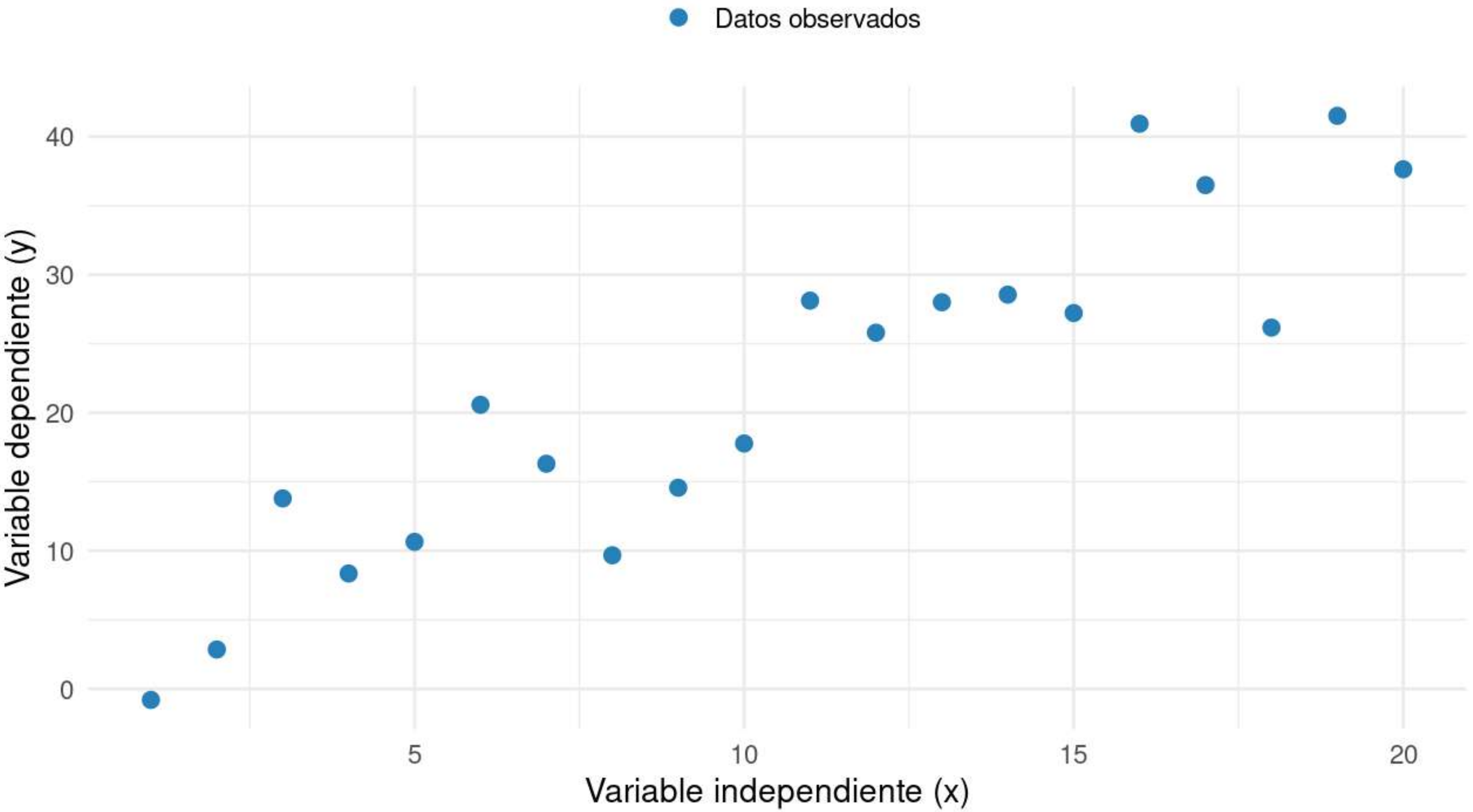
- Conocemos la distribución de los datos: Gaussiana o Normal.

- Los errores son *iid*: "*independent and identically distributed*", en nuestro caso con una distribución normal con media en 0.

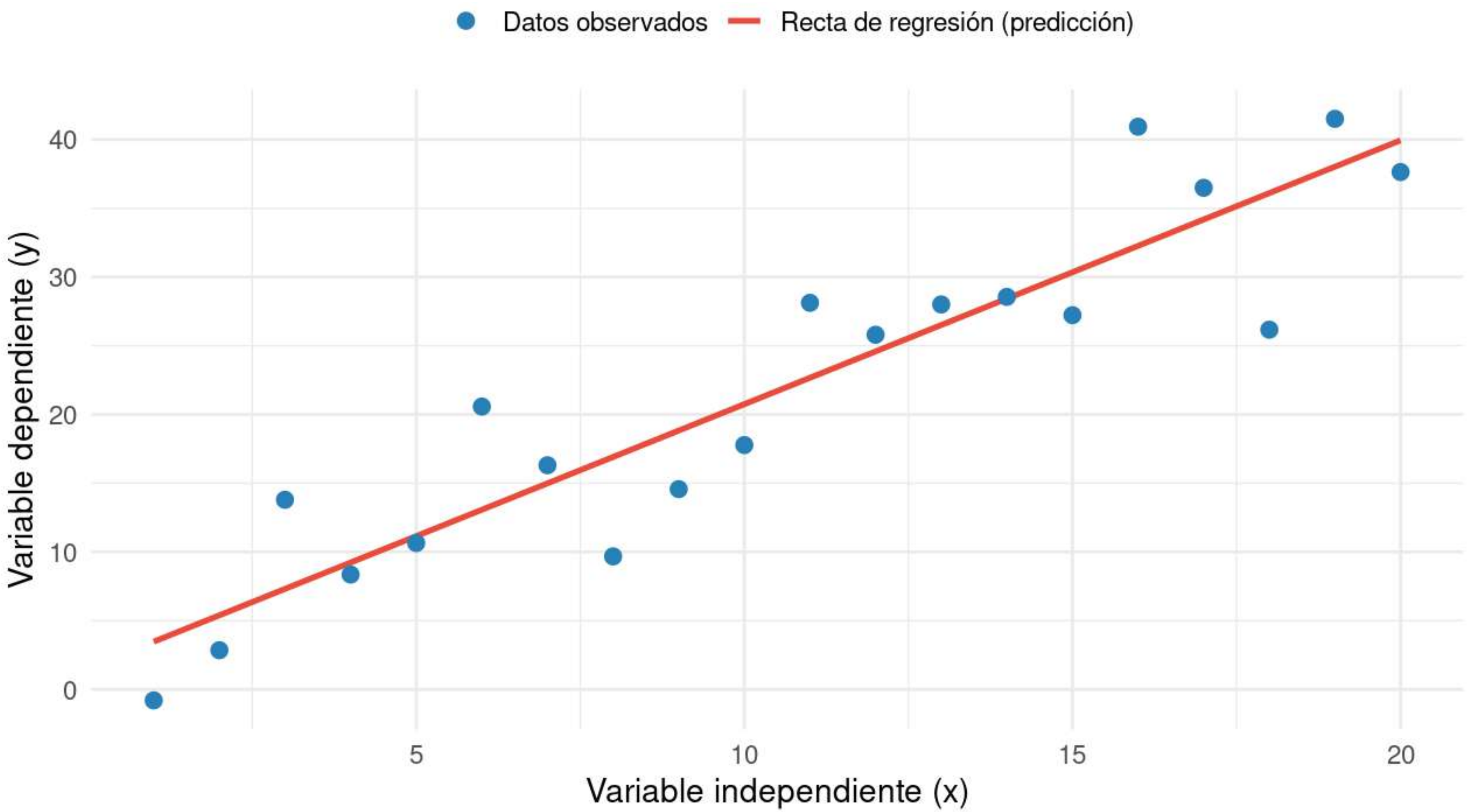
- La varianza se mantiene constante: homocedasticidad.



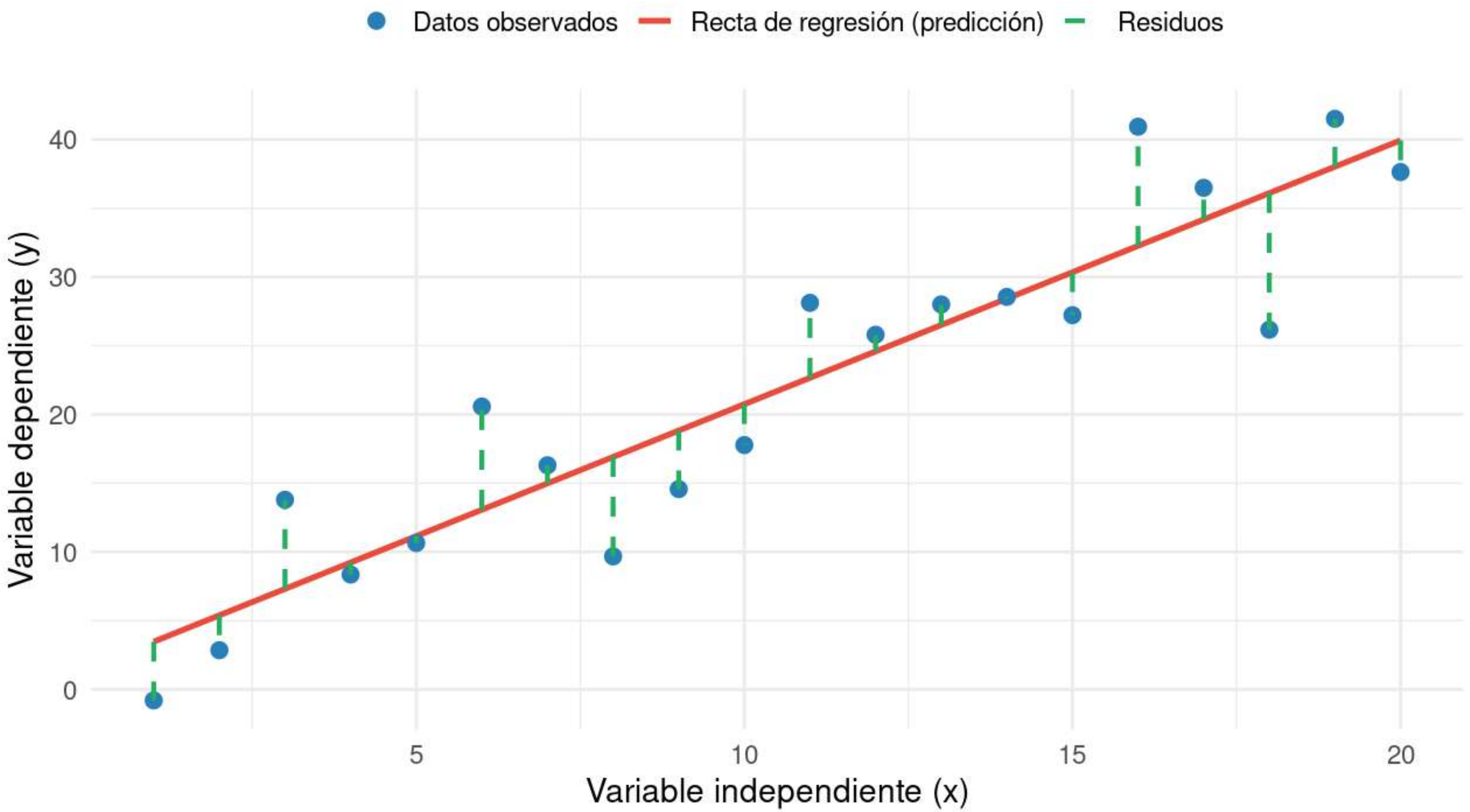
El modelo lineal: los residuos



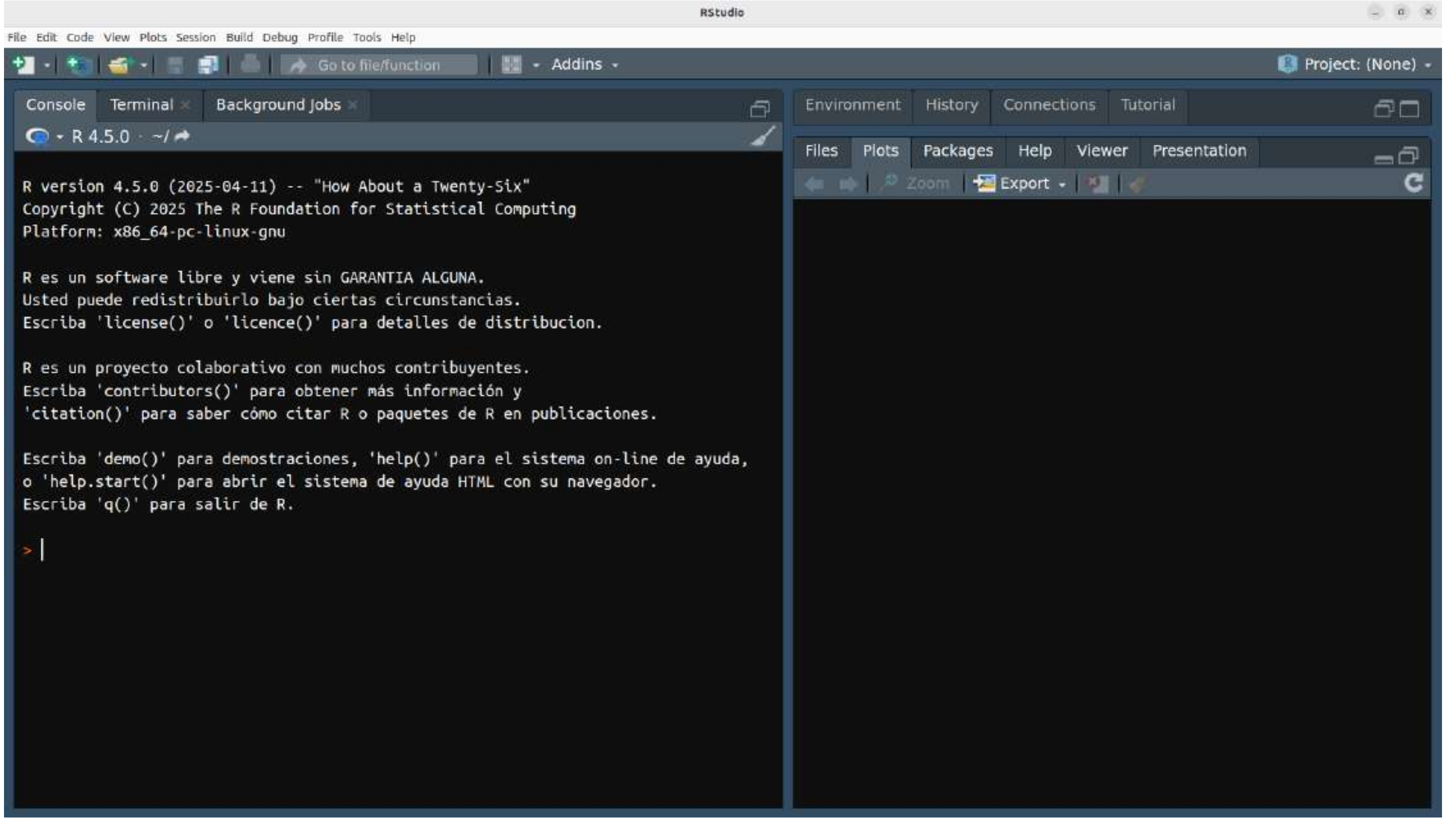
El modelo lineal: los residuos



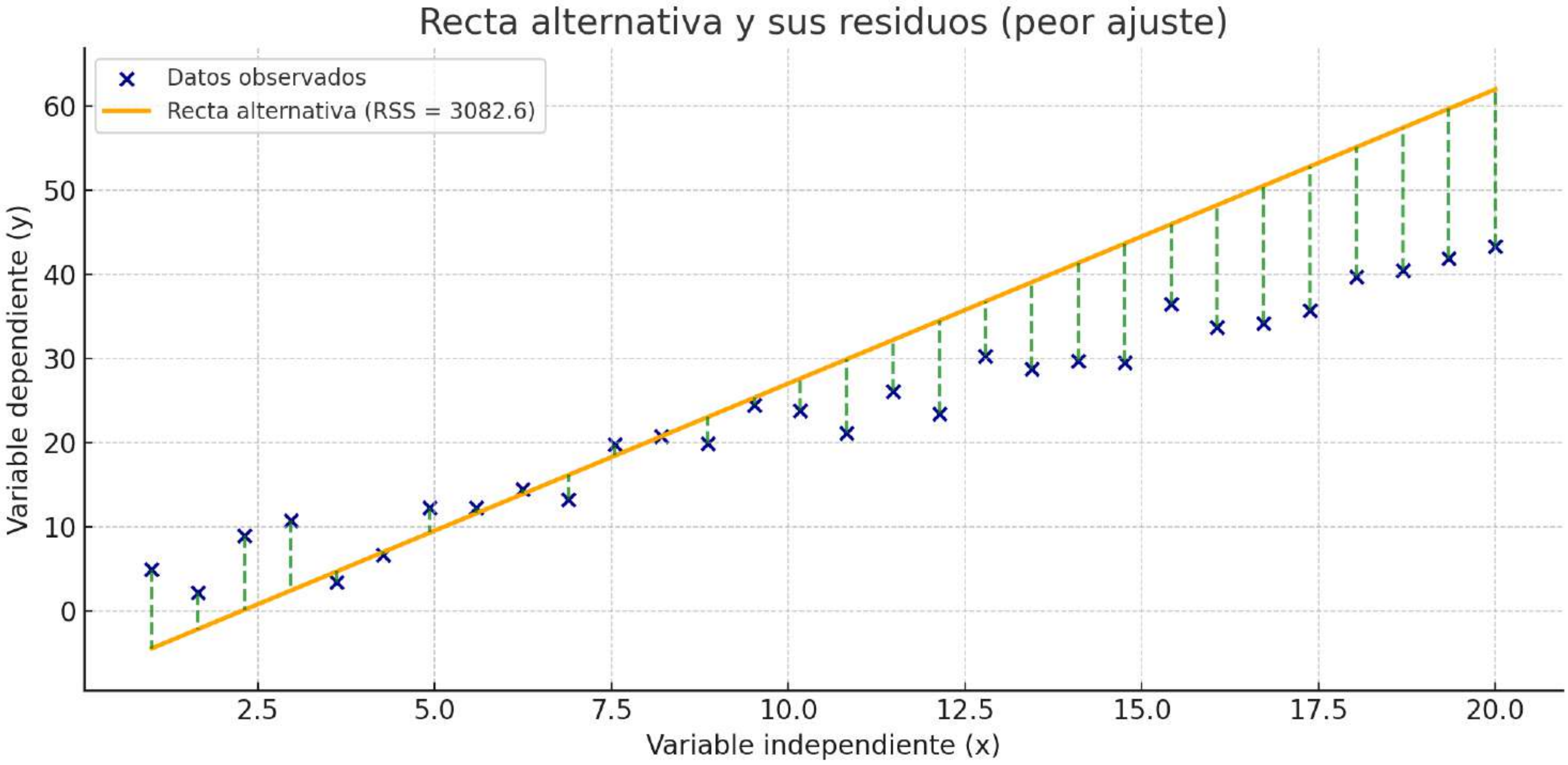
El modelo lineal: los residuos



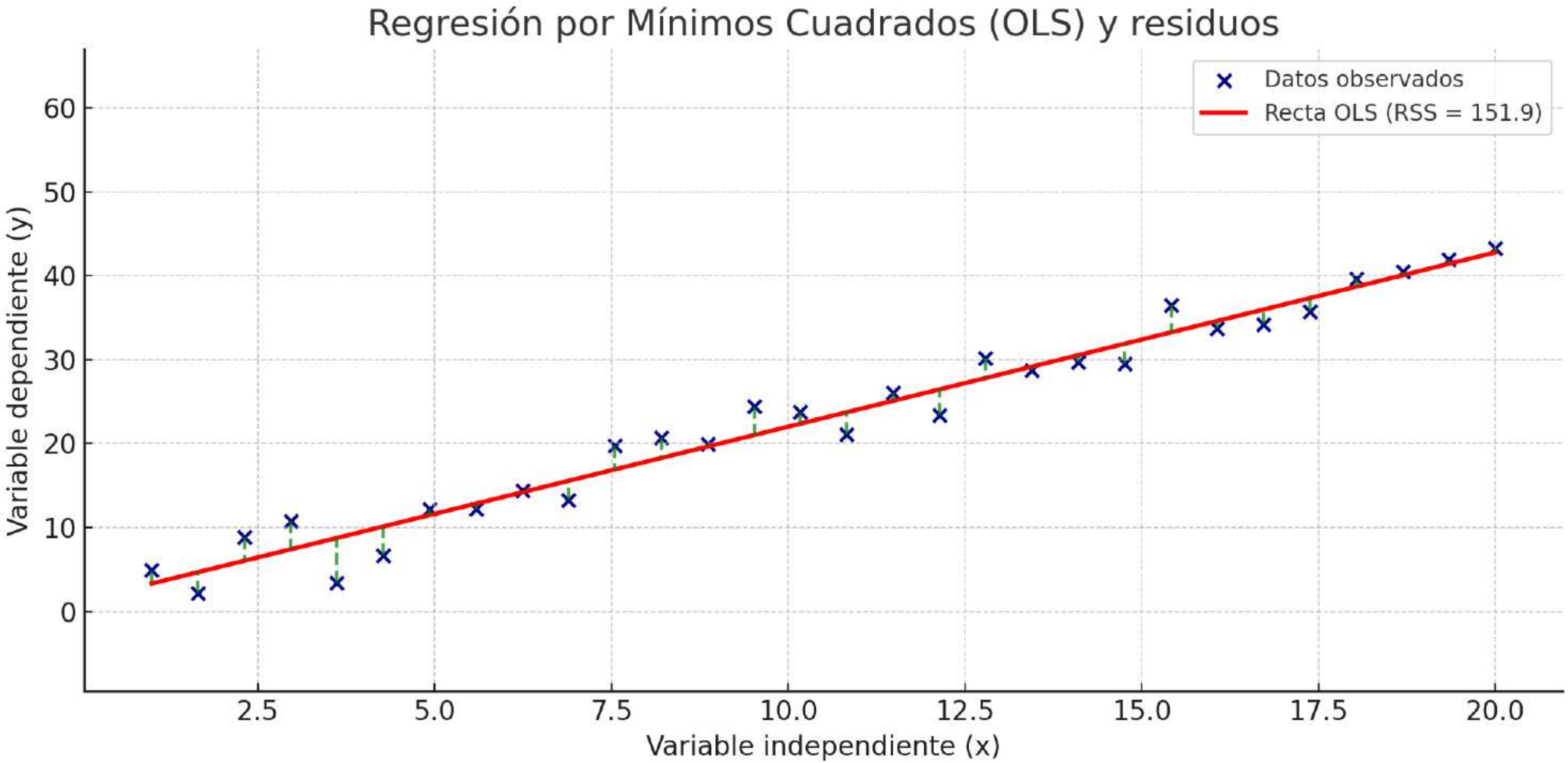
El modelo lineal: los residuos



El modelo lineal: mínimos cuadrados ordinarios (OLS)



El modelo lineal: mínimos cuadrados ordinarios (OLS)



El modelo lineal: mínimos cuadrados ordinarios (OLS)

- El método de mínimos cuadrados ordinarios (*ordinary least squares*, OLS) busca estimar los parámetros β_0, β_1,…, β_n de un modelo lineal de forma que los valores ajustados o predichos estén lo más cerca posible de los valores observados.

El modelo lineal: mínimos cuadrados ordinarios (OLS)

- El método de mínimos cuadrados ordinarios (*ordinary least squares*, OLS) busca estimar los parámetros β_0, β_1,…, β_n de un modelo lineal de forma que los valores ajustados o predichos estén lo más cerca posible de los valores observados.

- En otras palabras, OLS minimiza la suma de los cuadrados de los residuos:

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2$$

Recap...

muestra

modelo/ inferencia

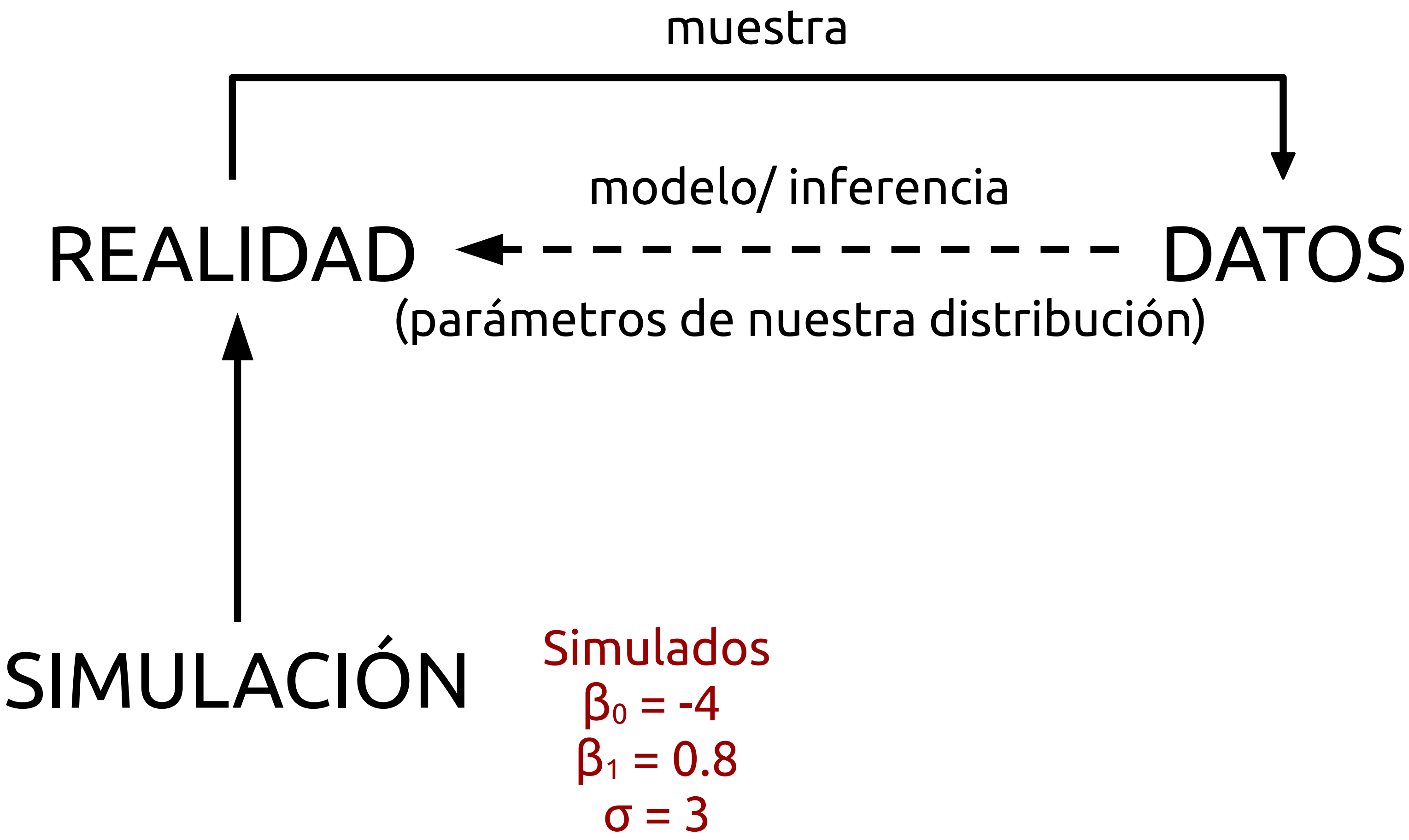
REALIDAD ← - - - - - - - - - - - - DATOS

(parámetros de nuestra distribución)

SIMULACIÓN

Número de parámetros simulados?

Recap...


muestra
modelo/ inferencia
REALIDAD
DATOS
(parámetros de nuestra distribución)
SIMULACIÓN
Simulados
$\beta_0 = -4$
$\beta_1 = 0.8$
$\sigma = 3$

Recap...

muestra

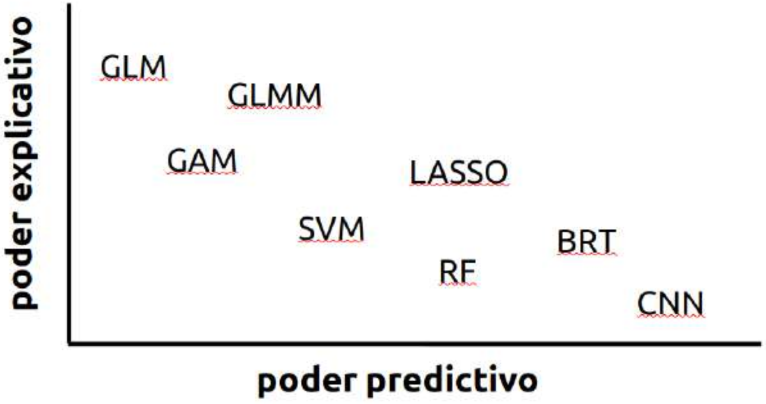
modelo/ inferencia

REALIDAD ← - - - - - - - - - - - - DATOS

(parámetros de nuestra distribución)

SIMULACIÓN

Simulados
$\beta_0 = -4$
$\beta_1 = 0.8$
$\sigma = 3$



Recap...

muestra

modelo/ inferencia

REALIDAD ← - - - - - - - - - - - - DATOS

(parámetros de nuestra distribución)

SIMULACIÓN

Simulados
$\beta_0 = -4$
$\beta_1 = 0.8$
$\sigma = 3$

Recap...

muestra

modelo/ inferencia

REALIDAD ← - - - - - - - - - - - - DATOS

(parámetros de nuestra distribución)

Recuperados

β_0 = ?

β_1 = ?

σ = ?

SIMULACIÓN

Simulados

$\beta_0 = -4$

$\beta_1 = 0.8$

$\sigma = 3$

Ejercicio propuesto (5 min)

- Simula una variable predictora usando una distribución uniforme o una distribución normal, pero que tenga unos valores coherentes con lo que quieres simular (temperatura, precipitación, cobertura de bosque, peso, etc.).

- Crea un predictor lineal.

- Genera observaciones a partir de una distribución lineal cuya media sea el predictor lineal y la desviación estándar que elijas.

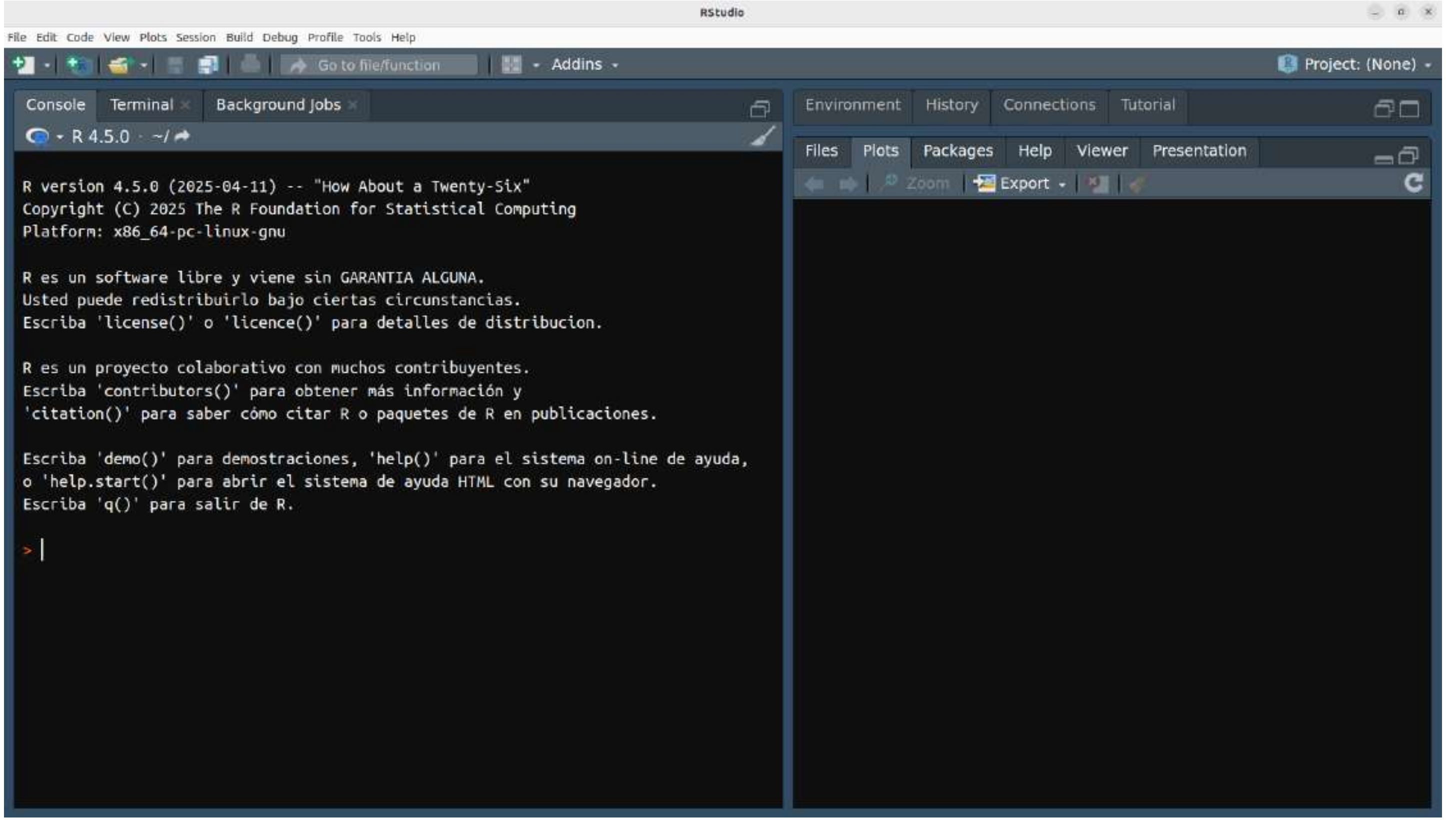
- Ajusta un modelo lineal usando la función glm().

Sobre el P-valor...

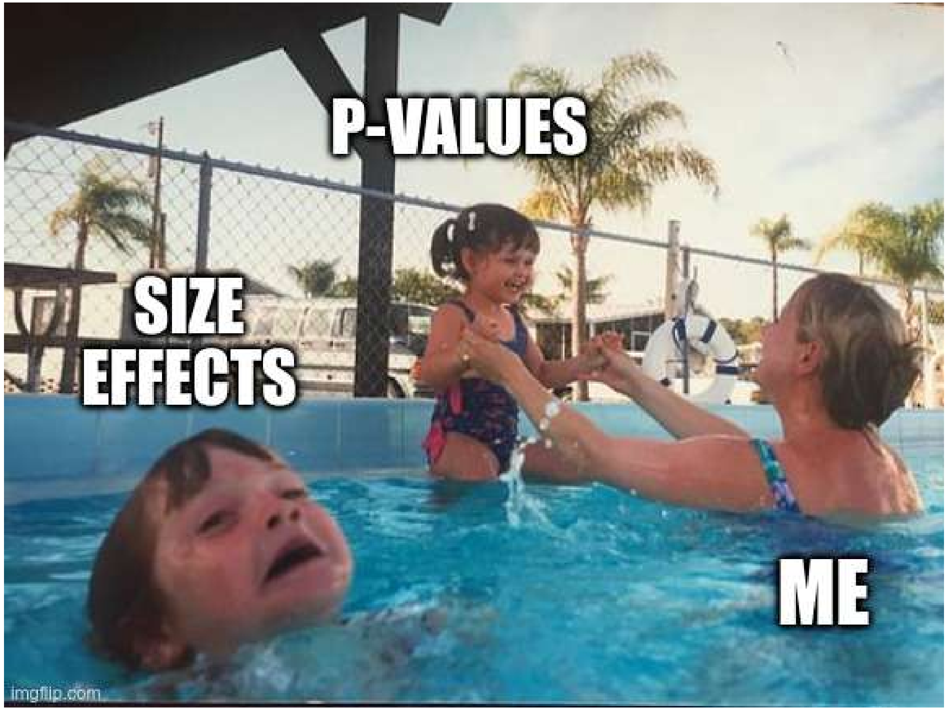
- El p-valor es la probabilidad de obtener unos datos tan extremos como los observados, si la hipótesis nula (el no-efecto del predictor) fuera cierta.

- “Si el predictor no tuviese efecto ($\beta_1 = 0$ realmente), observar un resultado igual o más extremo que este ocurriría el p-valor % de las veces”

El modelo lineal: gráficos diagnóstico



Sobre tamaño del efecto…



Sobre tamaño del efecto…

- Es el cambio esperado en la variable respuesta por cada unidad de cambio en la variable predictora.

- En un modelo lineal simple, es literalmente la definición de “pendiente de una recta”.

Sobre tamaño del efecto…

- Es el cambio esperado en la variable respuesta por cada unidad de cambio en la variable predictora.

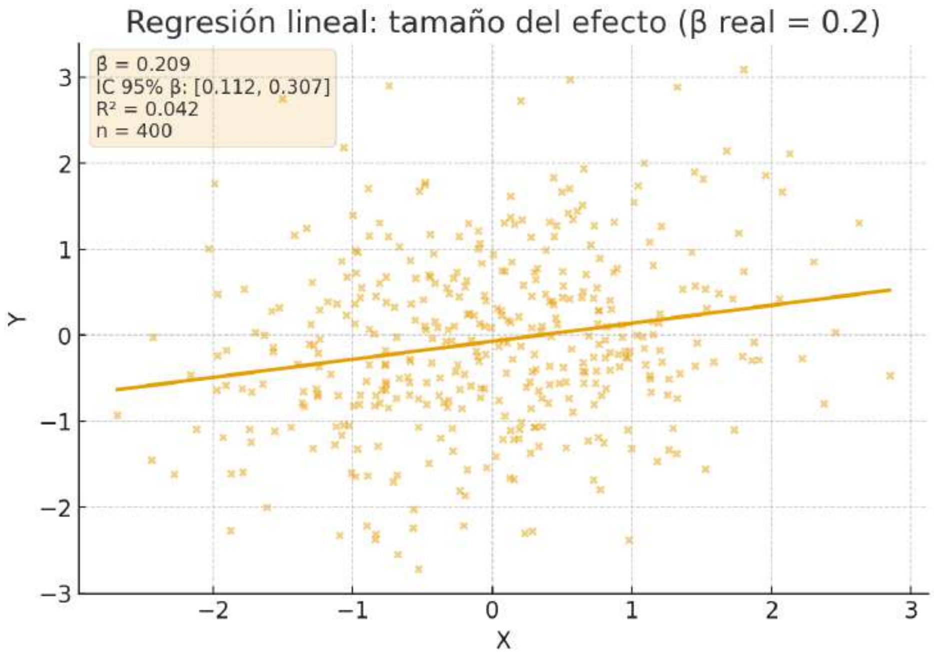
- En un modelo lineal simple, es literalmente la definición de “pendiente de una recta”.

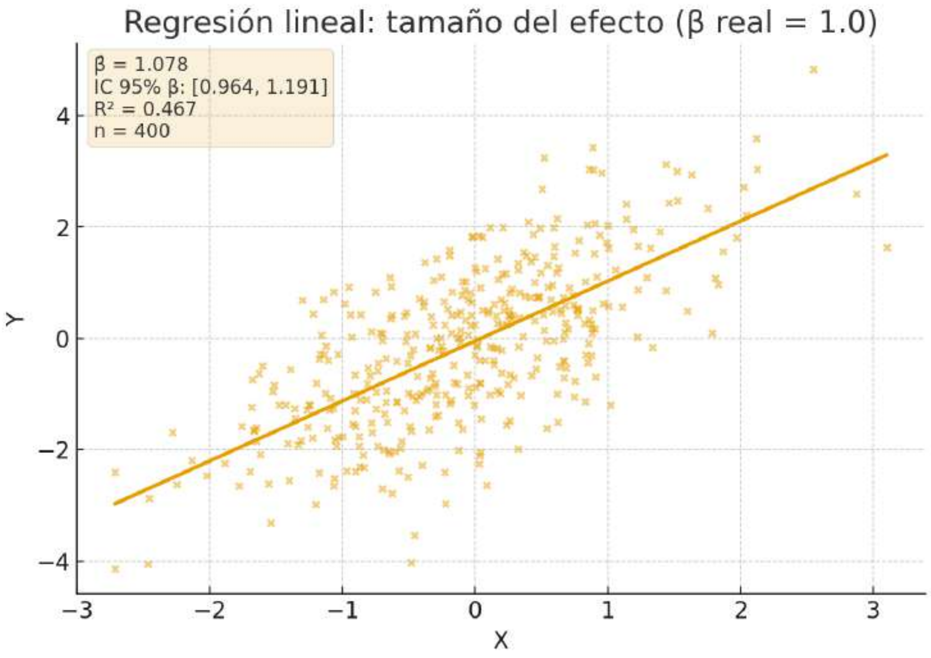
$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \boxed{\beta_1} X_i$$

$$\boxed{\beta_1} = \frac{\Delta \mu}{\Delta X} = \frac{\Delta\, weight}{\Delta\, latitud}$$

Sobre tamaño del efecto…





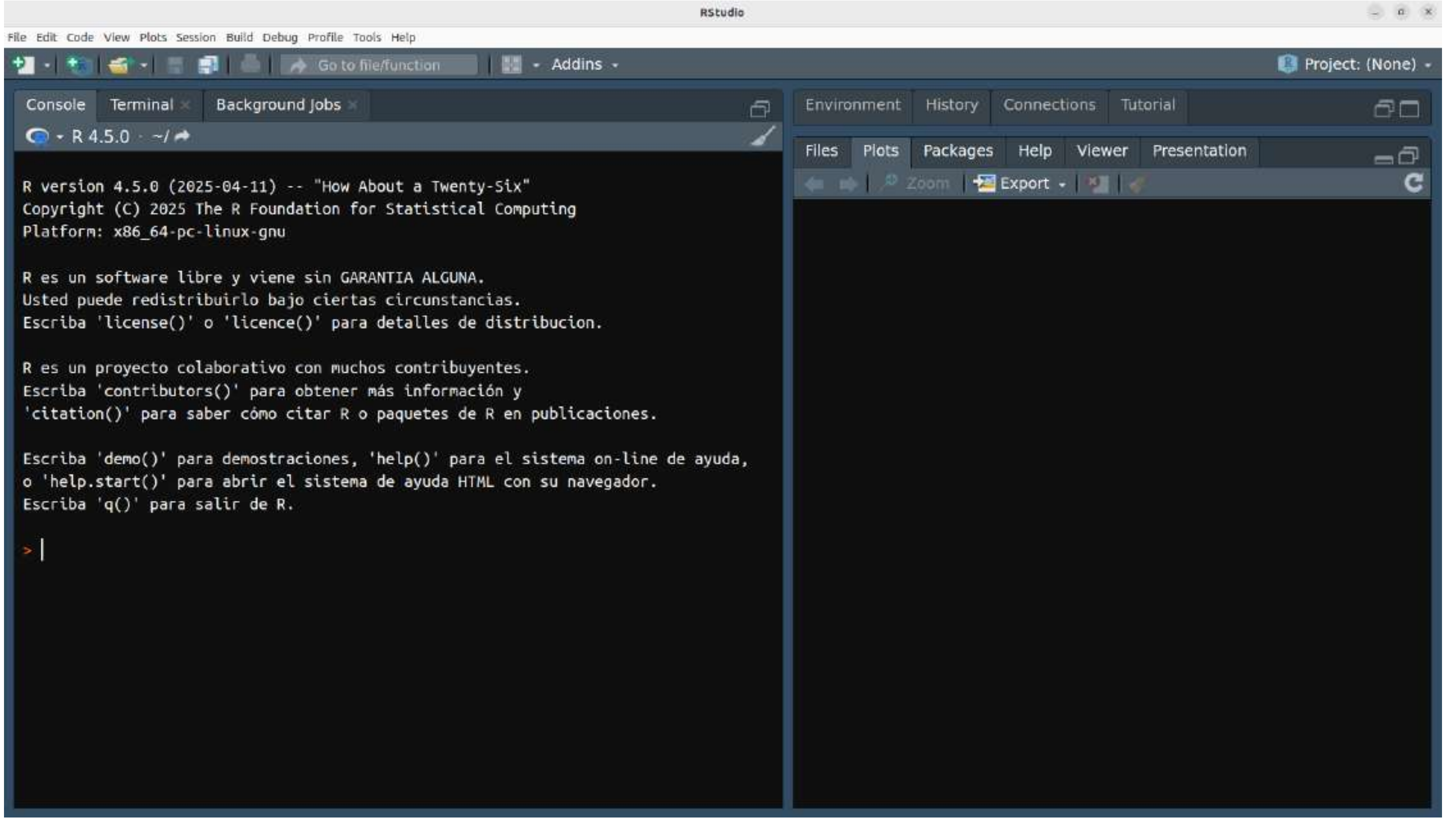
Sobre tamaño del efecto…

- Es el cambio esperado en la variable respuesta por cada unidad de cambio en la variable predictora.

- En un modelo lineal simple, es literalmente la definición de “pendiente de una recta”.

- Es muy importante tener en cuenta las unidades en las que están los predictores, puesto que afecta directamente al coeficiente que lo acompaña.

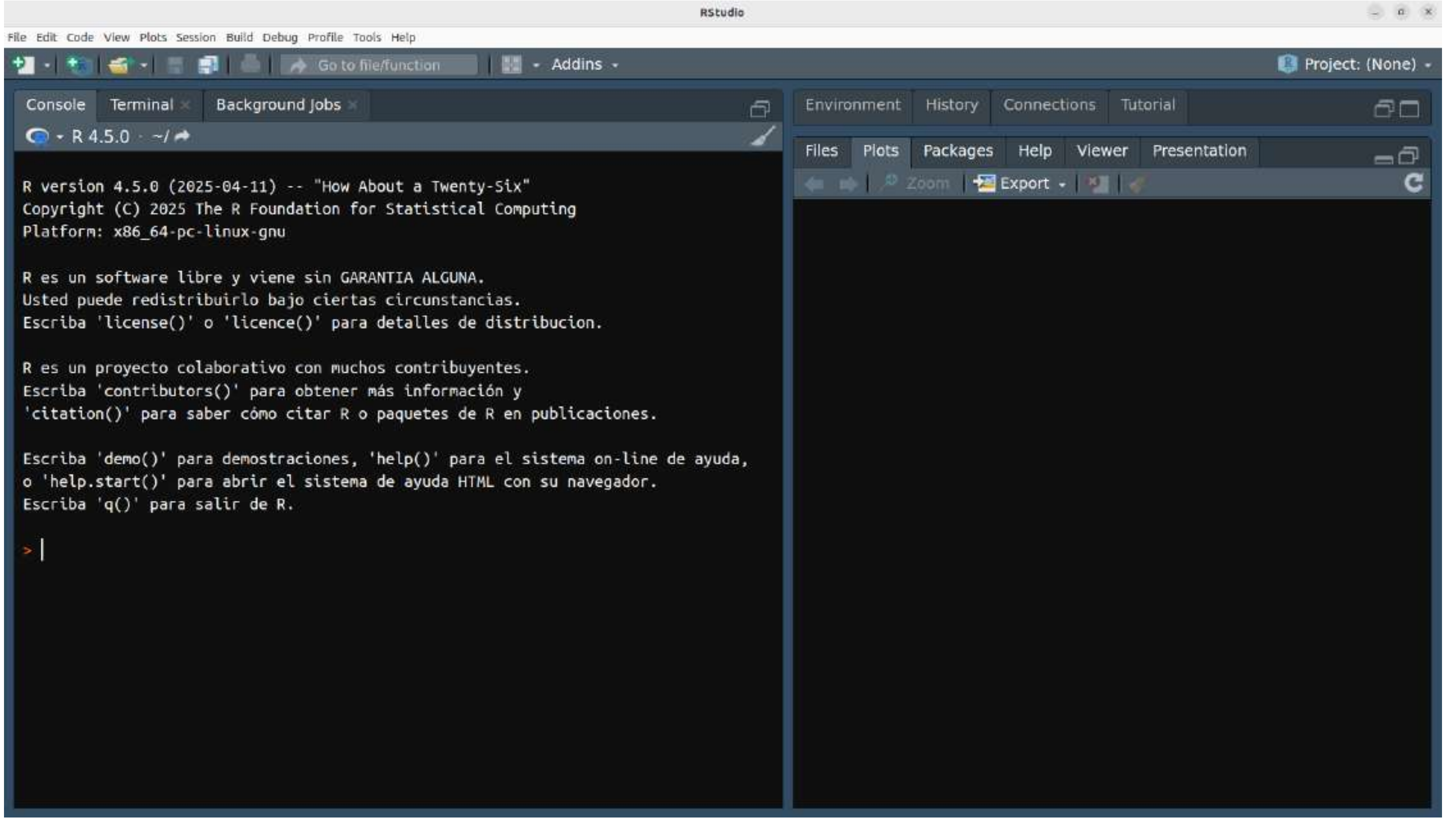
El modelo lineal: gráficos diagnóstico



Gráficos diagnóstico

Gráfico	Eje X	Eje Y	Qué diagnostica	Qué asunción comprueba
1️⃣ Residuals vs Fitted	Valores ajustados	Residuos	No linealidad / heterocedasticidad	Linealidad, homocedasticidad
2️⃣ Q–Q plot	Cuantiles teóricos	Cuantiles de residuos	Normalidad	Normalidad de los errores
3️⃣ Scale–Location	Valores ajustados	√	Residuos est.	
4️⃣ Residuals vs Leverage	Leverage	Residuos est.	Influencia / outliers	Independencia, influencia

El modelo lineal: gráficos diagnóstico



Bondad de ajuste

- Medida que indica cuánto coinciden los valores **predichos** por un modelo con los valores **observados**, reflejando su capacidad para **describir adecuadamente los datos**.

Bondad de ajuste

- Medida que indica cuánto coinciden los valores **predichos** por un modelo con los valores **observados**, reflejando su capacidad para **describir adecuadamente los datos**.

- En un modelo lineal, podemos usar el **R^2** o el **σ residual**, que evalúa la **varianza explicada por el modelo**.

Bondad de ajuste

- Medida que indica cuánto coinciden los valores **predichos** por un modelo con los valores **observados**, reflejando su capacidad para **describir adecuadamente los datos**.

- En un modelo lineal, podemos usar el **R^2** o el **σ residual**, que evalúa la **varianza explicada por el modelo**.

- En un modelo lineal generalizado es algo más complejo. La métrica fundamental es la **devianza**, que compara tu modelo con un modelo que fuese "perfecto", esto es, que ajustase perfectamente todos tus datos (modelo saturado).

→ **Devianza pequeña** → tu modelo se acerca al saturado → **mejor ajuste**

→ **Devianza grande** → tu modelo se aleja del saturado → **peor reflejo de los datos**

Bondad de ajuste

- También puede usarse un pseudo R^2 que, basado en la devianza, informa sobre la ganancia de poder predictivo de tu modelo respecto al modelo nulo:

$$R^2_{\mathrm{McFadden}} = 1 - \frac{D_{\mathrm{modelo}}}{D_{\mathrm{nulo}}}$$

- D_{modelo}: deviance del modelo ajustado
- D_{nulo}: deviance del modelo con solo intercepto

Valor de R^2_{McFadden}	**Interpretación aproximada**
0.0 – 0.2	Ajuste pobre
0.2 – 0.4	Ajuste razonable
0.4 – 0.6	Ajuste bueno
> 0.6	Ajuste excelente

El modelo lineal: bondad de ajuste