Análisis Estadístico de Datos

Sesión 2: Probabilidad y Pensamiento Bivariado

Profesor: Nicolás López

2026-02-27

Puente con la Sesión 1

Resumen Sesión 1

En la primera sesión abordamos los fundamentos del análisis estadístico univariado:

- **R y RStudio**: lenguaje, entorno y ecosistema `tidyverse`
- **Estructuras de datos**: vectores, data frames, tipos de variable
- **Clasificación de variables**: cuantitativas vs. cualitativas; continuas vs. discretas
- **Distribución univariada**: histograma, boxplot, descripción visual
- **Medidas de tendencia central**: media, mediana, moda
- **Medidas de dispersión**: varianza, desviación estándar, rango intercuartílico
- **Forma**: asimetría y curtosis
- **Atipicidades**: detección e interpretación de valores extremos
- **Covarianza**: primera aproximación a la relación entre dos variables

¿Por qué ir más allá de lo univariado?

La estadística descriptiva univariada nos dice **cómo se comporta una sola variable**, pero el mundo real es **multidimensional**:

Un país puede producir mucho carbón — pero ¿cuánto consume?
¿Existe una relación sistemática entre producción y consumo?

Para responder a estas preguntas necesitamos:

1. Un marco probabilístico formal (**funciones de densidad**)
2. Herramientas para estudiar la **relación entre variables** (pensamiento bivariado)

Estos son los temas de la **Sesión 2**.

Tabla de contenidos

1. Dependencias y datos

Paquetes necesarios

```
install.packages("tidyverse")
install.packages("ggExtra")
```

Cargamos las dependencias y los datos de carbón de la sesión anterior:

```
library(tidyverse)
data_charcoal  = read_csv("UNdata_Charcoal.csv",
                          show_col_types = FALSE)
charcoal_chh19 = data_charcoal %>%
  filter(Year == 2019 &
    Commodity == "Charcoal - Consumption by households") %>%
  select(-Commodity)
charcoal_prd19 = data_charcoal %>%
  filter(Year == 2019 &
    Commodity == "Charcoal - Production") %>%
  select(-Commodity)
```

2. Probabilidad

Experimentos aleatorios

Un **experimento aleatorio** es aquel cuyo resultado no puede predecirse con certeza, pero cuyo comportamiento a largo plazo presenta patrones regulares.

Ejemplos:

- Lanzar una moneda: $S = \{Cara, Sello\}$
- Lanzar un dado: $S = \{1, 2, 3, 4, 5, 6\}$
- Seleccionar un país y medir su producción de carbón

El conjunto de todos los posibles resultados es el **espacio muestral** S.

Conexión con los datos
Al filtrar los datos de producción de carbón para 2019 obtenemos una **muestra aleatoria** de países. Uzbekistán no figura en los datos — su ausencia es parte de la aleatoriedad del proceso de recolección.

Variables aleatorias

Una **variable aleatoria** X asigna un valor numérico a cada resultado del experimento:

$$Cara_1 \to 1, \quad Cara_2 \to 2, \quad \ldots, \quad Cara_6 \to 6$$

Ejemplo con los datos del curso:

Si $X =$ edad (en años) de los estudiantes, y el estudiante #3 (Pepito) tiene 27 años:

$$x_3 = 27$$

Con $n = 13$ estudiantes, la **muestra aleatoria** es $x_1, x_2, \ldots, x_{13}$ — realizaciones de las variables aleatorias $X_1, \ldots, X_{13}$.

Distinción clave
X_i (mayúscula) = variable aleatoria (antes de observar)
x_i (minúscula) = valor observado (realización concreta)

Tipos de variables aleatorias

Tipo	Descripción	Ejemplo
Discreta	Valores contables	N. de caras en 10 lanzamientos
Continua	Valores en $\mathbb{R}$	Producción de carbón (miles ton.)

- Variables **discretas**: función másica de probabilidad (fmp)
- Variables **continuas**: función de densidad de probabilidad (fdp)

Nota
En este curso trabajaremos principalmente con **variables continuas** y sus distribuciones de densidad.

3. Función de densidad

Modelos matemáticos del mundo real

La ciencia construye modelos matemáticos para describir y predecir fenómenos:

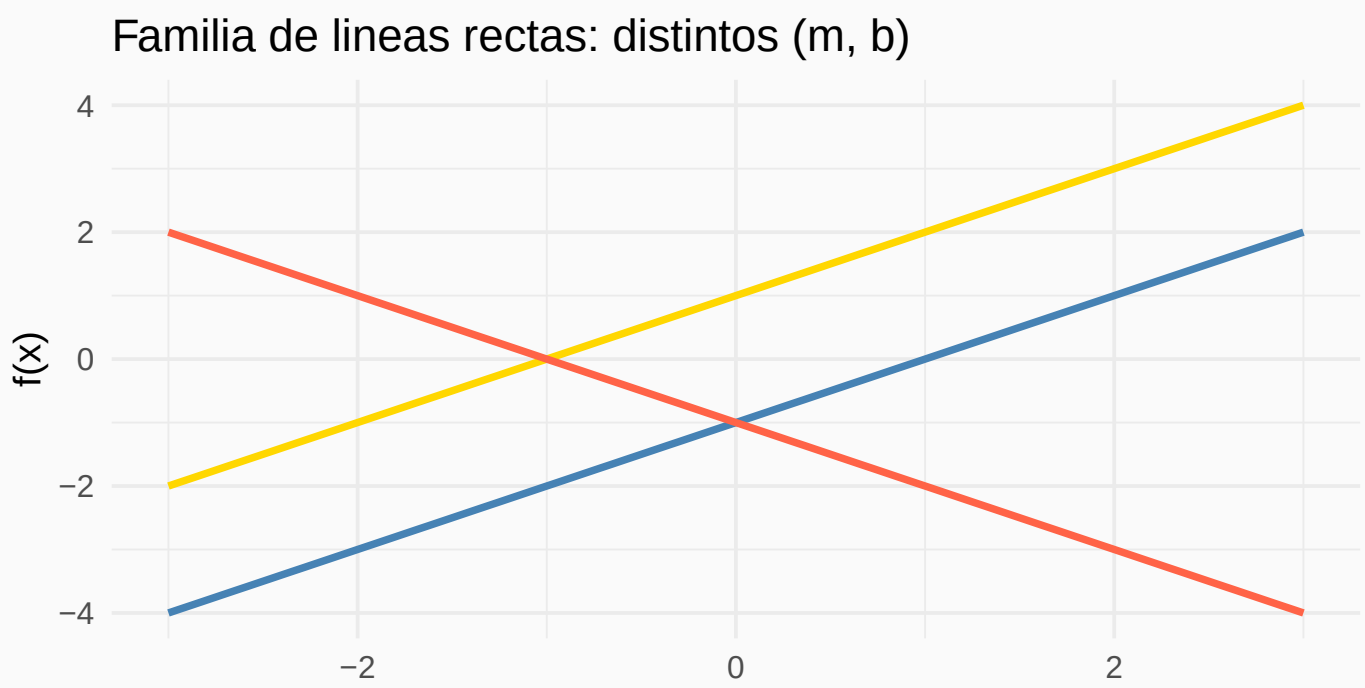
Modelo	**Ecuación**
Teorema de Pitágoras	$a^2 = b^2 + c^2$
Relatividad especial	$E = mc^2$
Línea recta	$f(x \mid m, b) = mx + b$

La **función de densidad de probabilidad** sigue la misma lógica: es un modelo matemático que describe cómo se distribuyen los valores de una variable aleatoria continua, indexado por un conjunto de **parámetros**.

La familia de líneas rectas como motivación

Una familia de funciones indexada por parámetros (m, b) — distintos valores generan distintas funciones:

```
f_lineal = function(x, m, b) { m * x + b }
ggplot() + xlim(-3, 3) +
  geom_function(fun=f_lineal, args=list(m=1,  b=1),  color="gold",      linewidth=1) +
  geom_function(fun=f_lineal, args=list(m=1,  b=-1), color="steelblue", linewidth=1) +
  geom_function(fun=f_lineal, args=list(m=-1, b=-1), color="tomato",    linewidth=1) +
  theme_minimal() +
  labs(y="f(x)", title="Familia de lineas rectas: distintos (m, b)")
```



Densidad paramétrica: la distribución normal

De manera análoga, la **distribución normal** es una familia indexada por (μ, σ):

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < +\infty$$

Parámetros:

- $\mu \in \mathbb{R}$: media — controla la **ubicación** de la campana
- $\sigma > 0$: desviación estándar — controla el **ancho** de la campana

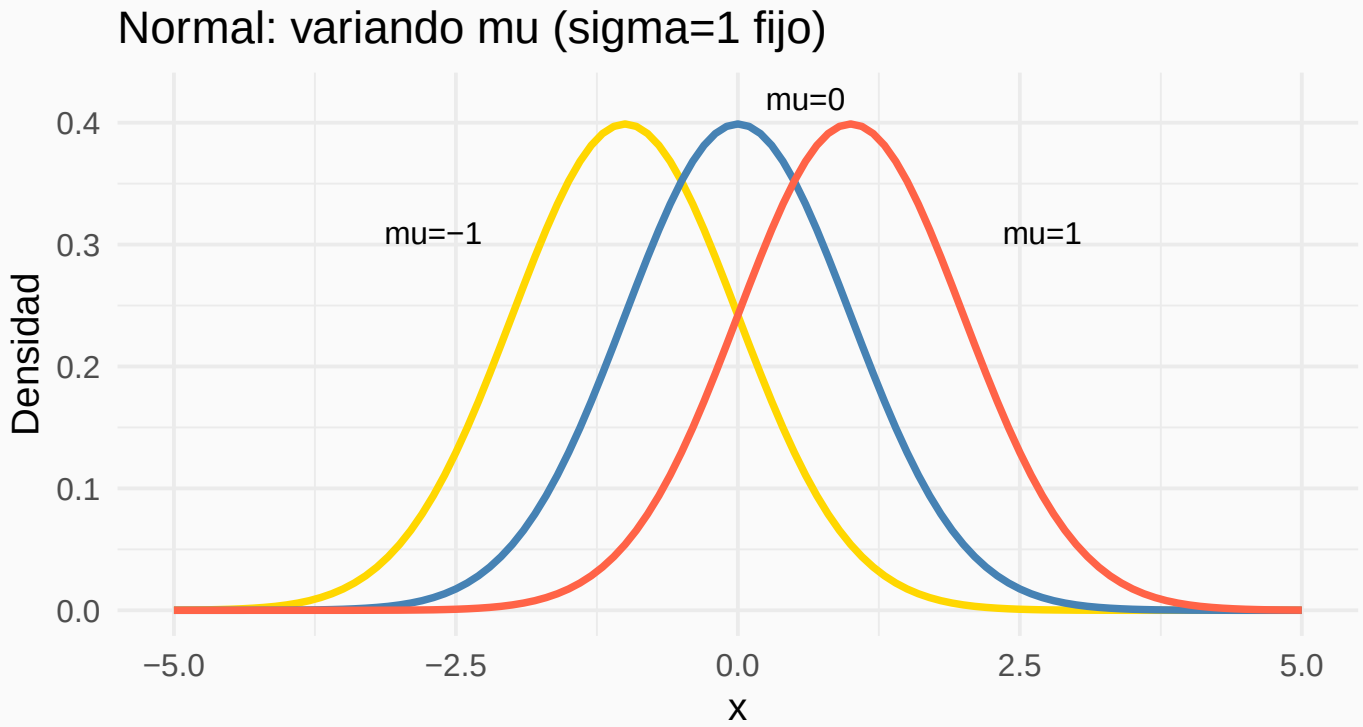
Propiedades clave:

- Simétrica alrededor de μ; forma de campana (*bell curve*)
- $\int_{-\infty}^{+\infty} f(x \mid \mu, \sigma)\, dx = 1$

Efecto del parámetro μ

Tres elementos de la familia con $\sigma = 1$ fijo, usando `dnorm()`:

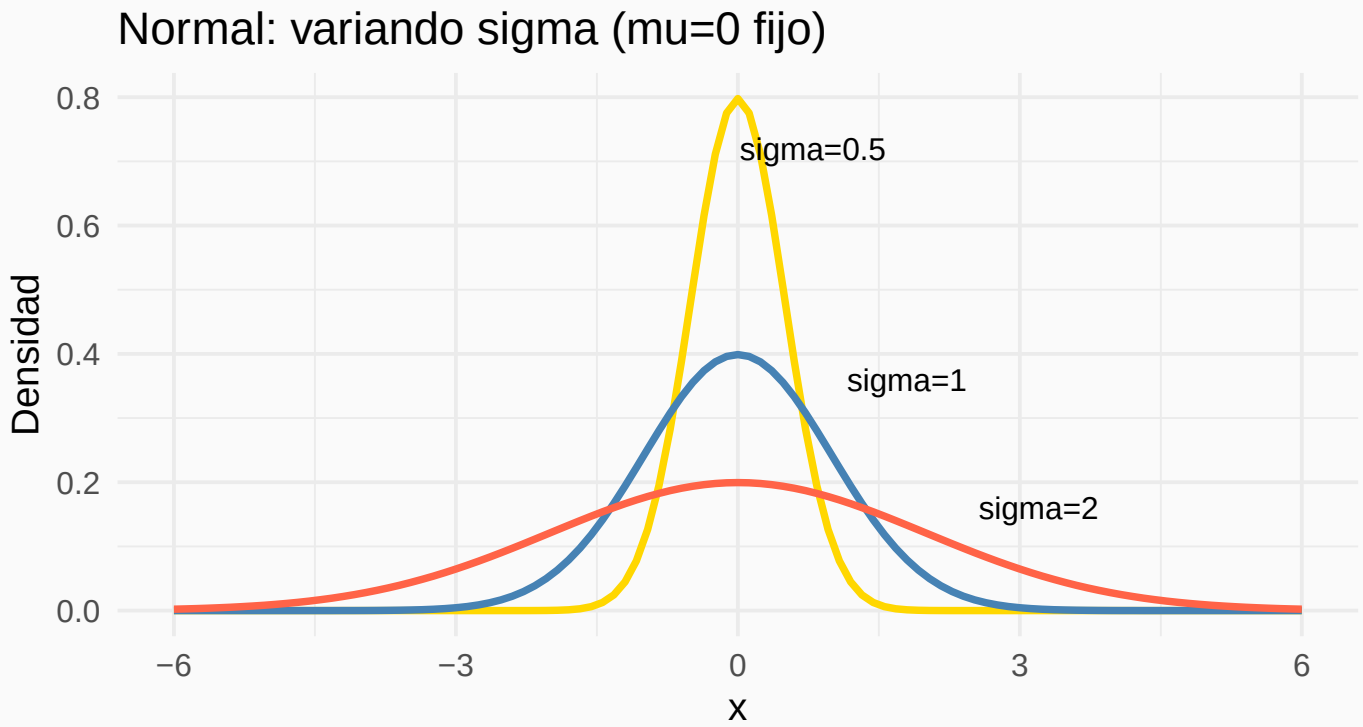
```
# dnorm(x, mean, sd) es la funcion nativa de R para la densidad normal
ggplot() + xlim(-5, 5) +
  geom_function(fun=dnorm, args=list(mean=-1, sd=1), color="gold",      linewidth=1) +
  geom_function(fun=dnorm, args=list(mean=0,  sd=1), color="steelblue", linewidth=1) +
  geom_function(fun=dnorm, args=list(mean=1,  sd=1), color="tomato",    linewidth=1) +
  annotate("text", x=c(-2.7, 0.6, 2.7), y=c(0.31, 0.42, 0.31),
           label = c("mu=-1", "mu=0", "mu=1"), size=3) +
  theme_minimal() +
  labs(y="Densidad", title="Normal: variando mu (sigma=1 fijo)")
```



Efecto del parámetro σ

Tres elementos de la familia con $\mu = 0$ fijo:

```
ggplot() + xlim(-6, 6) +
  geom_function(fun=dnorm, args=list(mean=0, sd=0.5), color="gold",      linewidth=1) +
  geom_function(fun=dnorm, args=list(mean=0, sd=1),   color="steelblue", linewidth=1) +
  geom_function(fun=dnorm, args=list(mean=0, sd=2),   color="tomato",    linewidth=1) +
  annotate("text", x=c(0.8, 1.8, 3.2), y=c(0.72, 0.36, 0.16),
           label = c("sigma=0.5", "sigma=1", "sigma=2"), size=3) +
  theme_minimal() +
  labs(y="Densidad", title="Normal: variando sigma (mu=0 fijo)")
```



La distribución normal estándar

La distribución normal **estándar** corresponde a $\mu = 0$ y $\sigma = 1$:

$$Z \sim \mathcal{N}(0,1) \quad \Longleftrightarrow \quad f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}$$

Regla empírica (68-95-99.7):

Intervalo	Probabilidad
$\mu \pm \sigma$	$\approx 68\%$
$\mu \pm 2\sigma$	$\approx 95\%$
$\mu \pm 3\sigma$	$\approx 99.7\%$

Conexión con la Sesión 1

Los **outliers** se detectan precisamente cuando los datos caen fuera de $\mu \pm 3\sigma$ — ahora tenemos un marco probabilístico para detección de outliers.

Importancia de la distribución normal

Omnipresencia de la distribución normal
Muchas técnicas estadísticas se basan en la distribución normal porque muchos fenómenos, al ser medidos, se aglomeran simétricamente en torno a un valor central — siguen aproximadamente esta distribución.

¿Por qué es tan importante? El Teorema Central del Límite (TCL)

El TCL establece que la **suma (o promedio) de muchas variables aleatorias independientes** tiende a una distribución normal, sin importar la distribución original de cada variable:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Esto la convierte en el modelo de probabilidad **más importante** de la estadística clásica.

El área bajo la curva como probabilidad

El área bajo la fdp entre a y b representa la probabilidad del intervalo:

$$P(a \leq X \leq b) = \int_a^b f(x \mid \mu, \sigma)\, dx$$

Propiedades:

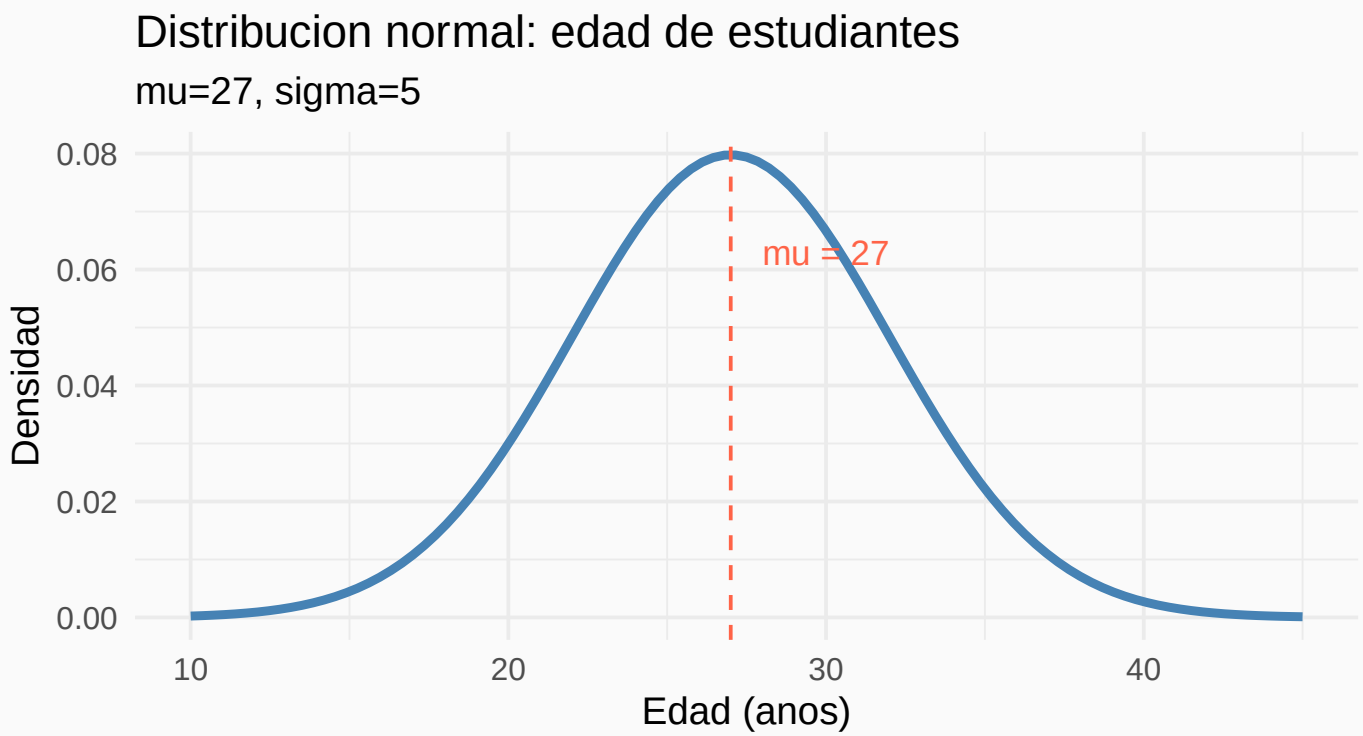
$$\int_{-\infty}^{+\infty} f(x \mid \mu, \sigma)\, dx = 1 \qquad \text{(certeza total)}$$

$$\int_{-\infty}^{\mu} f(x \mid \mu, \sigma)\, dx = 0.5 \qquad \text{(simetría)}$$

Ejemplo: Si $X =$ edad de los estudiantes $\sim \mathcal{N}(27, 5)$, entonces $P(X \leq 27) = 0.5$.

Visualización: normal para edades del curso

```r
ggplot() + xlim(10, 45) +
  geom_function(fun=dnorm, args=list(mean=27, sd=5),
                color="steelblue", linewidth=1.2) +
  geom_vline(xintercept=27, linetype="dashed", color="tomato") +
  annotate("text", x=30, y=0.063,
           label="mu = 27", color="tomato", size=3.5) +
  theme_minimal() +
  labs(y="Densidad", x="Edad (anos)",
       title="Distribucion normal: edad de estudiantes",
       subtitle="mu=27, sigma=5")
```

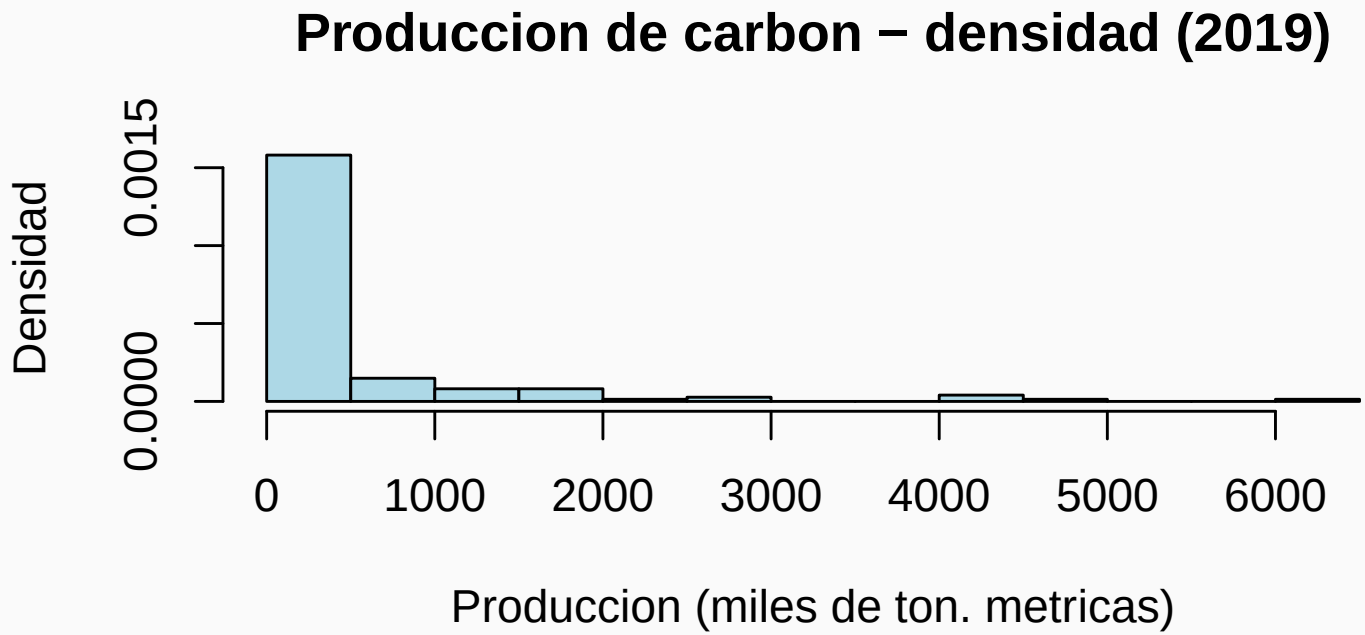


4. Estimación de densidad

El histograma como estimador de densidad

El histograma es la primera aproximación a la fdp subyacente. Con `prob=TRUE` el eje y pasa de frecuencia absoluta a **densidad** (área total $=1$):

```
h_dens = hist(charcoal_prd19$Quantity, prob=TRUE,
  main  = "Produccion de carbon - densidad (2019)",
  xlab  = "Produccion (miles de ton. metricas)",
  ylab  = "Densidad", col="lightblue")
```



Verificando que el área total = 1

```
l_base   = diff(h_dens$breaks)  # base de cada rectangulo
l_altura = h_dens$density       # altura estimada (densidad)
sum(l_base * l_altura)          # suma de areas = 1
```

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

Esta es una propiedad fundamental de cualquier fdp:

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$

Esto permite medir el chance de ocurrencia de cualquier evento como el **área bajo la curva** en el intervalo correspondiente.

Probabilidades desde el histograma

Probabilidad de seleccionar un país con producción entre 1000 y 1500 miles de ton.:

$$P(1000 \leq X \leq 1500) \approx \text{base} \times \text{altura} = 500 \times \hat{f}(1250)$$

```
base_1000_1500 = 500
altu_1000_1500 = h_dens$density[h_dens$mids == 1250]
prob_A_hist    = base_1000_1500 * altu_1000_1500
paste0("P(1000 <= X <= 1500) aprox. ",
       round(100 * prob_A_hist, 2), "%")
```

```
## [1] "P(1000 <= X <= 1500) aprox. 4.05%"
```

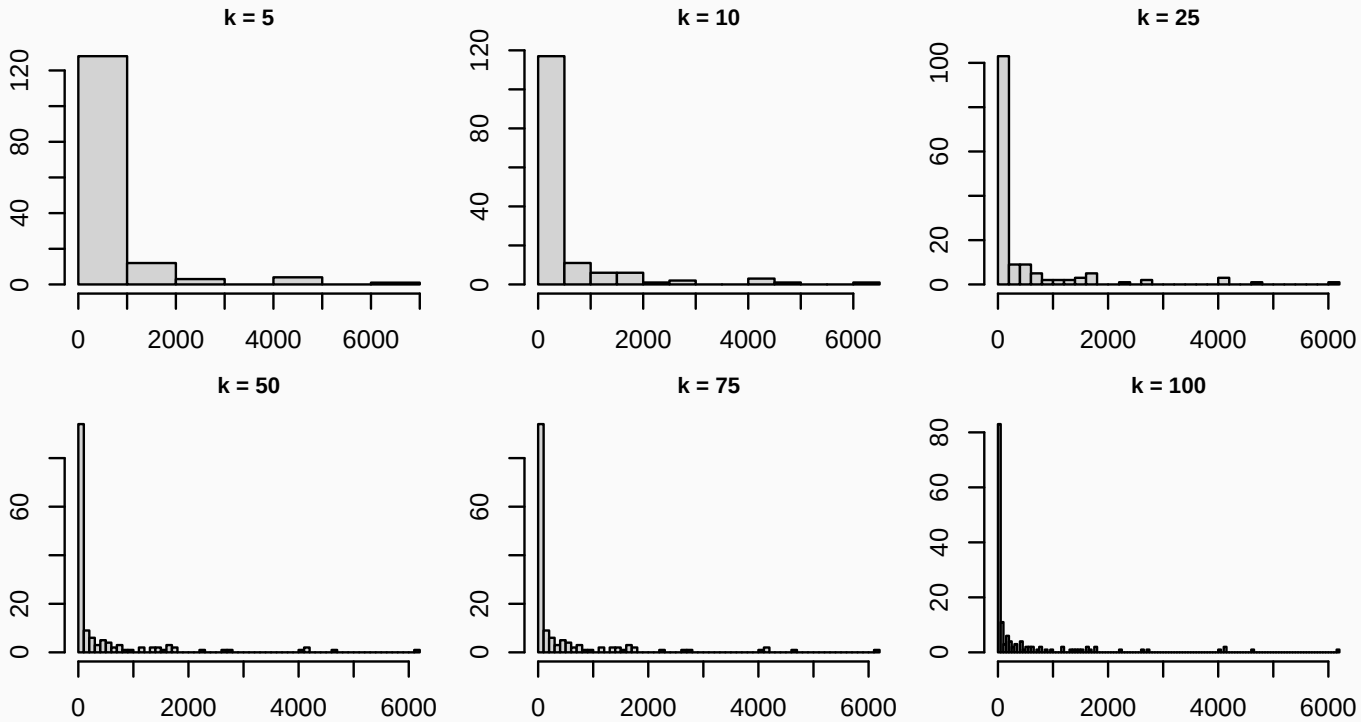
Ejercicio 1

Ejercicio 1
Calcule la probabilidad del evento $B =$ "producción de carbón entre 0 y 500 miles de ton." usando la fdp estimada por el histograma.

```
## [1] "P(0 <= X <= 500) aprox. 79.05%"
```

Problema central del histograma: el número de clases k

El histograma aproxima la fdp, pero el resultado depende críticamente de k:



Con k pequeño se **sobresuaviza**; con k grande se **sobreajusta**. ¿Existe un método que evite esta decisión arbitraria?

¿Por qué el histograma tiene este problema?

El histograma estima la densidad en cada punto x contando observaciones **dentro de un bin fijo** centrado en el intervalo de clase:

$$\hat{f}_{\text{hist}}(x) = \frac{\text{obs. en el bin que contiene } x}{n \cdot \text{ancho del bin}}$$

Consecuencias:

- La estimación es **constante** dentro de cada intervalo — salto brusco en los bordes
- Un punto x obtiene la misma densidad que todos sus vecinos de intervalo, incluso si están muy lejos de él
- Dos puntos en el borde de intervalos adyacentes reciben densidades muy distintas aunque sean casi iguales

Insight clave

El histograma es un estimador **local** muy crudo: solo considera si un punto está *dentro o fuera* del bin, sin gradación por distancia.

De bloques a promedios ponderados: la idea de la KDE

La **Estimación Kernel de la Densidad** (KDE) corrige esto: en lugar de asignar peso binario (dentro/fuera), asigna **pesos decrecientes con la distancia** a cada observación respecto al punto x que se está estimando:

Histograma	**KDE**
Peso $= 1$ si está en el bin	Peso $= K\left(\frac{x - x_i}{h}\right)$
Peso $= 0$ si no está	Peso decrece suavemente con $\lvert x - x_i \rvert$
Resultado: escalones	Resultado: curva continua y suave

La función $K(\cdot)$ que asigna esos pesos se llama **función kernel**.

Definición formal de la KDE

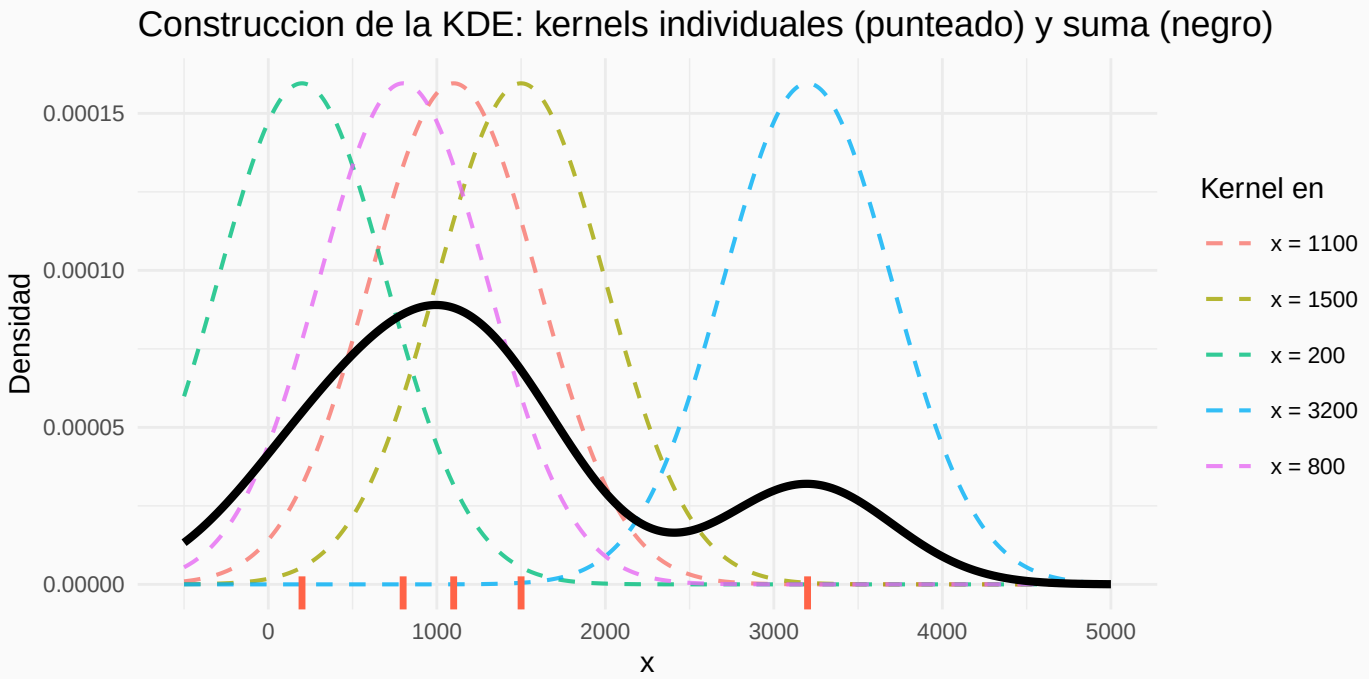
$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Donde:

- n: tamaño muestral
- $h > 0$: **ancho de banda** (*bandwidth*) — controla el suavizado
- $K(\cdot)$: función **kernel** — simétrica, $\int K(u)\, du = 1$
- x_i: observaciones de la muestra

Intuición geométrica: se coloca una campana pequeña (kernel) centrada en cada observación x_i, y la curva KDE es la **suma** de todas esas campanas, normalizada para que el área total sea 1.

Construcción visual de la KDE

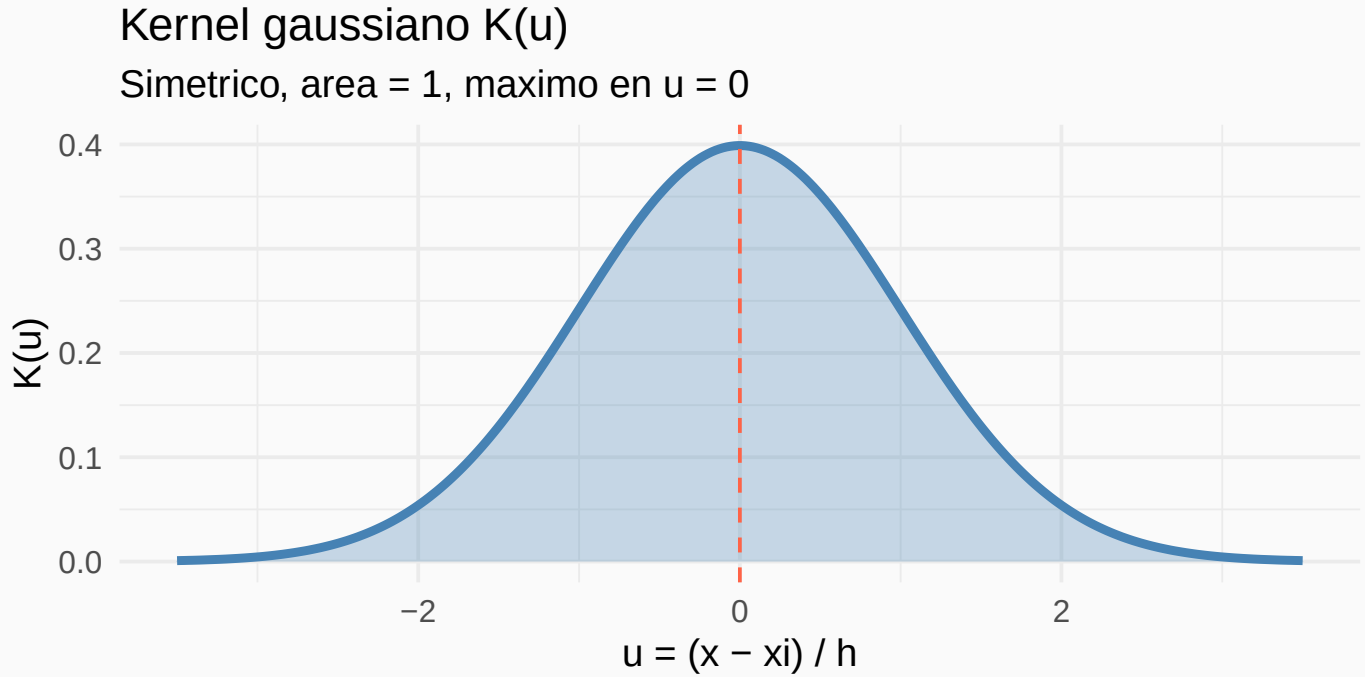


Cada línea punteada = kernel gaussiano centrado en una observación (marcas rojas en el eje x). La curva negra es su suma normalizada.

La función kernel: el kernel gaussiano

El kernel más utilizado en la práctica (y por defecto en R) es el **kernel gaussiano**:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$



El argumento $u = (x - x_i)/h$ mide cuántas unidades de ancho de banda separan el punto de estimación x de la observación x_i.

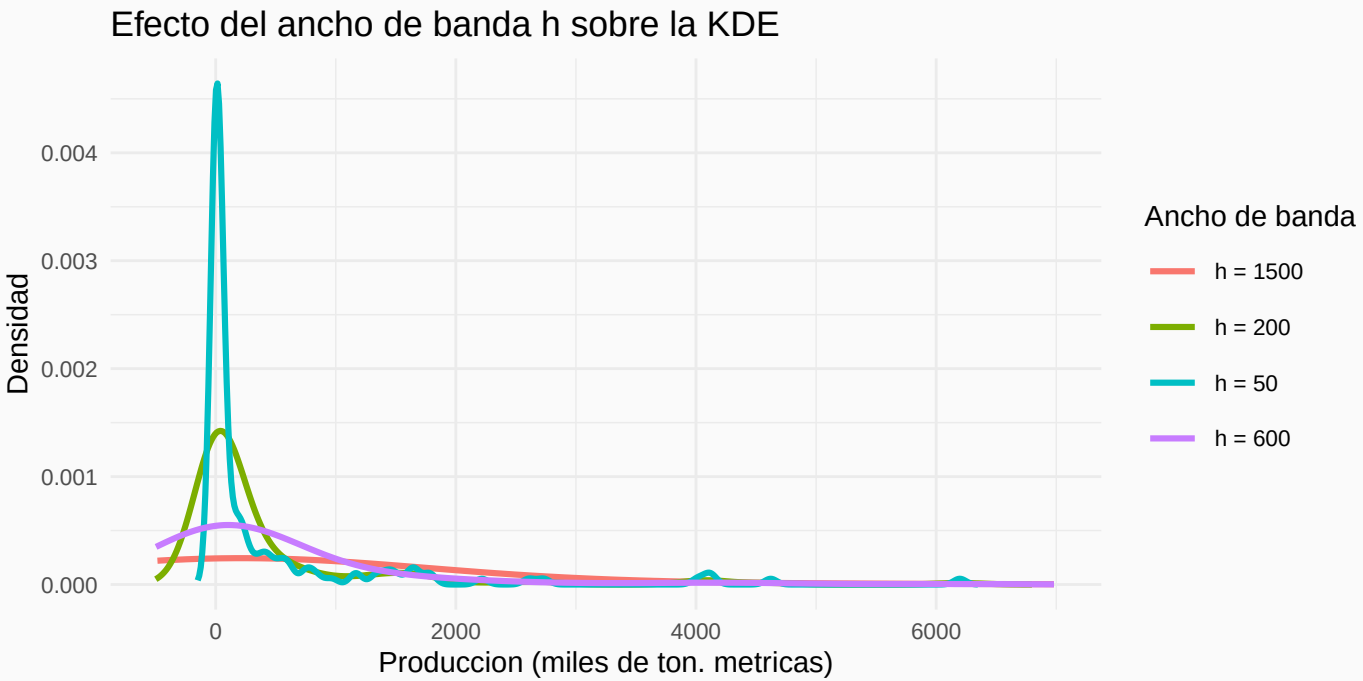
Otros kernels disponibles en R

Aunque el gaussiano es el más común, `density()` admite otros kernels mediante el argumento `kernel=`:

Kernel	**Argumento en R**	**Soporte**
Gaussiano	`"gaussian"` (defecto)	$(-\infty, +\infty)$
Epanechnikov	`"epanechnikov"`	$[-1, 1]$
Rectangular	`"rectangular"`	$[-1, 1]$
Triangular	`"triangular"`	$[-1, 1]$
Biweight	`"biweight"`	$[-1, 1]$

En la práctica, la **elección del kernel importa poco** comparada con la elección del **ancho de banda** h, que es el parámetro verdaderamente crítico de la KDE.

El ancho de banda h: el parámetro clave



h pequeño: **sobreajuste** (ruidoso, sigue cada punto individualmente). h grande: **sobresuavizado** (borra estructura real de los datos).

Selección automática del ancho de banda

El ancho de banda enfrenta el mismo dilema que k en el histograma, pero existen criterios objetivos para seleccionarlo:

Regla de Silverman (1986) — el default de R:

$$h_{\text{Silverman}} = 0.9 \cdot \min\left(s, \frac{IQR}{1.34}\right) \cdot n^{-1/5}$$

donde s es la desviación estándar muestral e IQR el rango intercuartílico.

```
# bw.nrd0() implementa la regla de Silverman en R
h_nrd0 = bw.nrd0(charcoal_prd19$Quantity)
h_sj   = bw.SJ(charcoal_prd19$Quantity)   # Sheather-Jones (1991)
round(c(Silverman = h_nrd0, Sheather_Jones = h_sj), 2)
```

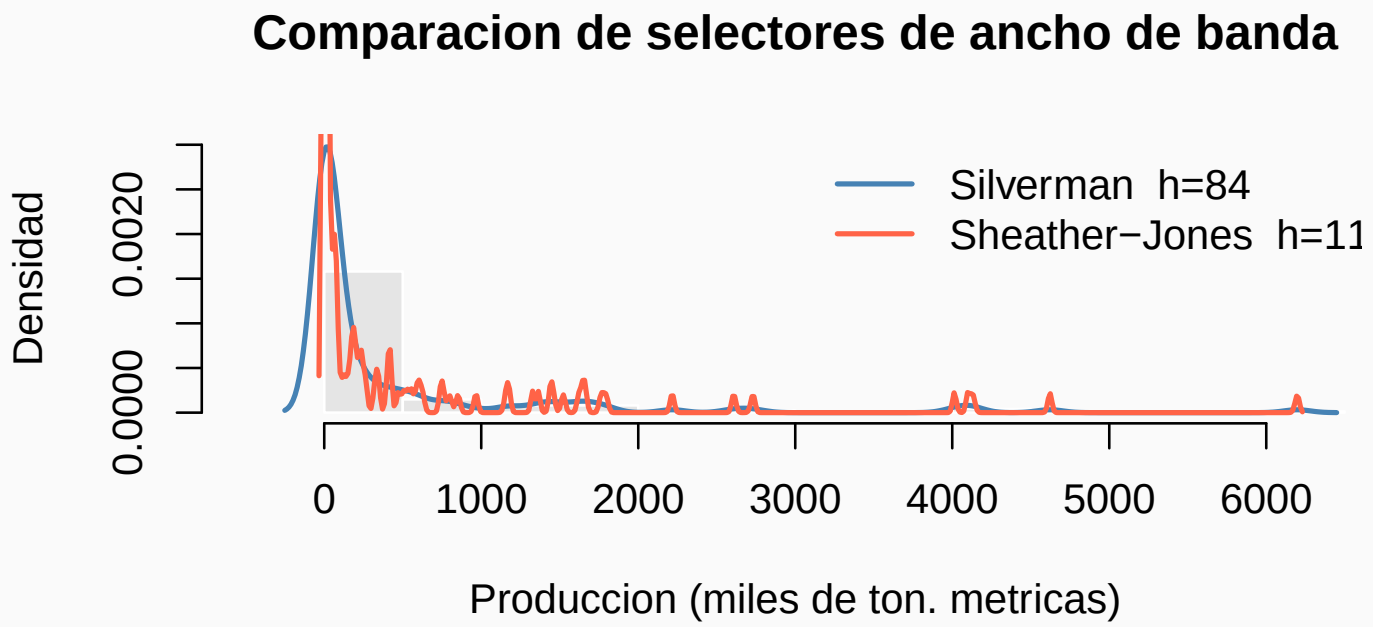
```
##      Silverman Sheather_Jones
##          83.72          11.10
```

Nota práctica

Para distribuciones muy asimétricas (como la nuestra), Silverman puede sobresuavizar. `bw.SJ()` (Sheather-Jones) es considerado el selector más robusto en la práctica general. Es bueno comparar los dos y seleccionar con criterio experto.

Comparando selectores de ancho de banda

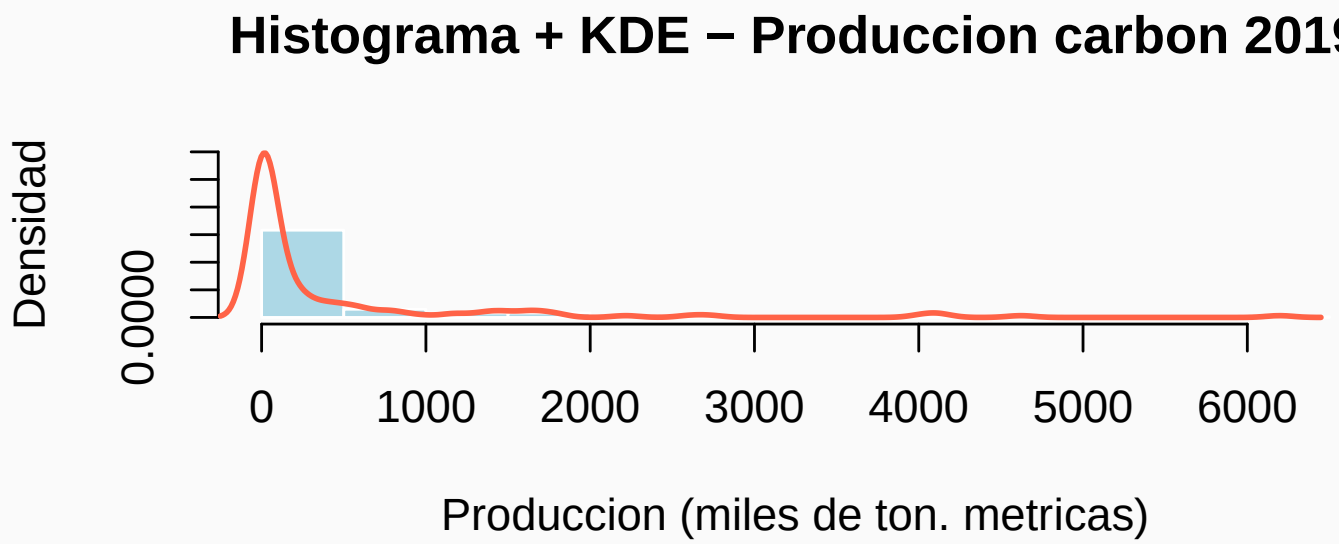
```r
x_plot = charcoal_prd19$Quantity

hist(x_plot, prob=TRUE, xlim=c(-500, 6500), ylim=c(0, 0.0030),
     main="Comparacion de selectores de ancho de banda",
     xlab="Produccion (miles de ton. metricas)", ylab="Densidad",
     col="grey90", border="white")
lines(density(x_plot, bw=h_nrd0), col="steelblue", lwd=2)
lines(density(x_plot, bw=h_sj),   col="tomato",    lwd=2)
legend("topright", bty="n", lwd=2,
       col=c("steelblue", "tomato"),
       legend=c(paste0("Silverman  h=", round(h_nrd0, 0)),
                paste0("Sheather-Jones  h=", round(h_sj, 0))))
```



KDE en R: sintaxis completa de `density()`

```r
density(
  x,                    # vector de datos
  bw     = "nrd0",      # selector de h (o valor numerico directo)
  kernel = "gaussian",  # funcion kernel
  n      = 512,         # puntos de evaluacion en la grilla
  from   = ...,         # limite inferior del dominio
  to     = ...          # limite superior del dominio
)
```

```r
hist(charcoal_prd19$Quantity, prob=TRUE,
     ylim=c(0, 0.0030), xlim=c(-4, 6500),
     main="Histograma + KDE - Produccion carbon 2019",
     xlab="Produccion (miles de ton. metricas)",
     ylab="Densidad", col="lightblue", border="white")
lines(density(charcoal_prd19$Quantity), lwd=2, col="tomato")
```



Interpretación del KDE para los datos de carbón

La curva KDE (roja) revela características que el histograma puede enmascarar:

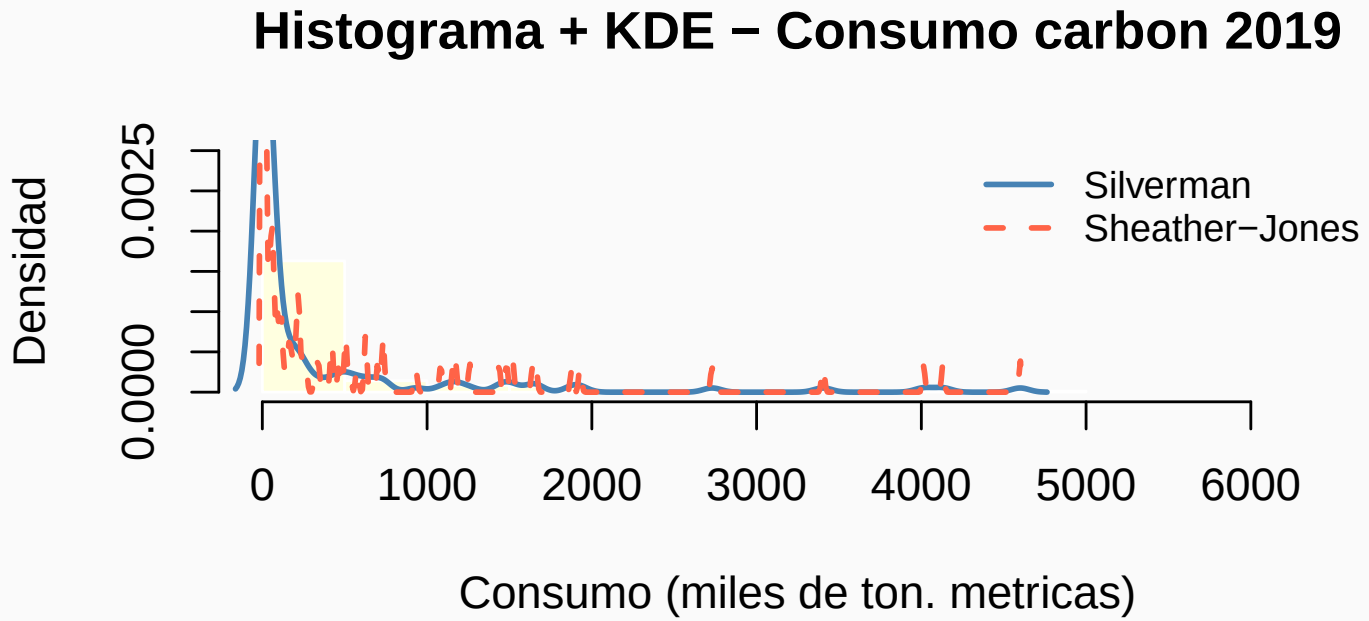
- El **primer intervalo** del histograma mostraba densidad homogénea artificial — la KDE muestra alta concentración real cerca de cero (alta asimetría positiva)
- La **probabilidad del evento** A estaba sobreestimada con el histograma
- La KDE asigna densidad a valores negativos: limitación para variables con soporte $[0, +\infty)$

Conexión con la Sesión 1

La producción de carbón presentaba una distribución altamente **sesgada a la derecha** — la KDE captura esto mejor que el histograma con pocos intervalos. Es consistente con la asimetría y los outliers observados con boxplot y medidas de dispersión en la Sesión 1.

Ejercicio 2

Para `charcoal_chh19`, elabore el histograma de densidad junto con la KDE para la variable `Quantity`. ¿Es similar a la estimación para `charcoal_prd19`? Pruebe también con `bw.SJ()` y compare visualmente ambos selectores.



Ejercicio 3 (reto)

Ejercicio 3 — Reto

Calcule el área bajo la curva KDE para `charcoal_chh19`. Aproxime el área como suma de rectángulos usando los objetos retornados por `density()`. Recuerde que `density()` devuelve una lista con componentes `$x` (grilla de puntos) y `$y` (densidad estimada).

```
## [1] "Area total bajo la curva: 0.9991"
```

¿Para qué sirven las funciones de densidad?

Las fdp $f(x \mid \theta)$ resumen el **comportamiento aleatorio** de la variable y son la base de la estadística inferencial:

Propiedad	**Expresión**
Area total $= 1$	$\int_{-\infty}^{+\infty} f(x)\,dx = 1$
Probabilidad intervalo	$P(a \leq X \leq b) = \int_a^b f(x)\,dx$
Simetria (Normal)	$P(X \leq \mu) = P(X \geq \mu) = 0.5$

- Valores **cercanos a la media** tienen mayor **verosimilitud**
- Son la base de las **pruebas de hipótesis** e **intervalos de confianza**

5. Pensamiento bivariado

Del escalar al vector

Hasta ahora cada observación era un **escalar**: $x_i \in \mathbb{R}$

$$x_2 = 1159.8 \quad \text{(Angola: produccion de carbon)}$$

```
charcoal_prd19$Quantity[2]
```

```
## [1] 1159.8
```

```
charcoal_prd19$Country_Area[2]
```

```
## [1] "Angola"
```

En el análisis **bivariado**, cada observación es un **vector de dimensión 2**:

$$(x_i, y_i) = (\text{produccion}_i,\ \text{consumo}_i)$$

Construcción del dataset bivariado

```
charcoal_bivar =
  charcoal_prd19 %>%
    select(Country_Area, Charcoal_Production = Quantity) %>%
  inner_join(
    charcoal_chh19 %>%
      select(Country_Area, Charcoal_Consumption = Quantity),
    by = "Country_Area")

charcoal_bivar %>% filter(Country_Area == "Argentina")
```

```
## # A tibble: 1 x 3
##   Country_Area Charcoal_Production Charcoal_Consumption
##   <chr>                      <dbl>                <dbl>
## 1 Argentina                    411                  247
```

Argentina produce **411** pero consume **247** miles de ton. — exporta carbón.

Notación estadística bivariada

La muestra bivariada de tamaño n se expresa como:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

donde:

- $x_i =$ producción de carbón del país i (miles de ton. métricas)
- $y_i =$ consumo de carbón del país i (miles de ton. métricas)

Distribución conjunta vs. distribución marginal
Marginal: comportamiento de X o de Y de manera independiente

Conjunta: comportamiento de (X, Y) en conjunto — captura la **relación entre variables**

El análisis bivariado busca entender esta relación conjunta.

Repaso: Covarianza

De la Sesión 1 recordamos que la **covarianza** mide la dirección de la relación lineal entre dos variables:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Signo	**Interpretación**
$\text{Cov} > 0$	Cuando X sube, Y tiende a subir
$\text{Cov} < 0$	Cuando X sube, Y tiende a bajar
$\text{Cov} \approx 0$	Sin relación lineal aparente

Limitación: la covarianza depende de las unidades de medida.\ **Solución**: la correlación de Pearson $r \in [-1, 1]$.

Correlación de Pearson

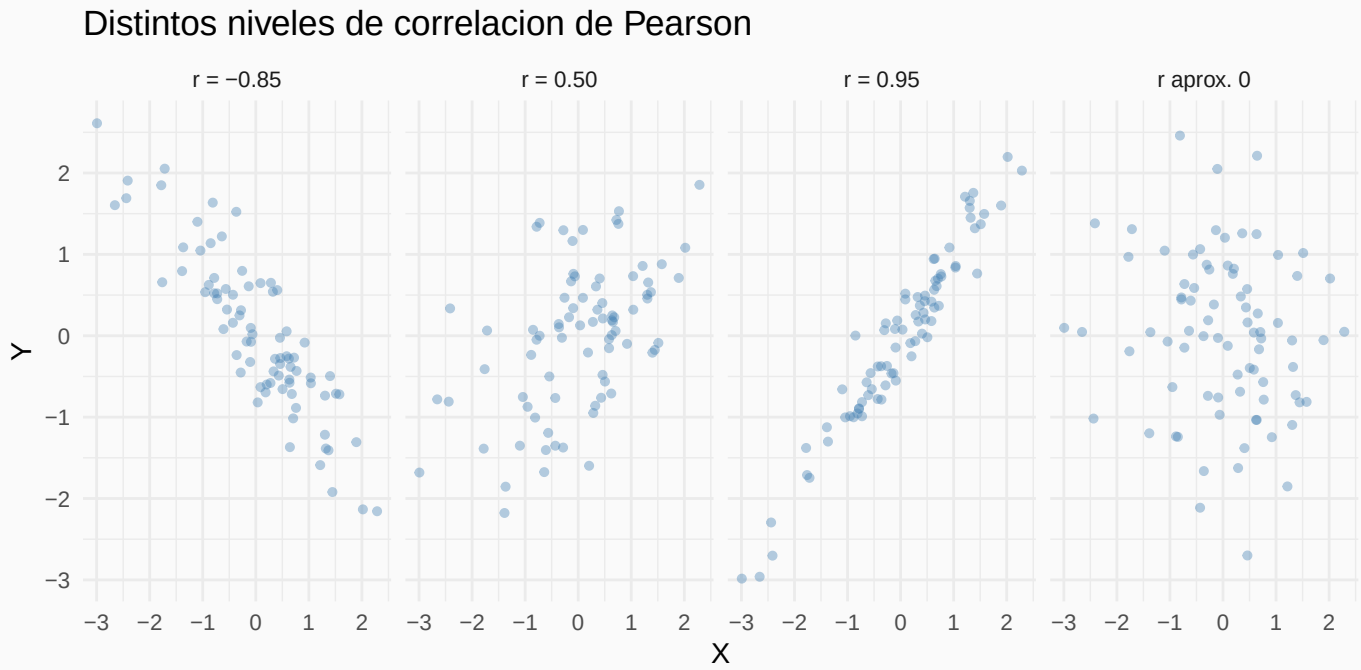
La **correlación de Pearson** estandariza la covarianza por el producto de las desviaciones estándar, haciéndola **adimensional**:

$$r = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \in [-1, 1]$$

```
# Funciones nativas de R para covarianza y correlacion
cov_xy = cov(charcoal_bivar$Charcoal_Production,
             charcoal_bivar$Charcoal_Consumption)
cor_xy = cor(charcoal_bivar$Charcoal_Production,
             charcoal_bivar$Charcoal_Consumption)
round(c(Covarianza = cov_xy, Correlacion = cor_xy), 3)
```

```
##  Covarianza Correlacion
##  752861.088       0.857
```

Interpretación de distintos niveles de correlación



$r = 1$: perfecto positivo $r = -1$: perfecto negativo $r = 0$: sin relacion lineal

Correlación en los datos de carbón

```
ggplot(charcoal_bivar,
       aes(x=Charcoal_Production, y=Charcoal_Consumption)) +
  geom_point(alpha=0.5, color="steelblue") +
  geom_smooth(method="lm", color="tomato", se=FALSE, linewidth=0.8) +
  annotate("text", x=4200, y=350,
           label=paste0("r = ", round(cor_xy, 3)),
           size=3.5, color="tomato") +
  theme_minimal() +
  labs(x="Produccion (miles ton.)", y="Consumo (miles ton.)",
       title="Correlacion de Pearson - carbon 2019")
```

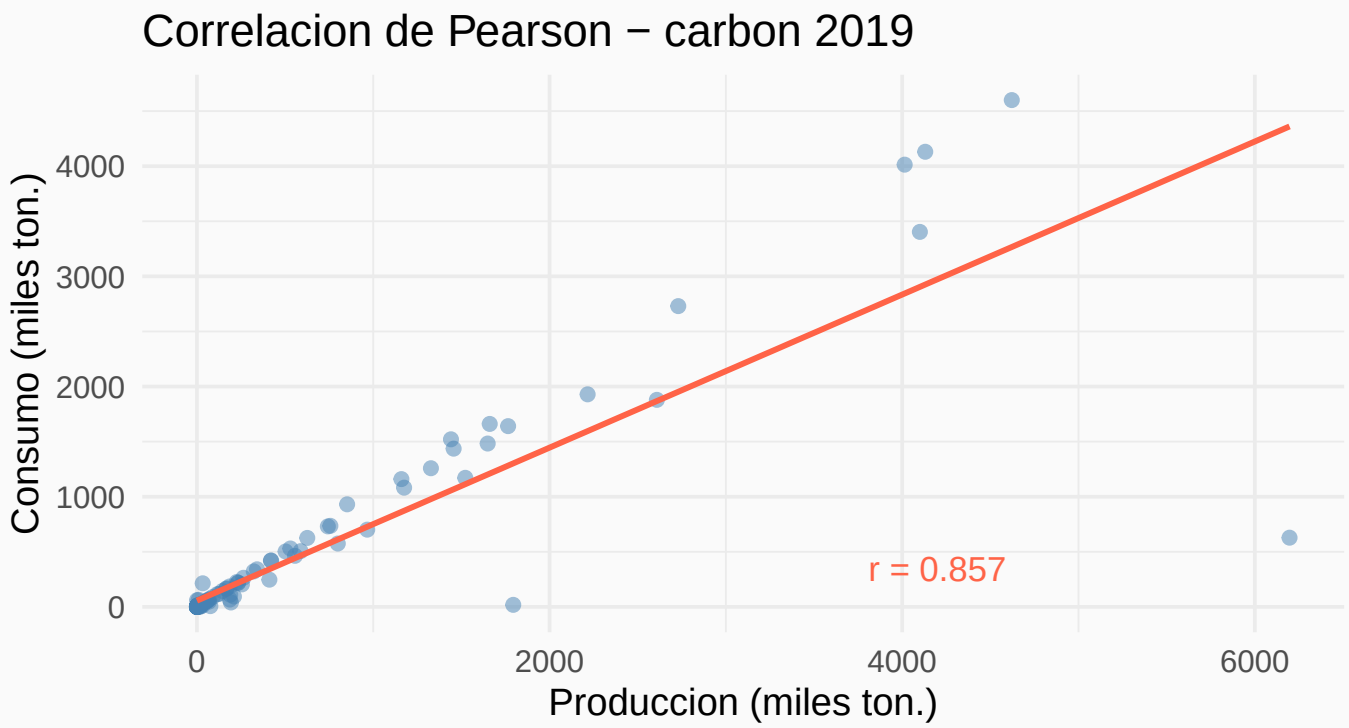
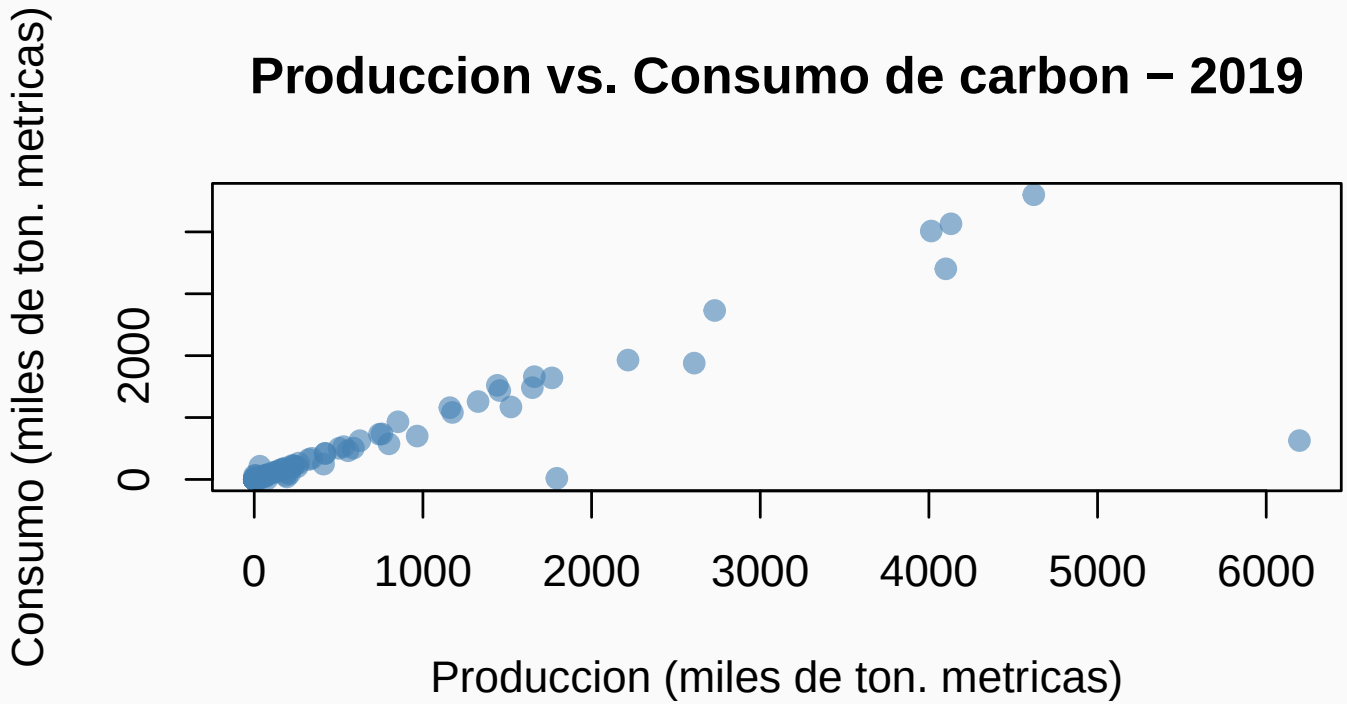


Diagrama de dispersión

```r
# Version con gráficos base
plot(charcoal_bivar$Charcoal_Production,
     charcoal_bivar$Charcoal_Consumption,
     main = "Produccion vs. Consumo de carbon - 2019",
     xlab = "Produccion (miles de ton. metricas)",
     ylab = "Consumo (miles de ton. metricas)",
     pch  = 19, col = adjustcolor("steelblue", 0.6))
```



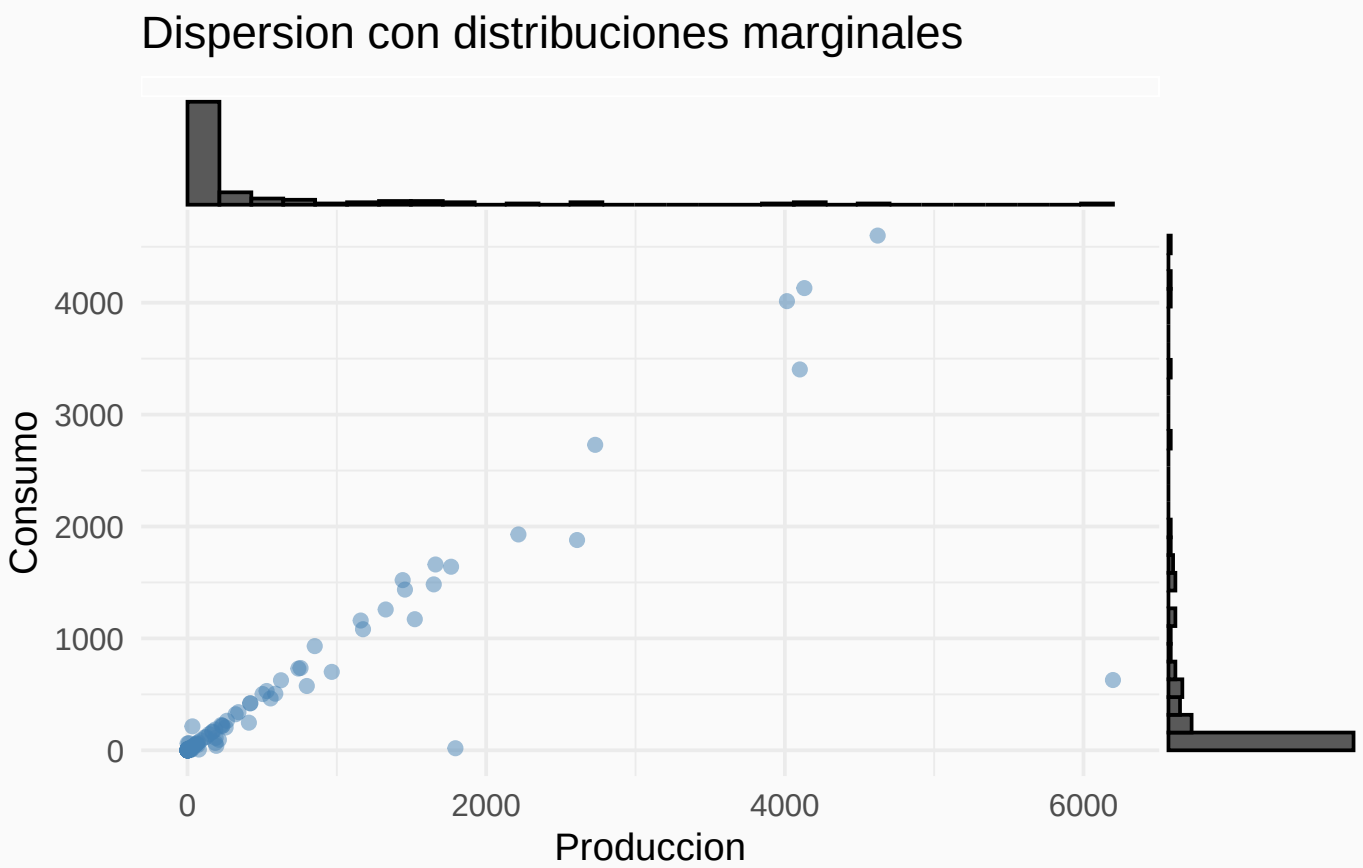
¿Qué detectamos en un diagrama de dispersión?

Patron	Descripcion
Lineal positivo/negativo	Tendencia clara ($\|r\|$ alto)
No lineal (cuadratico)	Relacion compleja
Nube sin forma	Sin relacion lineal
Embudo (heterosced.)	Varianza no constante
Puntos aislados	Outliers bivariados

Los datos de carbón muestran una **relación lineal positiva** clara, con cierta **heteroscedasticidad** para valores altos de producción.

Diagrama de dispersión con distribuciones marginales

```
library(ggExtra)
g_base = ggplot(charcoal_bivar,
                aes(x=Charcoal_Production, y=Charcoal_Consumption)) +
  geom_point(alpha=0.5, color="steelblue") +
  theme_minimal() +
  labs(x="Produccion", y="Consumo",
       title="Dispersion con distribuciones marginales")
ggMarginal(g_base, type="histogram")
```



Histogramas marginales: interpretación

El gráfico combina dos perspectivas complementarias:

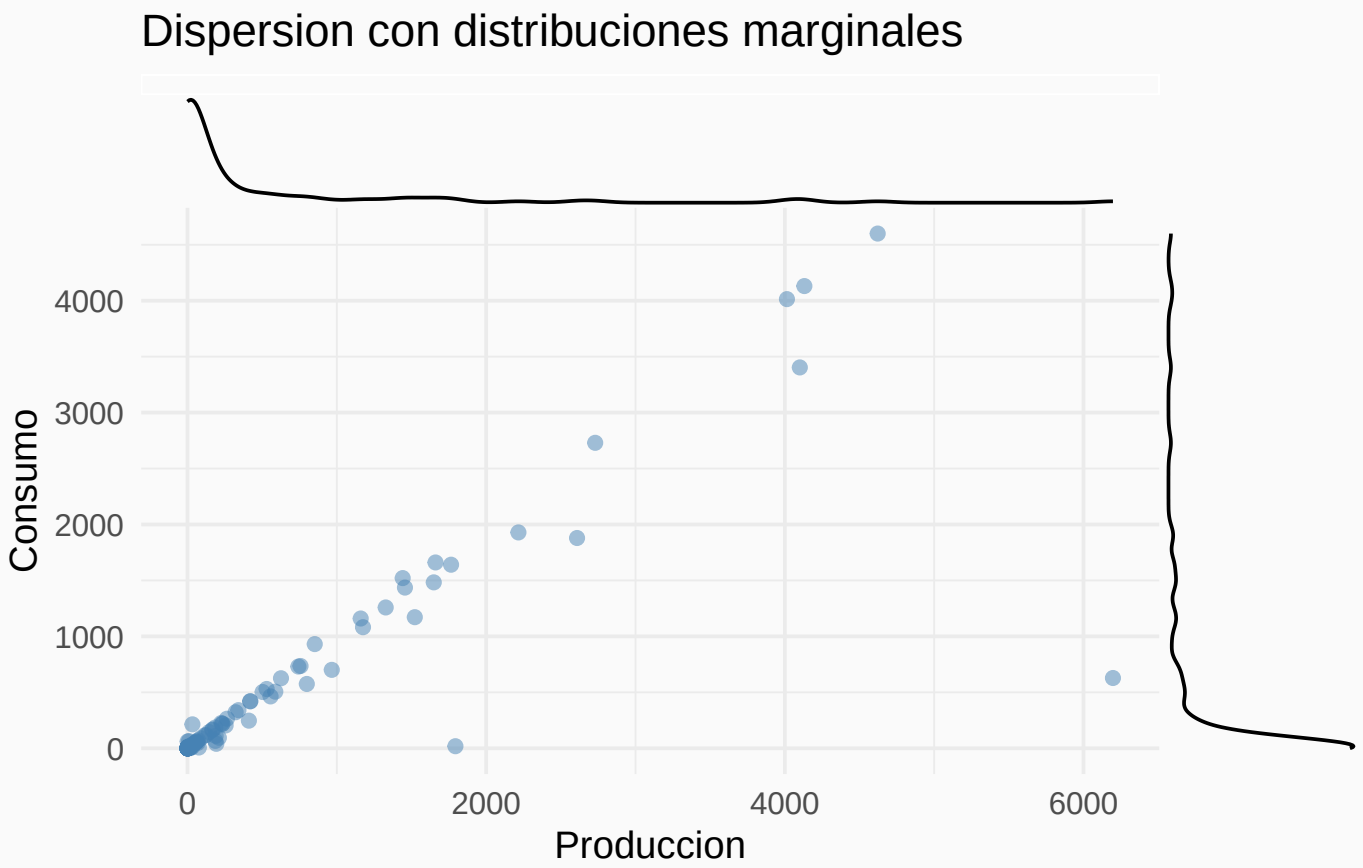
- **Centro**: relación **conjunta** (X, Y) — pensamiento bivariado
- **Margen superior**: distribución **marginal** de Y (consumo)
- **Margen lateral**: distribución **marginal** de X (producción)

Conexión con la Sesión 1

Ambas distribuciones marginales son altamente **asimétricas a la derecha** (como vimos con histograma y boxplot en la Sesión 1): la mayoría de países presentan bajos niveles de producción y consumo, y unos pocos exhiben valores muy elevados.

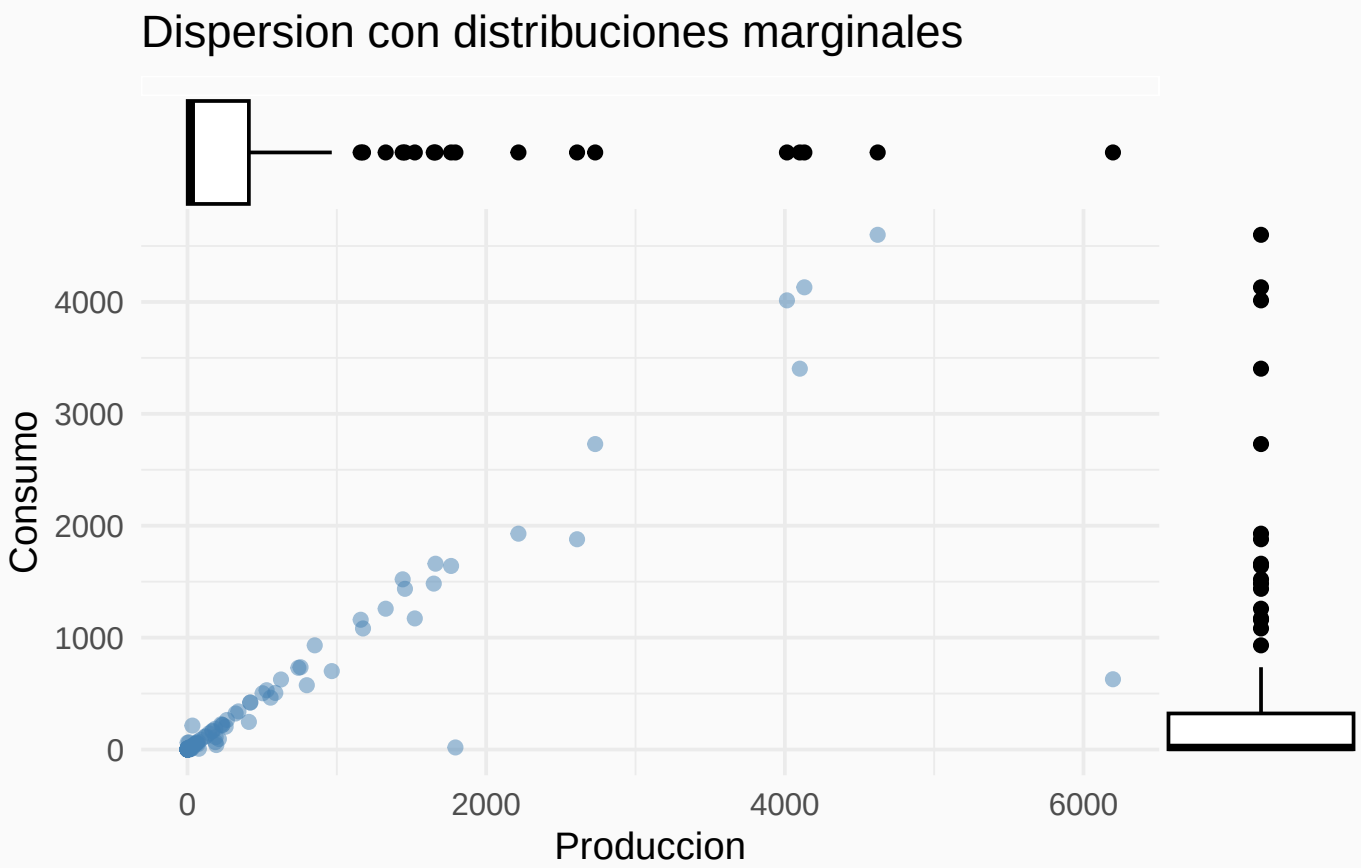
Dispersión con KDE marginal

```
ggMarginal(g_base, type="density")
```



Dispersión con boxplots marginales

```
ggMarginal(g_base, type="boxplot")
```



Los boxplots confirman la presencia de **valores atípicos** en ambas variables, consistente con la fuerte asimetría observada en la Sesión 1.

6. Síntesis

Conclusiones de la Sesión 2

Probabilidad y funciones de densidad:

- Las variables aleatorias formalizan la incertidumbre en los datos
- La **distribución normal** es el modelo central del curso, indexada por (μ, σ); su ubicuidad se explica por el TCL
- El **histograma** y la **KDE** son estimadores empíricos de la fdp subyacente
- El área bajo la curva cuantifica **probabilidades**

Pensamiento bivariado:

- Las observaciones se extienden de escalares a **vectores** (x_i, y_i)
- La **correlación de Pearson** r cuantifica la intensidad de la relación lineal
- El **diagrama de dispersión** es la herramienta gráfica fundamental
- Las distribuciones **marginales** complementan la visión conjunta

Conexión entre sesiones

Sesion 1	Sesion 2	Sesion 3
Distribución univariada (hist., boxplot, medidas desc.)	Funcion de densidad (Normal, KDE)	Distribución Normal Multivariada $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Covarianza (escalar)	Pensamiento bivariado (r, diagrama de dispersion)	Matriz de covarianza $\boldsymbol{\Sigma}$ (estructura)

Puente hacia la Sesión 3

¿Qué sigue? Distribución Normal Multivariada

En la Sesión 3 extenderemos todo lo aprendido al caso **multivariado** ($p > 2$ variables):

Univariado	**Multivariado**
$X \sim \mathcal{N}(\mu, \sigma^2)$	$\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Media $\mu \in \mathbb{R}$	Vector de medias $\boldsymbol{\mu} \in \mathbb{R}^p$
Varianza $\sigma^2 > 0$	Matriz de covarianza $\boldsymbol{\Sigma}$ $(p \times p)$
Campana en $\mathbb{R}$	Elipsoide en $\mathbb{R}^p$
Regla $\mu \pm 3\sigma$	Distancia de Mahalanobis

La fdp Normal Multivariada

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

- $\boldsymbol{\mu}$: vector de medias (centro de la distribución)
- $\boldsymbol{\Sigma}$: matriz de covarianza (forma y orientación del elipsoide)
- $(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$: **distancia de Mahalanobis** — generalización de la distancia euclidea

Conexion hacia adelante

El diagrama de dispersión bivariado de esta sesión es la **proyeccion 2D** de esta distribución — hemos preparado el terreno visual y conceptual para la Sesion 3.

Preguntas para la Sesión 3

Al concluir la Sesión 2 es natural preguntarse:

1. ¿Cómo se generaliza la campana normal a 3, 10 o 100 dimensiones?
2. ¿Cómo resumimos la relación entre **múltiples** variables con $\boldsymbol{\Sigma}$?
3. ¿Qué rol juega la **forma** de $\boldsymbol{\Sigma}$ en la distribución conjunta?
4. ¿Cómo detectamos **outliers multivariados** (invisibles en las marginales)?
5. ¿Cómo evaluamos si un conjunto de datos sigue una distribución normal multivariada?

Estas preguntas guiarán la **Sesión 3: Distribución Normal Multivariada**.

Anexo

Función de distribución acumulada (FDA)

La FDA $F(x)$ calcula probabilidades acumuladas hacia la izquierda:

$$F(x \mid \mu, \sigma) = \int_{-\infty}^{x} f(t \mid \mu, \sigma)\, dt = P(X \leq x)$$

En R se calcula directamente con `pnorm(x, mean, sd)`:

```
x = seq(-4, 4, length=200)
# Nombres de columna sin caracteres especiales
df_fda = tibble(x=x,
  "sigma=0.5" = pnorm(x, 0, 0.5),
  "sigma=1"   = pnorm(x, 0, 1),
  "sigma=2"   = pnorm(x, 0, 2)) %>%
  pivot_longer(-x, names_to="Parametro", values_to="Fx")
ggplot(df_fda, aes(x=x, y=Fx, color=Parametro)) +
  geom_line(linewidth=1) + theme_minimal() +
  labs(title="FDA Normal con distintos sigma (mu=0)", y="F(x)")
```

FDA Normal con distintos sigma (mu=0)

Referencias

- Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2012). *Introducción a la probabilidad y estadística*. Cengage Learning.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6a ed.). Pearson.
- Wickham, H., & Grolemund, G. (2017). *R for Data Science*. O'Reilly Media.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.