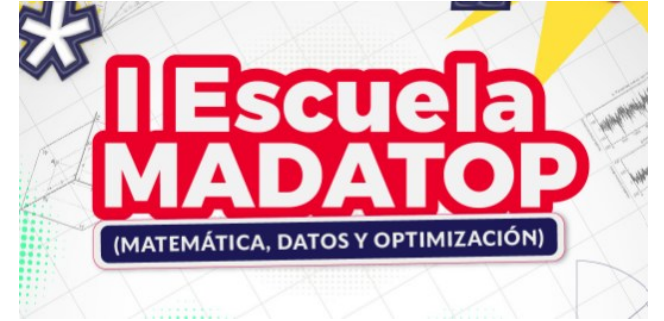




UNL. FACULTAD DE  
INGENIERÍA QUÍMICA

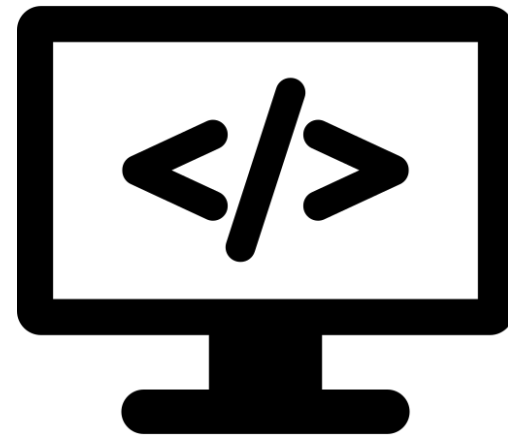


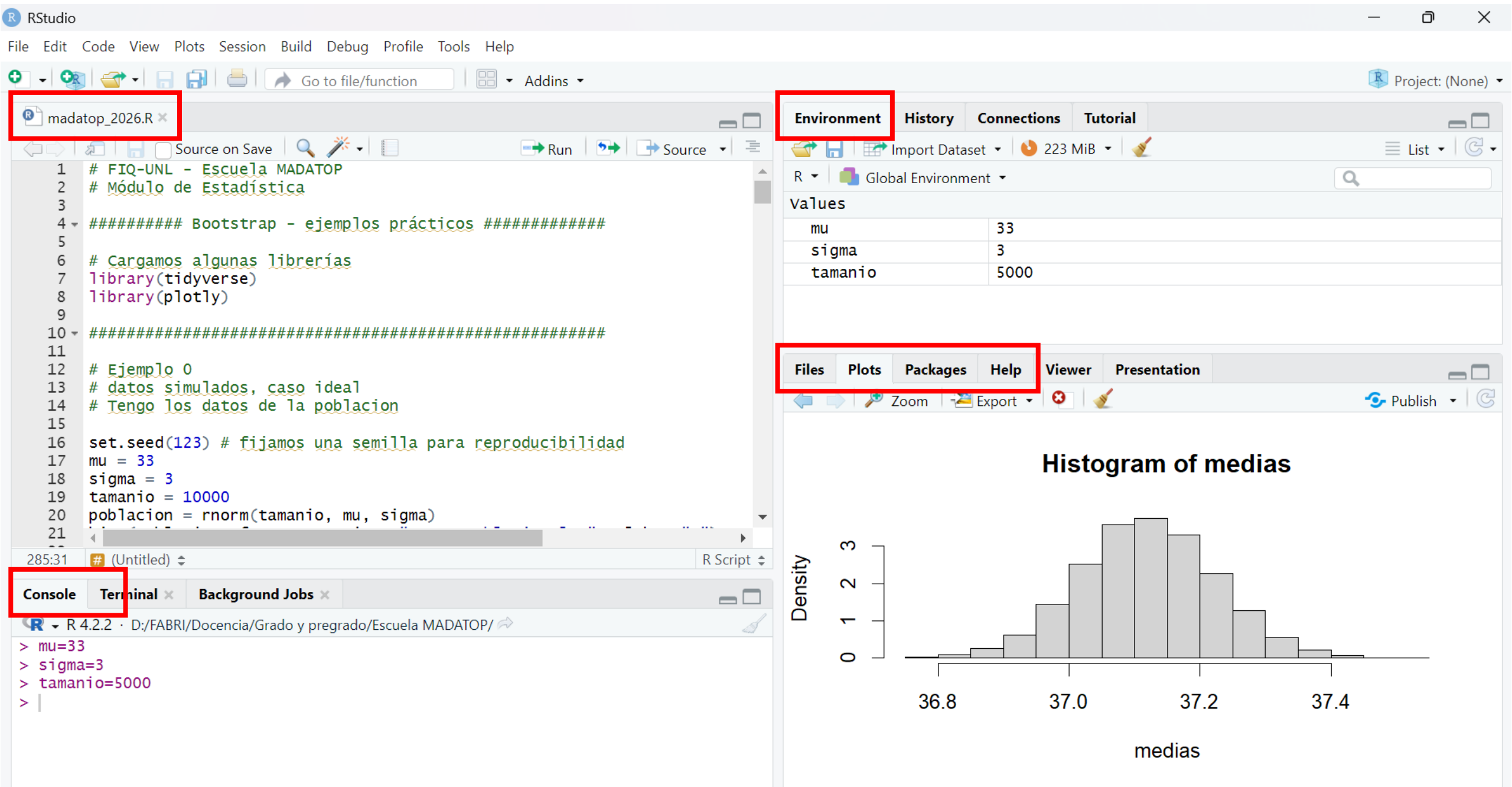
# *Bootstrap* – Parte 2

---

Curso de Estadística – MADATOP 2026

Andrea Bergesio – Fabricio Chiappini – Stefania D'lorio





Un paréntesis entre la clase de la mañana y  
ahora...



# Variable aleatoria y distribución (modelo)

Recordemos notación y algunas definiciones

- Parámetro  $\theta$ : propiedad de la población (en general desconocida). Ejemplo: media poblacional  $\mu$
- Estadístico: una cantidad que se calcula a partir de los datos muestrales. Ejemplo:  $\bar{X}$
- Estimador  $\hat{\theta}$ : un estadístico específico que se usa para estimar un parámetro. Ejemplo:  $\bar{X} = \hat{\mu}$



Pregunta: ¿cuáles son fijos y cuáles variables aleatorias? ¿por qué?

- Función de densidad/probabilidad acumulada de la variable aleatoria  $X$

$$X \sim F_{\theta}$$

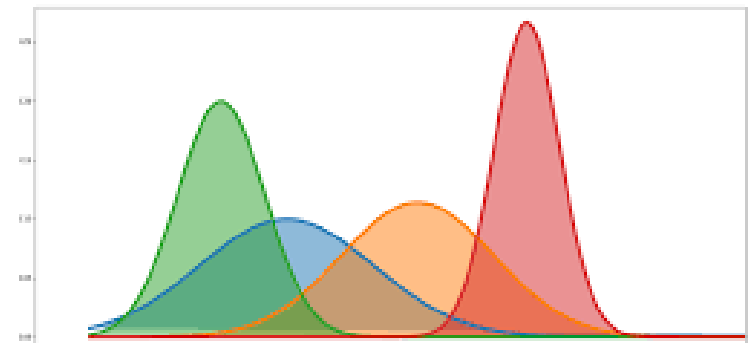
- Función de densidad/probabilidad

$$X \sim f(\theta)$$

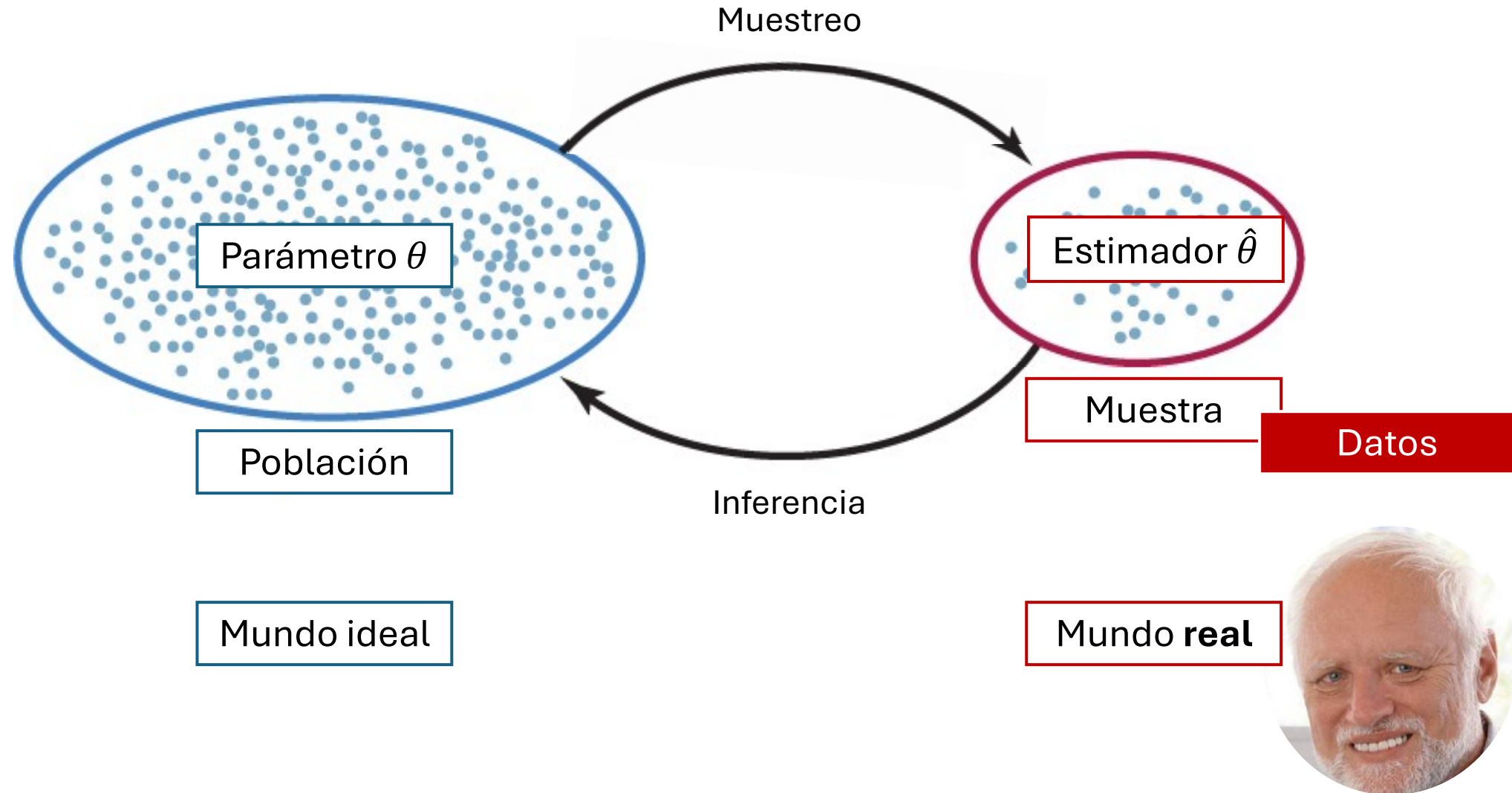


¿Una distribución famosa que conozcan?

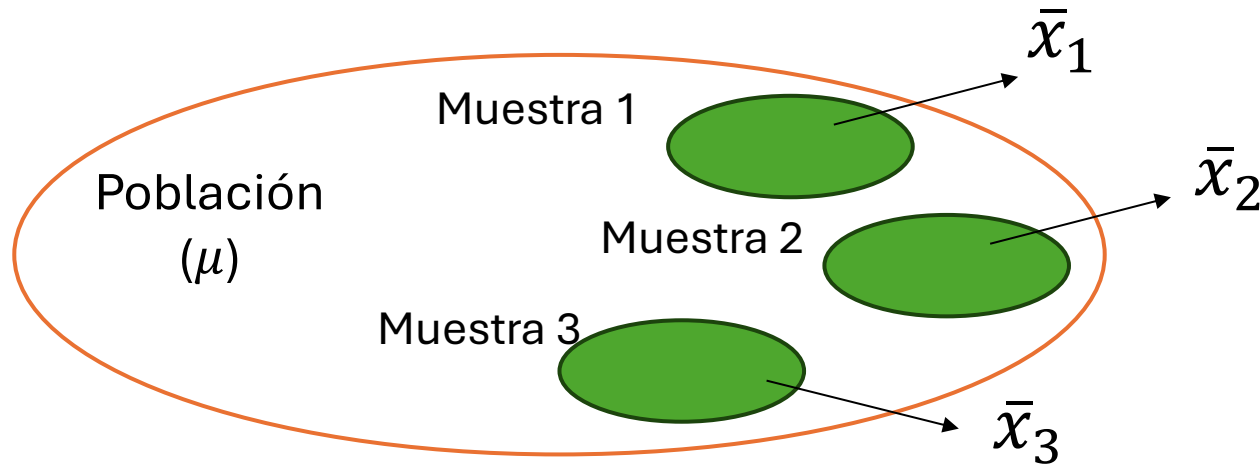
$$X \sim \mathcal{N}(\mu, \sigma)$$



# ¿Qué hacemos en estadística?



# Concepto de distribución muestral

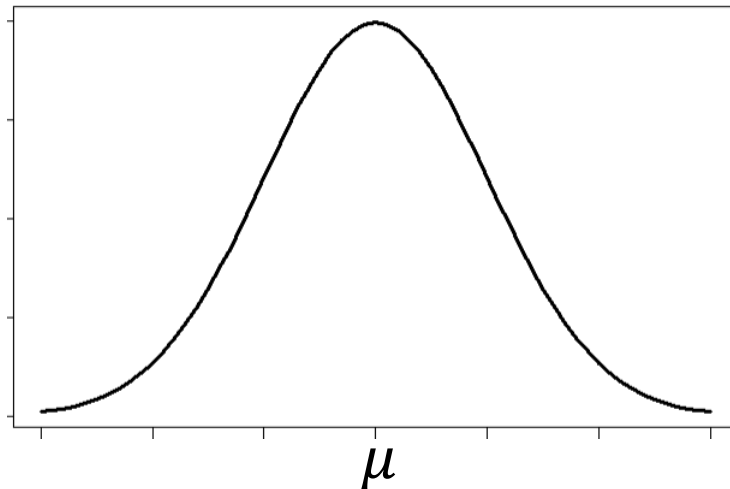


$\bar{X}$  es una variable aleatoria

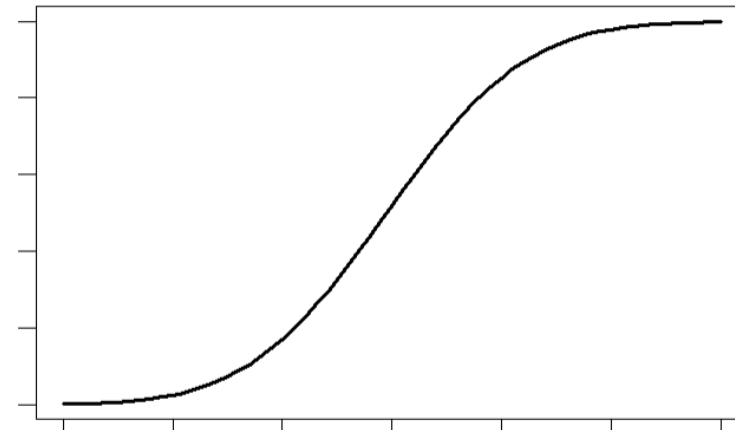
Si  $X$  es una variable aleatoria normal  $X \sim \mathcal{N}(\mu, \sigma)$ , entonces  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  tiene distribución

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Función de densidad

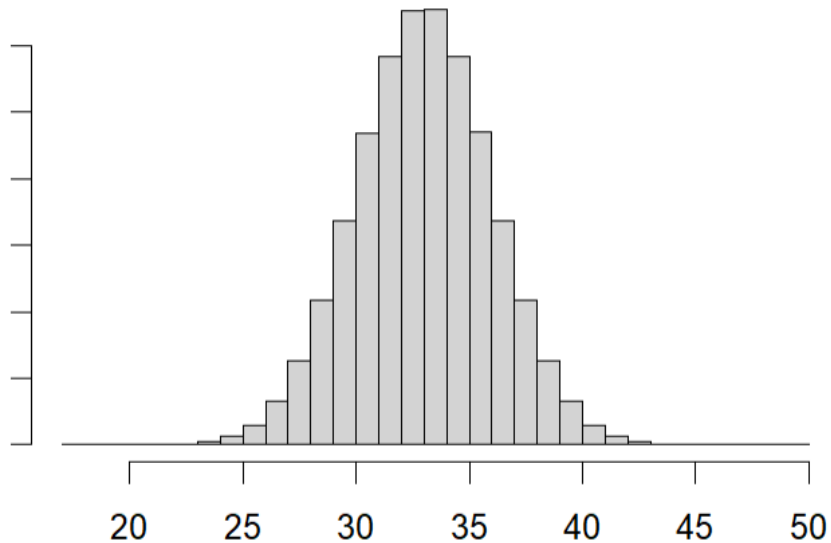


Función de densidad acumulada



# Simulemos datos poblacionales en R

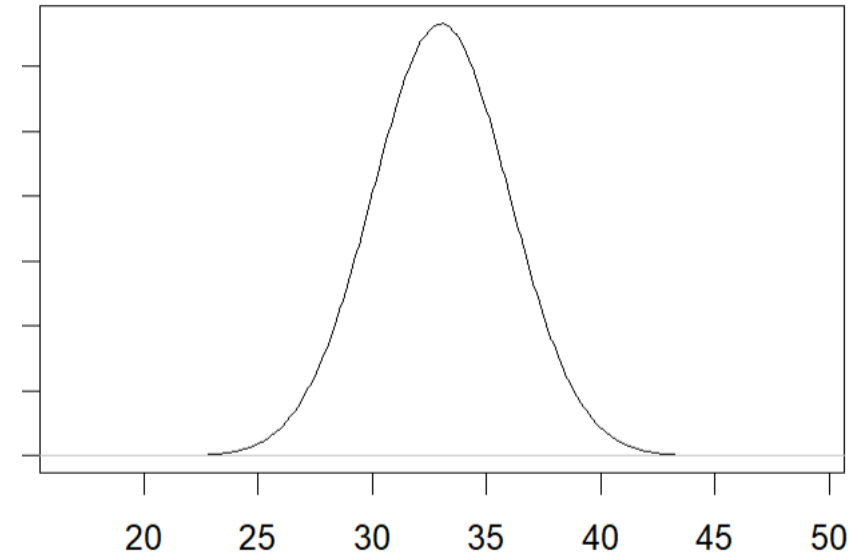
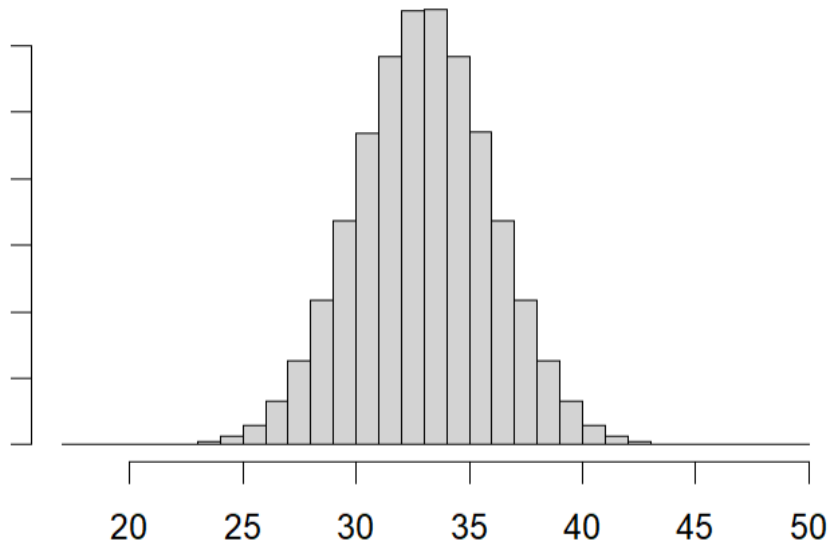
```
set.seed(123) # fijamos una semilla para reproducibilidad
mu = 33
sigma = 3
tamanio = 10000000
poblacion = rnorm(tamanio, mu, sigma)
→ hist(poblacion, freq = F, main = "Datos poblacionales", xlab = "X")
plot(density(poblacion))
```





# Simulemos datos poblacionales en R

```
set.seed(123) # fijamos una semilla para reproducibilidad
mu = 33
sigma = 3
tamanio = 10000000
poblacion = rnorm(tamanio, mu, sigma)
hist(poblacion, freq = F, main = "Datos poblacionales", xlab = "X")
→ plot(density(poblacion))
```



¿Necesitamos hacer inferencia?

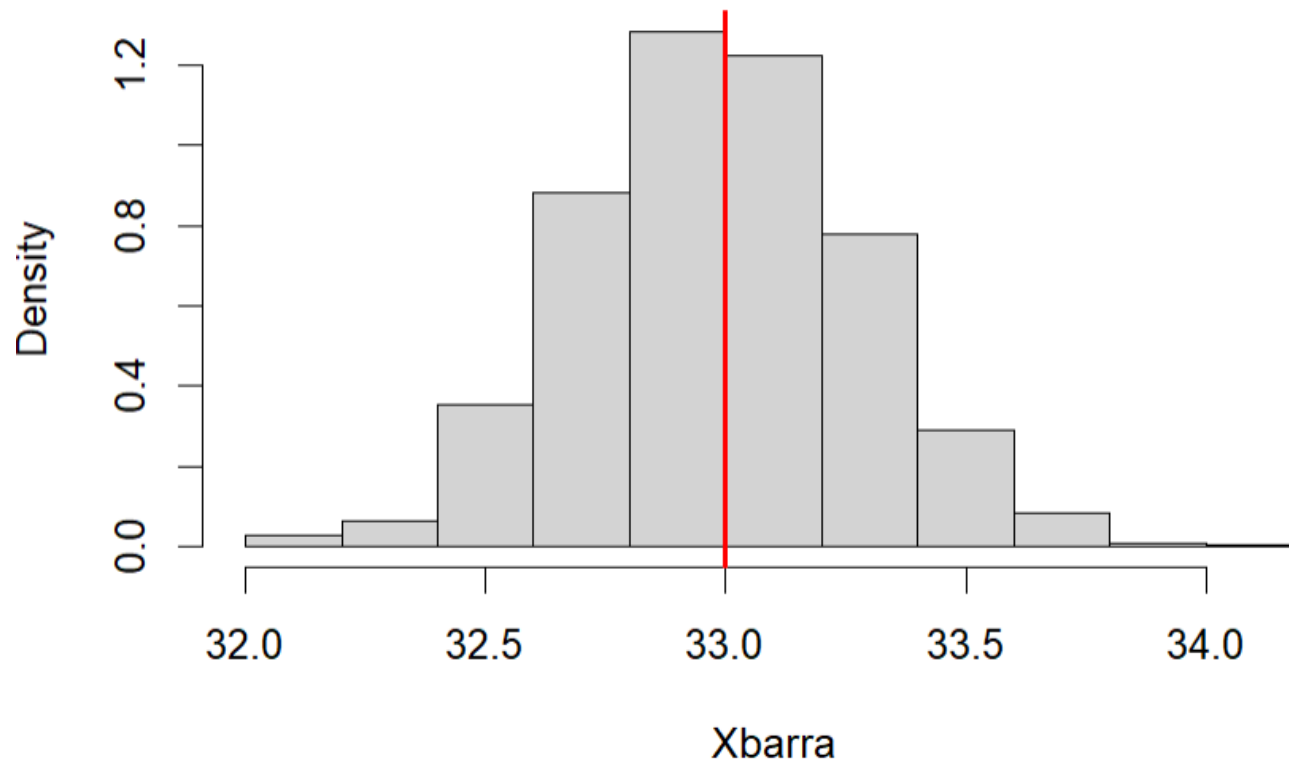
# Mirando la distribución muestral de $\bar{X}$ en R

```
nreps = 1000
n = 100
medias.muestrales = NULL

for (i in 1:nreps){
  medias.muestrales[i] = mean(sample(poblacion, n, replace = F))
}
```

# Distribución muestral de $\bar{X}$

Distribución muestral de Xbarra



- ¿Es la distribución de  $X$ ?
- ¿Qué aspecto tiene?

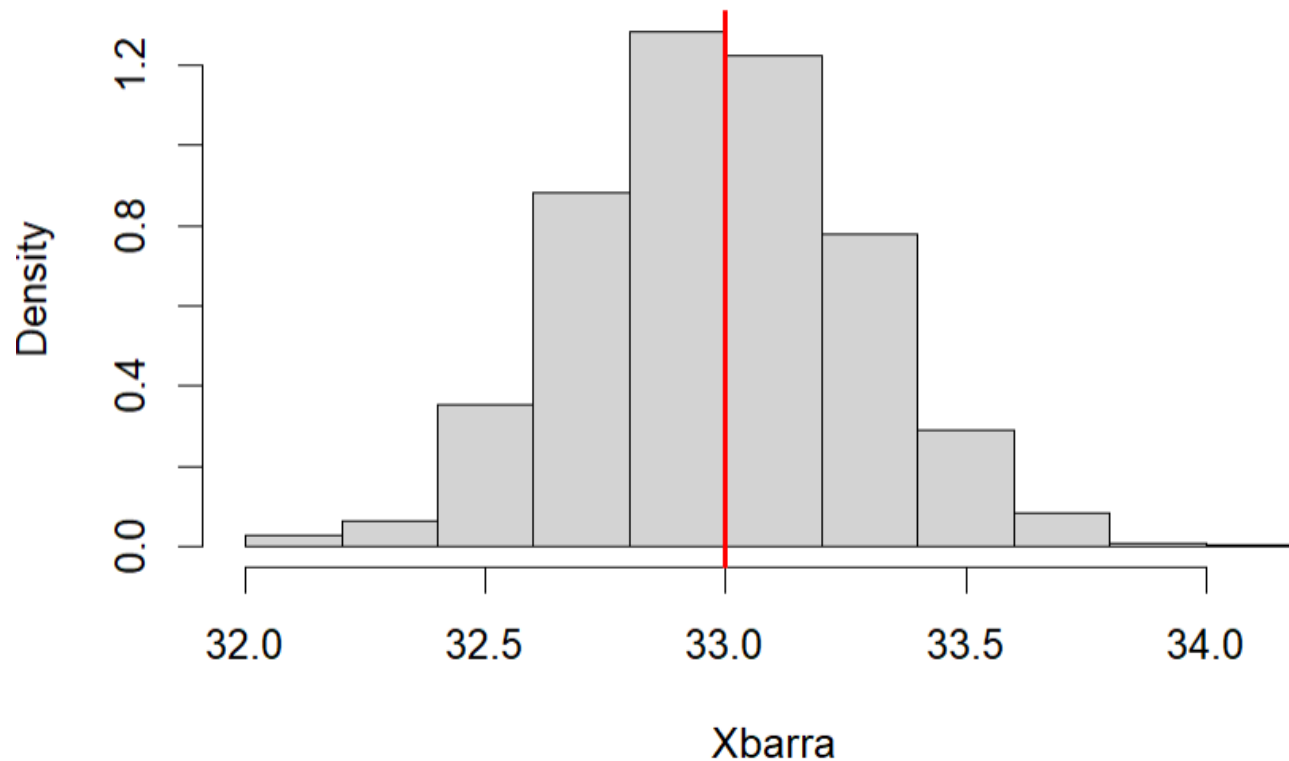
Recordemos:  $\mu = 33$ ,  $\sigma = 3$

## *Ejercicio 1*

1. Calcule la media de la distribución muestral de  $\bar{X}$ , ¿qué observa?
2. Calcule el desvío estándar de la distribución muestral de  $\bar{X}$ , ¿qué observa?
3. Repita la simulación anterior pero variando el tamaño muestral  $n$ , ¿qué observa?

# Distribución muestral de $\bar{X}$

Distribución muestral de Xbarra



- ¿Es la distribución de X?
- ¿Qué aspecto tiene?

Recordemos:  $\mu = 33, \sigma = 3$

```
> mean(medias.muestrales)
[1] 33.00266
> sd(medias.muestrales)
[1] 0.2909427
```

Resultado teórico

$$E(\bar{X}) = \mu$$

$$sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

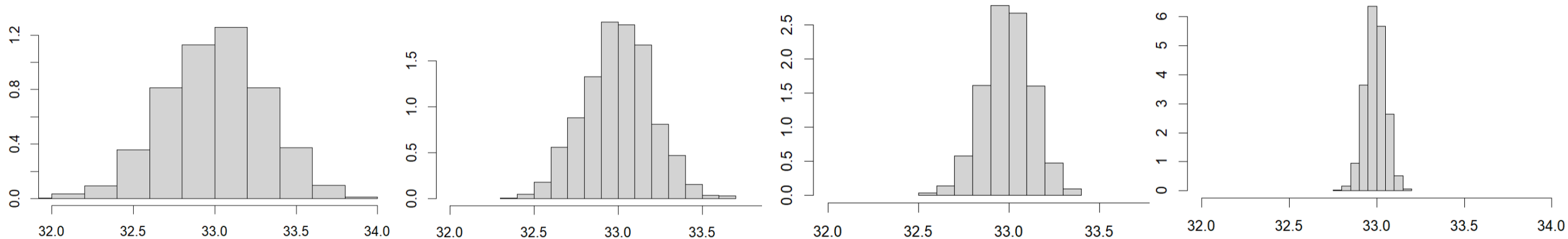


- ¿Podemos hacer esto en la vida?

# ¿Qué pasa si tomamos muestras más grandes?

→ `nreps = 1000`  
`n = 100`  
`medias.muestrales = NULL`

```
for (i in 1:nreps){  
  medias.muestrales[i] = mean(sample(poblacion, n, replace = F))  
}
```



# Población y muestra

## Una manera interactiva de verlo...

# Ejemplo 1

- Queremos estimar un parámetro (poblacional)
- NO tenemos los datos poblacionales, tenemos datos de una muestra aleatoria
- Sabemos que  $\bar{X}$  es una variable aleatoria
- Queremos no solamente estimar un valor (puntual) para  $\mu$ , sino además determinar con qué variabilidad = cuánta incertidumbre o cuánto error cometemos en nuestra estimación
- Estimación por intervalos

$$(\bar{X} - \text{ERROR}, \bar{X} + \text{ERROR})$$



# Ejemplo 1 – Cálculo de *ERROR*

- Opción 1: usando resultados teóricos

---

Si  $\bar{X}$  es la media de una muestra aleatoria de tamaño  $n$  de una población normal con varianza  $\sigma^2$  conocida, un intervalo de confianza de  $(1 - \alpha)100\%$  para  $\mu$  está dado por

$$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

donde  $z_{\alpha/2}$  es el valor de  $z$  que deja un área de  $\alpha/2$  a la derecha,

---

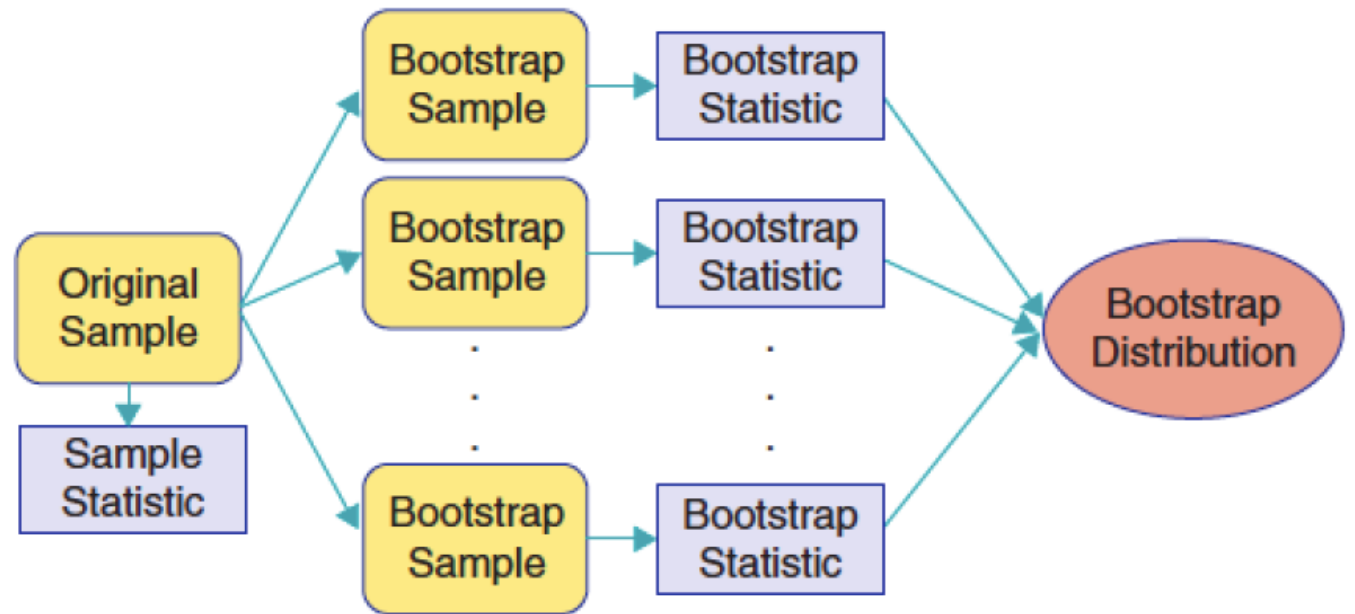


¿Qué ventajas y desventajas tendría usar la opción 1?

# Ejemplo 1 – Cálculo de *ERROR*

- Opción 2: usando los datos muestrales (método empírico)

bootstrap



# Ejemplo 1 – ¿Cómo lo hacemos en la PC?

Teoría Normal

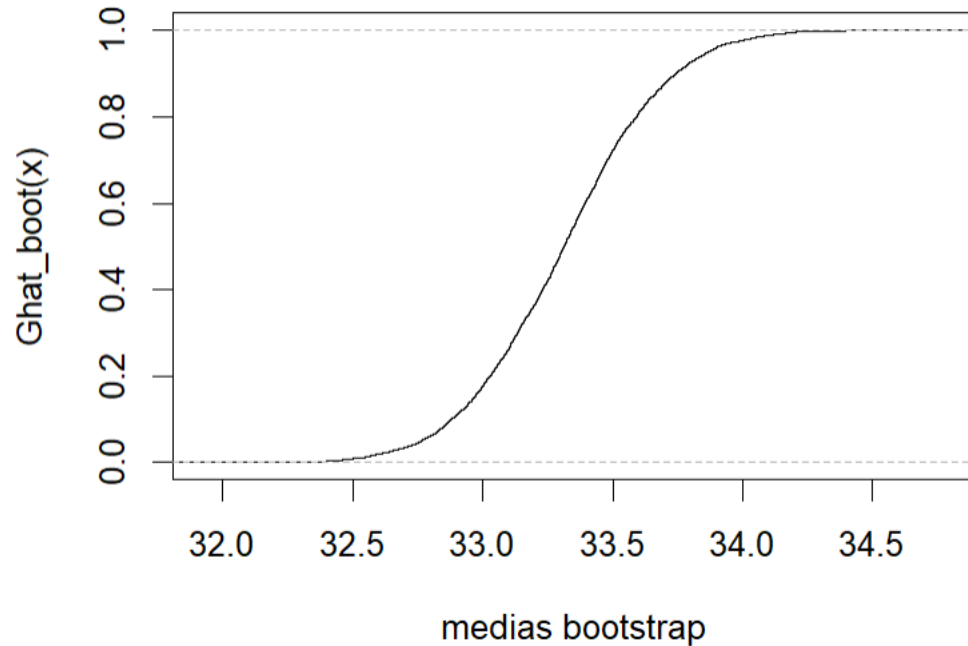
```
alfa = 0.05  
z = qnorm(1-alfa/2)  
mean(muestra) - z*sd(muestra)/sqrt(n)  
mean(muestra) + z*sd(muestra)/sqrt(n)
```

Bootstrap

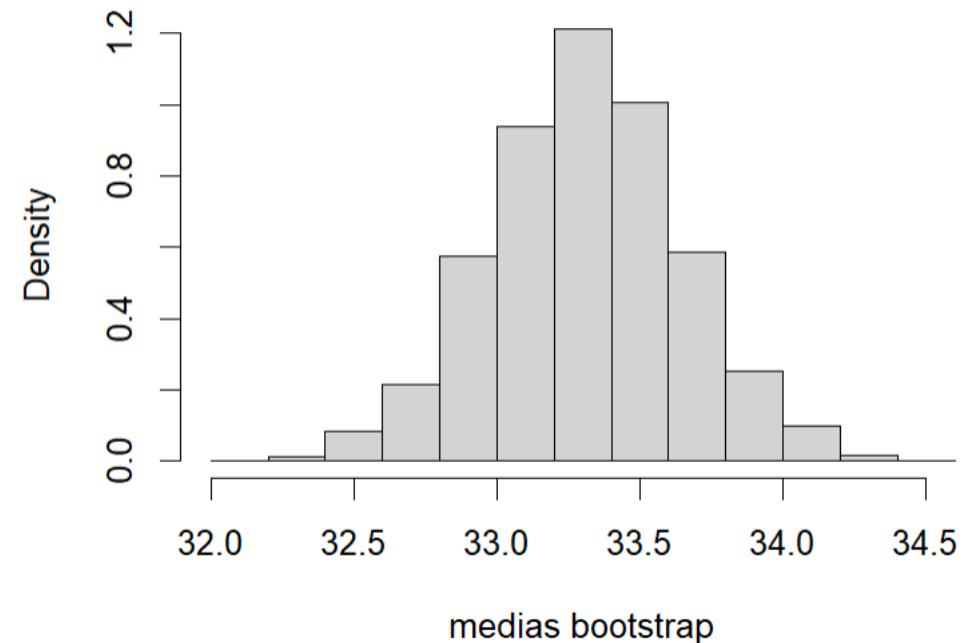
```
B = 5000  
medias.boot = NULL  
  
for (i in 1:B){  
  medias.boot[i] = mean(sample(muestra, n, replace = T))  
}  
  
hist(medias.boot, freq=F)  
abline(v = mean(muestra), col="red")  
mean(medias.boot)  
sd(medias.boot)  
  
c(mean(muestra)-2*sd(medias.boot), mean(muestra)+2*sd(medias.boot))  
quantile(medias.boot, c(0.025, 0.975))
```

# Ejemplo 1 – Resultados

Distribución acumulada boot (empírica)



Distribución bootstrap



¿A quién se parece?

- ¿Para qué podemos usar esta distribución?

```
> mean(medias.muestrales) [1] 33.00266
> mean(medias.boot)      [1] 33.3107
> mean(muestra)          [1] 33.31154
> sd(medias.muestrales)/sqrt(n) [1] 0.02909427
> sd(medias.boot)/sqrt(n)    [1] 0.03364247
```

# Ejemplo 1 – Resultados

- Veamos los intervalos

Normal

```
> round(c(Li, Ls),2)  
[1] 32.65 33.97
```

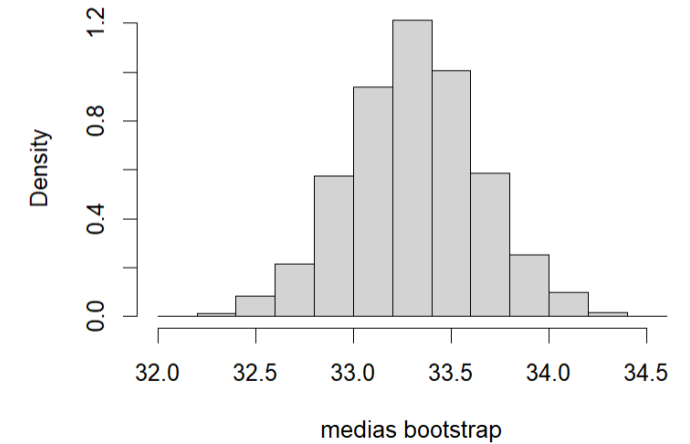
Asintótico

```
> round(c(mean(muestra)-2*sd(medias.boot), mean(muestra)+2*sd(medias.boot)),2)  
[1] 32.64 33.98
```

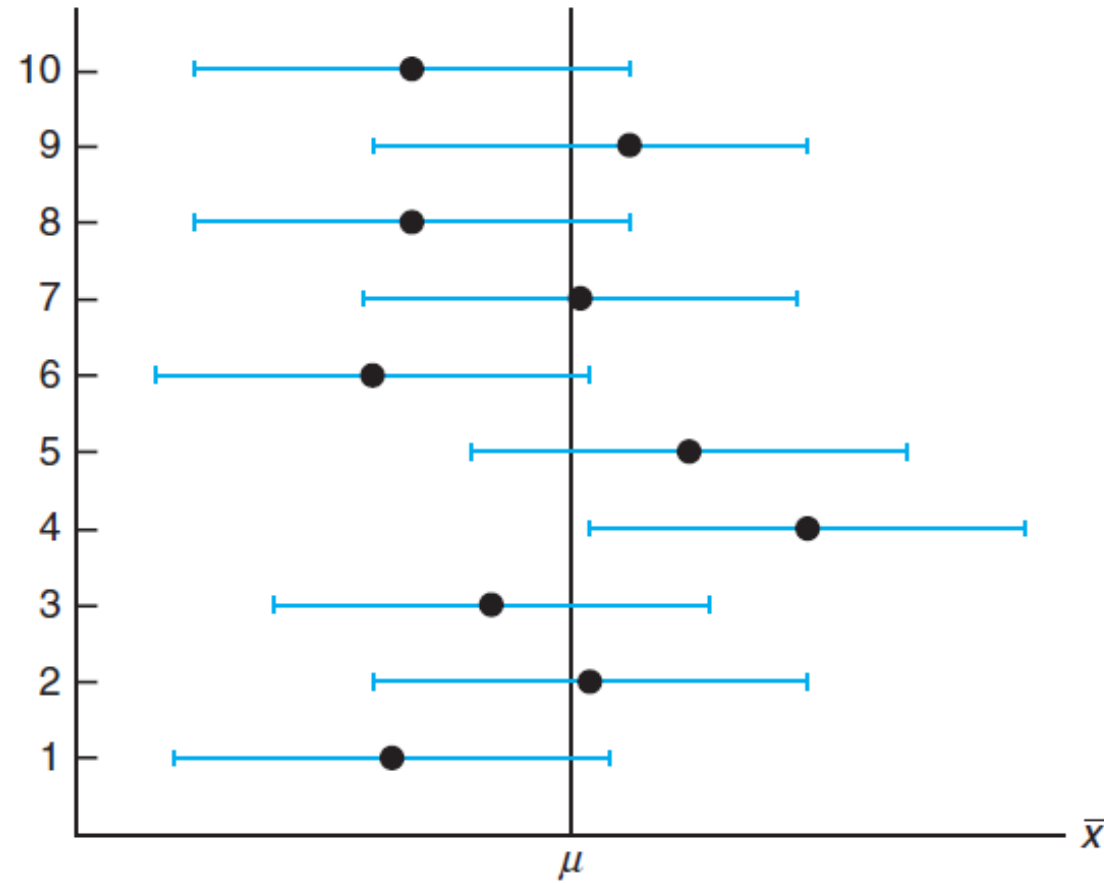
Percentílico

```
> round(quantile(medias.boot, c(0.025, 0.975)),2)  
2.5% 97.5%  
32.64 33.99
```

Distribución bootstrap



# Ejemplo 1 - ¿Qué representa el IC del $(1-\alpha)\%$ ?



# Ejemplo 1 - ¿Qué representa el IC del $(1-\alpha)\%$ ?

```
# Cálculo de coverage de IC del 95%
mu = 33          # media verdadera
sigma = 3        # desvío verdadero
n = 30           # tamaño muestral
M = 1000         # número de repeticiones del experimento
alfa = 0.05      # nivel de significancia
```

```
# Contador
```

```
coverN = numeric(M)
```

```
coverB = numeric(M)
```

```
# Loop 1 (generar una nueva muestra aleatoria)
```

```
for(i in 1:M){
```

```
  # Generar muestra
```

```
  x = rnorm(n, mean = mu, sd = sigma)
```

```
  # Estadísticos
```

```
  xbar = mean(x)
```

```
  s = sd(x)
```

```
  xboot = numeric(B)
```

```
  # Loop 2 (realizar bootstrap sobre la muestra actual)
```

```
  for(j in 1:B){
```

```
    xboot[j] = mean(sample(x, n, replace = T))
```

```
  }
```

```
# Intervalo de confianza (normal)
```

```
tcrit = qt(1 - alfa/2, df = n - 1)
```

```
LI_n = xbar - tcrit * s / sqrt(n)
```

```
LS_n = xbar + tcrit * s / sqrt(n)
```

```
# Intervalo de confianza (bootstrap)
```

```
intervalo = quantile(xboot, c(0.025, 0.975))
```

```
LI_b = intervalo[1]
```

```
LS_b = intervalo[2]
```

```
# Verificar si contiene la media verdadera
```

```
coverN[i] = (LI_n <= mu & mu <= LS_n)
```

```
coverB[i] = (LI_b <= mu & mu <= LS_b)
```

```
}
```

```
# Coverage empírico
```

```
mean(coverN)
```

```
mean(coverB)
```

# Verificamos *coverage* en la simulación anterior

```
> mean(coverN)
[1] 0.948
> mean(coverB)
[1] 0.931
```



¿Qué relación podemos establecer entre la varianza y el IC?

- ¿Qué significa que el coverage me dé por debajo del valor nominal?

$$\bar{X} \pm t_{\alpha} \frac{SE}{\sqrt{2}}$$

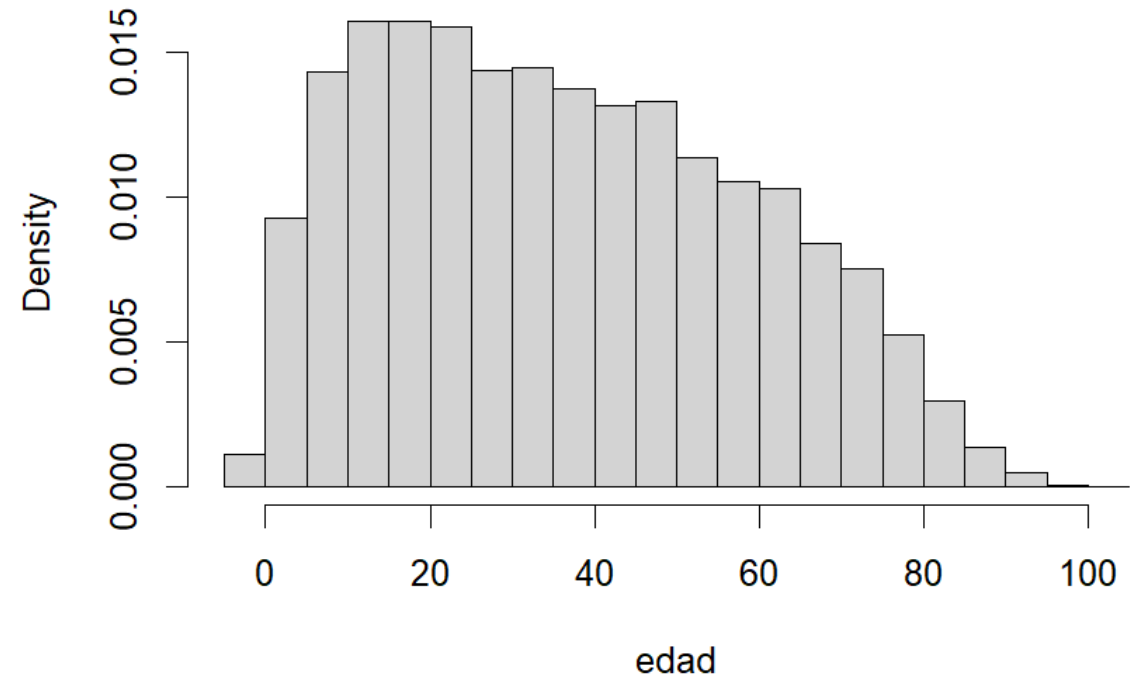


# Ejemplo 2 – datos reales

- Queremos estimar un parámetro (edad media de la población urbana argentina)
- NO tenemos los datos poblacionales, tenemos una muestra (aleatoria) → Encuesta permanente de hogares
- ¿Son normales?
- ¿Representan parámetros?

```
> mean(edad)
[1] 37.11988
> sd(edad)
[1] 22.22135
```

**Distribución de la variable edad**



## *Ejercicio 2*

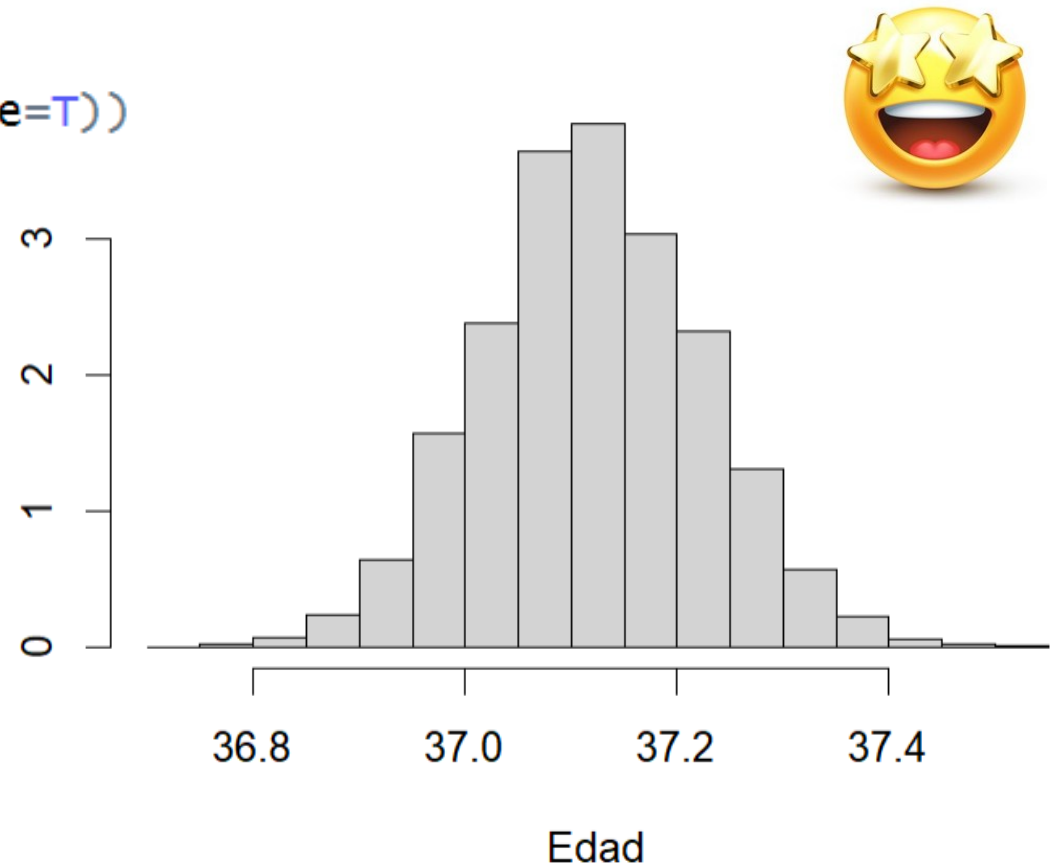
1. Obtener una aproximación de la distribución muestral de la edad media a partir de los datos muestrales (distribución bootstrap)
2. Calcular un IC para la media poblacional bajo normalidad
3. Calcular un IC para la media poblacional mediante bootstrap

# Hagamos bootstrap a ver qué pasa...

```
n = length(edad)
B = 5000
medias = NULL

for (i in 1:B){
  medias[i] = mean(sample(edad, n, replace=T))
}
```

```
> ICnorm
[1] 36.9 37.3
> ICB1
[1] 36.9 37.3
> ICB2
 2.5% 97.5%
36.9  37.3
```



Un resultado MUY importante de la estadística

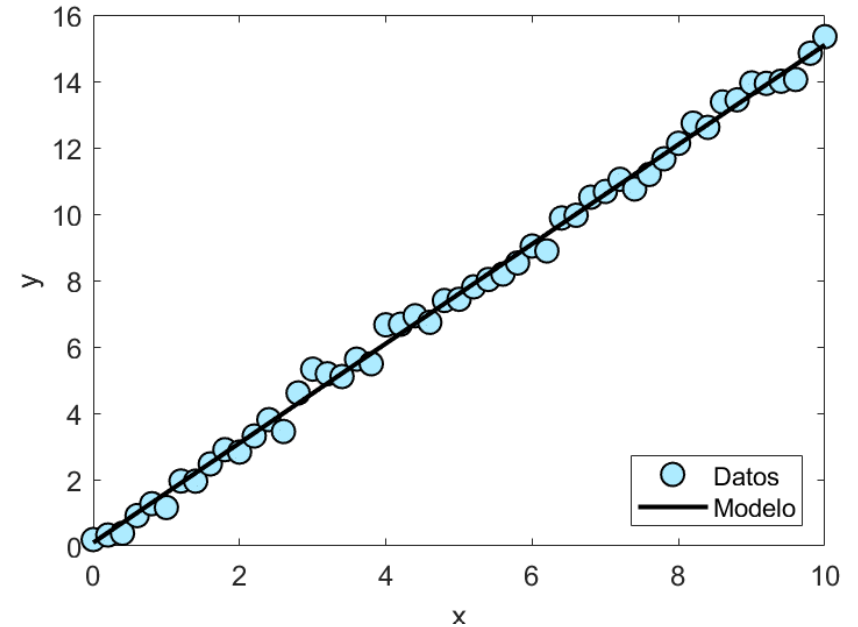
# Teorema Central del Límite (TCL)



# Ejemplo 3 – regresión

## A) regresión lineal simple (RLS)

- Tengo datos  $(x, y)$
- Modelo  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Uso el modelo estimado para predecir  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$
- Objetivo: estimar variabilidad de los estimadores, por ej.  $\text{var}(\hat{\beta}_1)$



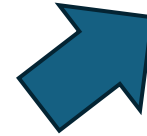
**Para qué?**

- Decir algo sobre  $\beta$  (poblacional), interpretar
- Dar un IC para la predicción

# Ejemplo 3A – RLS

- Opción 1: resultados teóricos

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{(n-2) \sum_{i=1}^n (x_i - \bar{x})^2}$$



$$\widehat{\text{Var}}(\hat{y}_0) = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\hat{y}_0 \pm t_{1-\alpha/2, n-2} \sqrt{s^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}$$

- Opción 2 ???



# Ejemplo 3A – RLS

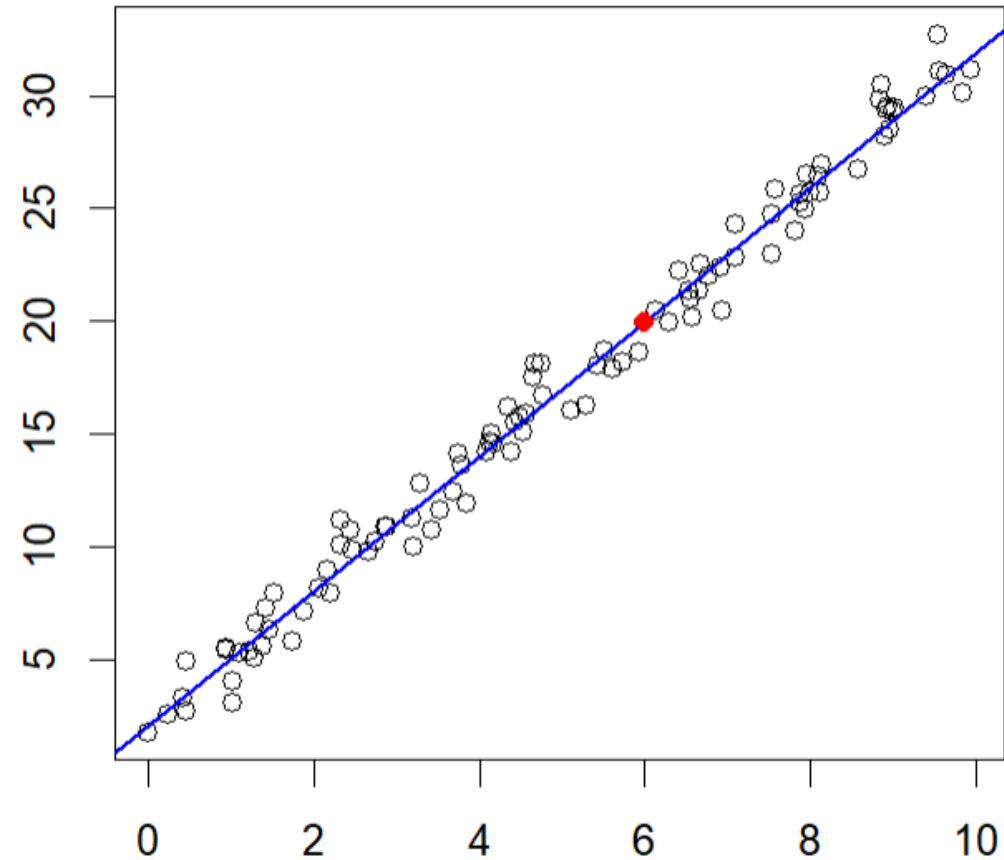
```
set.seed(123)
n = 100
beta0 = 2
beta1 = 3
sigma = 1

# Datos de entrenamiento
x = runif(n, 0, 10)
y = beta0 + beta1*x + rnorm(n, 0, sigma)

# Dato de predicción
x0 = 6

mod = lm(y ~ x)

plot(x, y)
```



# Ejemplo 3A – RLS

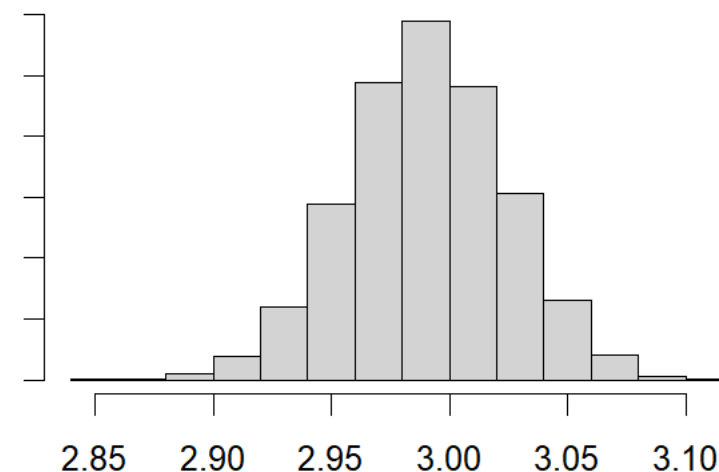
```
# Bootstrap para var(beta) y var(y0)
B = 5000
beta1_boot = numeric(B)
yhat_boot = numeric(B)

for(b in 1:B){
  ind = sample(1:n, size = n, replace = TRUE)
  x_b = x[ind]
  y_b = y[ind]
  mod_b = lm(y_b ~ x_b)
  beta1_boot[b] = coef(mod_b)[2]
  yhat_boot[b] = predict(mod_b, newdata = data.frame(x_b = x0))
}
```

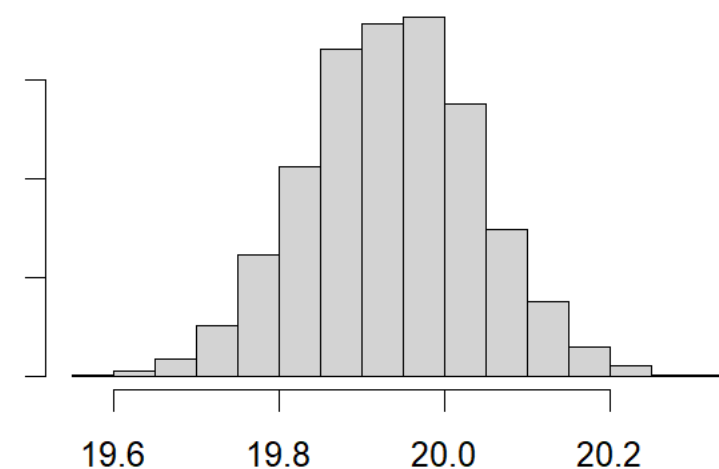
```
> varbeta1.norm
[1] 0.0012
> varbeta1.boot
[1] 0.0011
```

```
> round(IC_clasico,3)
      fit      lwr      upr
1 19.937 19.733 20.141
> round(IC_boot,3)
      2.5%   97.5%
19.732 20.142
```

beta1 boot



y0 boot





# Ejemplo 3 – regresión

## B) regresión lineal múltiple (RLM)

### Datos de vinos

Clarid	Aroma	Cuerpo	Sabor	Fuerza	Calidad	Region
1.7	7.7	9.6	8.7	2.9	16.1	Cuyo
1.5	5.9	8.7	7.0	3.0	15.6	Cuyo
1.8	5.5	9.6	5.6	3.5	15.5	Cuyo
1.4	7.1	8.9	5.8	3.1	15.5	Cuyo
1.9	6.4	9.4	6.6	3.8	15.1	Cuyo
1.0	6.8	9.0	6.0	3.2	14.9	Cuyo
1.0	5.1	9.3	4.5	3.6	14.4	Cuyo
1.0	4.3	8.9	4.7	3.9	13.9	Noroeste

## Ejemplo 3B) regresión lineal múltiple (RLM)

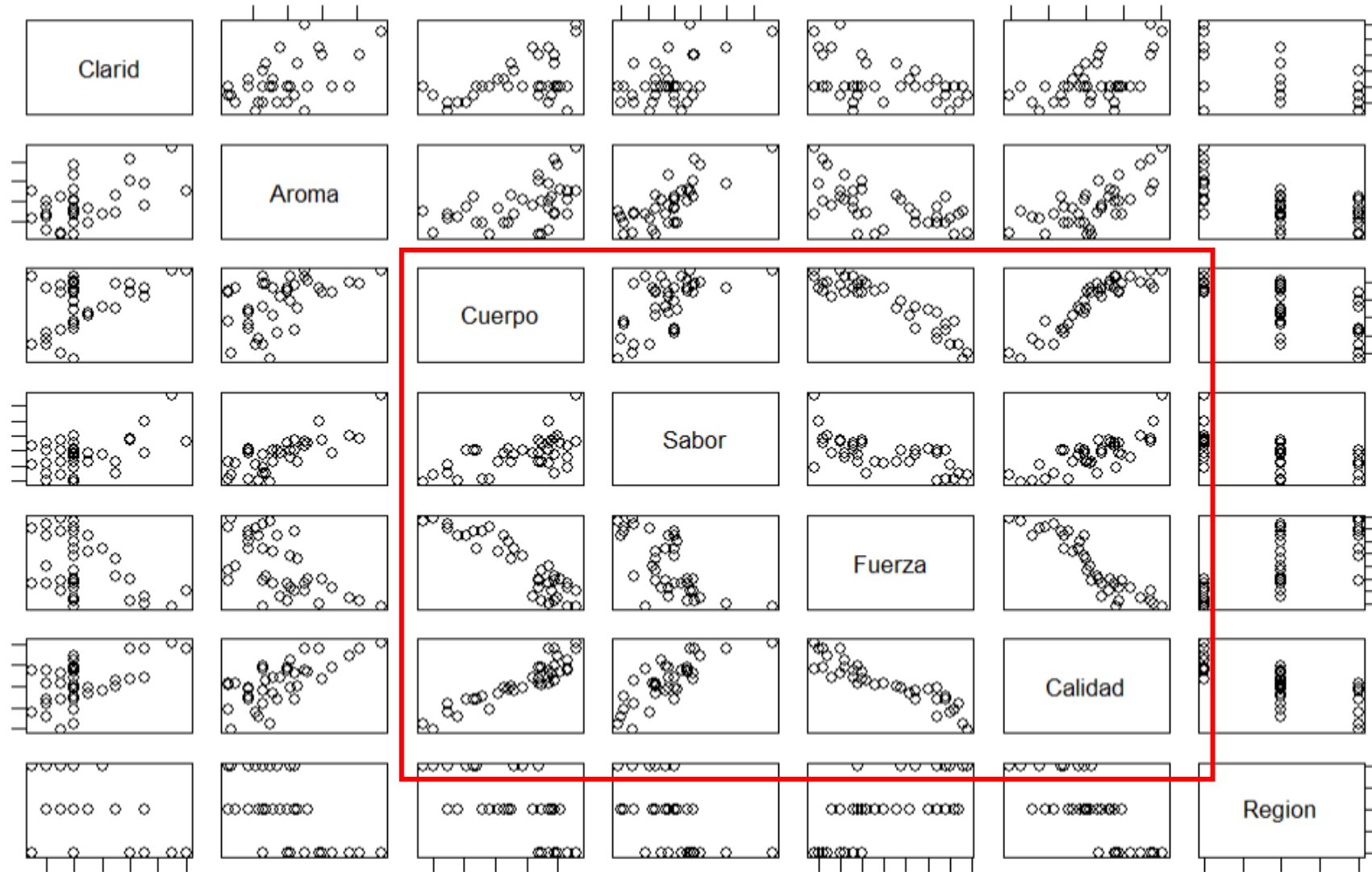
### Datos de vinos

Clarid	Aroma	Cuerpo	Sabor	Fuerza	Calidad	Region
1.7	7.7	9.6	8.7	2.9	16.1	Cuyo
1.5	5.9	8.7	7.0	3.0	15.6	Cuyo
1.8	5.5	9.6	5.6	3.5	15.5	Cuyo
1.4	7.1	8.9	5.8	3.1	15.5	Cuyo
1.9	6.4	9.4	6.6	3.8	15.1	Cuyo
1.0	6.8	9.0	6.0	3.2	14.9	Cuyo
1.0	5.1	9.3	4.5	3.6	14.4	Cuyo
1.0	4.3	8.9	4.7	3.9	13.9	Noroeste

- Tengo datos
- Ahora el modelo es  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} + \varepsilon_i$
- Objetivo (inferencial): el mismo que antes

## Ejemplo 3B) regresión lineal múltiple (RLM)

Datos de vinos - ¿Cómo lucen? ¿Parece tener sentido la regresión?



# Ejemplo 3B – RLM

```
modelo = lm(Calidad ~ Cuerpo + Sabor + Fuerza, data = datos)
summary(modelo)
plot(modelo)
pred_puntual = predict(modelo, dato_prueba, se.fit = TRUE, interval = "confidence")
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	7.3488	2.3805	3.087	0.004077	**
Cuerpo	0.6580	0.1715	3.836	0.000535	***
Sabor	0.4943	0.1130	4.373	0.000116	***
Fuerza	-0.5566	0.2207	-2.522	0.016661	*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5935 on 33 degrees of freedom

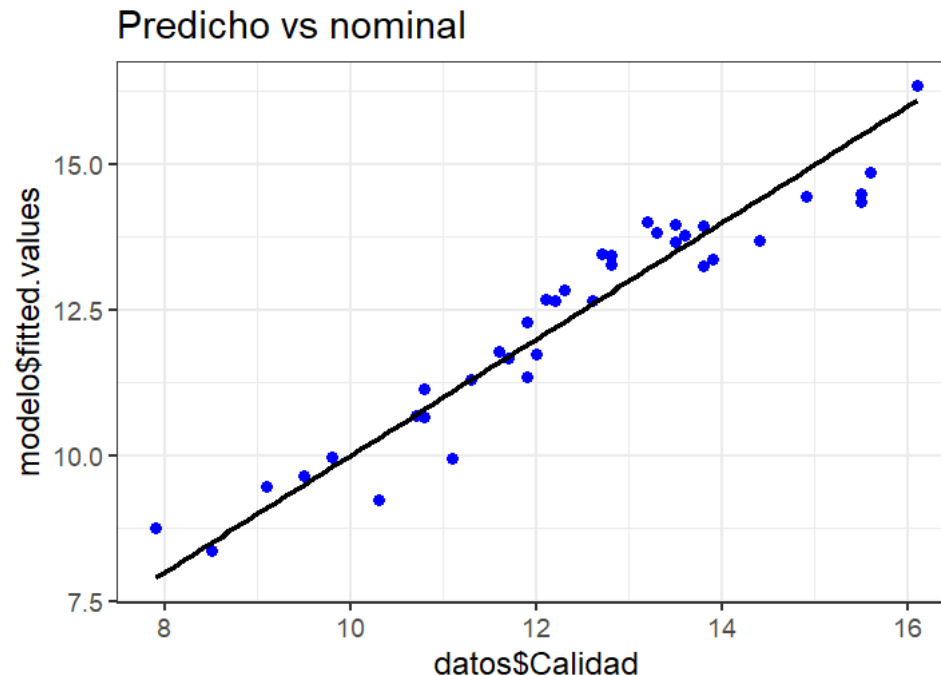
Multiple R-squared: 0.9193, Adjusted R-squared: 0.912

F-statistic: 125.3 on 3 and 33 DF, p-value: < 2.2e-16

# Ejemplo 3B – RLM

```
x1 = seq(from = min(datos$Calidad),to = max(datos$Calidad),by = 0.1)
rectaunidad = x1
```

```
ggplot() +
  geom_point(aes(x = datos$Calidad, y = modelo$fitted.values),col = 'blue') +
  geom_line(aes(x = x1, y = rectaunidad),col='black',size=1) +
  labs(title = "Predicho vs nominal") +
  theme_bw()
```



```

# Bootstrap
m = 5000
n = nrow(datos)
betas = matrix(0, ncol = 4, nrow = m)
pred_puntual_boot = matrix(NA, nrow = m, ncol = 1)

for (i in 1:m){
  train_idx = sample(1:n, size = n, replace = TRUE)
  datosB = datos[train_idx,]
  fit_boot = lm(Calidad ~ Cuerpo + Sabor + Fuerza, data = datosB)
  betas[i,] = fit_boot$coefficients[1:4]
  pred_puntual_boot[i] = predict(fit_boot, dato_prueba)
}

# Distrib bootstrap de beta y de y0
hist(betas[,1], freq = F)
hist(betas[,2], freq = F)
hist(betas[,3], freq = F)
hist(betas[,4], freq = F)

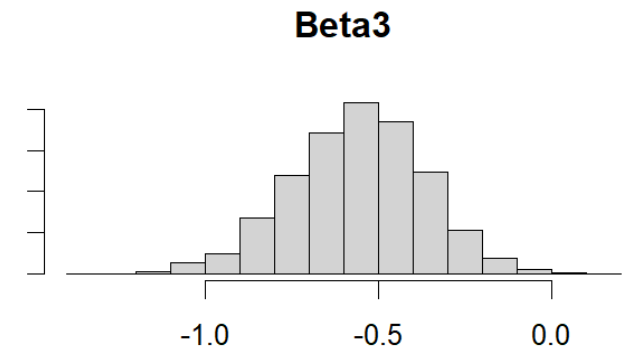
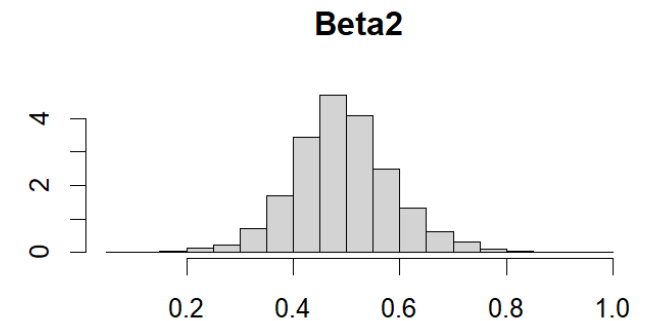
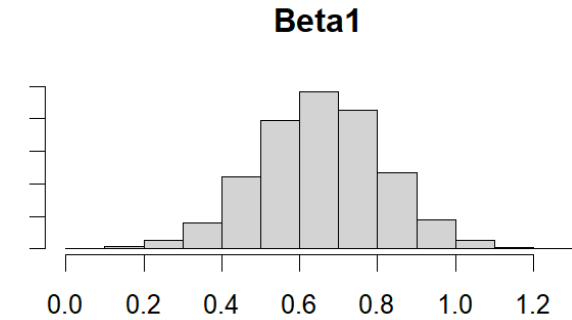
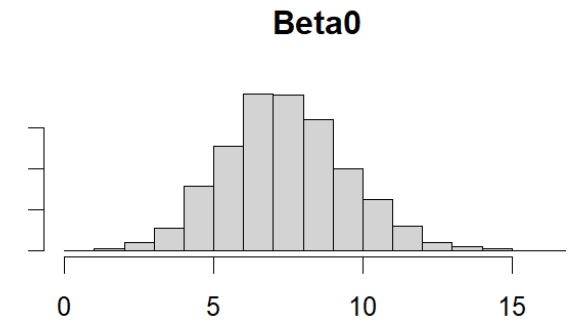
hist(pred_puntual_boot, freq = F)

# Error estándar de los beta estimados
resumen = summary(modelo)
sebetas_teor = as.numeric(resumen[["coefficients"]][,2])

sdbeta1 = sd(betas[,1])
sdbeta2 = sd(betas[,2])
sdbeta3 = sd(betas[,3])
sdbeta4 = sd(betas[,4])

# Error estándar de la predicción
sdteor = pred_puntual$se.fit
sdpred = sd(pred_puntual_boot)

```



# Ejemplo 3B – RLM

- Comparación

	Cantidad	Teoría	Bootstrap
Predicción		14.681	14.676
SE predicción		0.235	0.221

	Cantidad	Teoría	Bootstrap
beta1	7.349	7.401	
SE beta1	2.381	2.092	
beta2	0.658	0.654	
SE beta2	0.172	0.161	
beta3	0.494	0.493	
SE beta3	0.113	0.095	
beta4	-0.557	-0.561	
SE beta4	0.221	0.194	

# Para reflexionar...

- En modelos clásicos, las varianzas de los estimadores suelen obtenerse en forma analítica.
- En modelos más complejos, deducir expresiones cerradas puede ser muy difícil, tedioso o directamente impracticable.
- Bootstrap permite aproximar numéricamente distribuciones muestrales, errores estándar e intervalos de confianza.
- Particularmente útil en machine learning y modelos no lineales o “no clásicos”.

La filosofía profunda del bootstrap clásico es justamente:

“La muestra observada contiene suficiente información sobre la población como para usarla como una pseudo-población.”

Y por eso Efron originalmente lo formuló como un reemplazo computacional de derivaciones asintóticas difíciles.



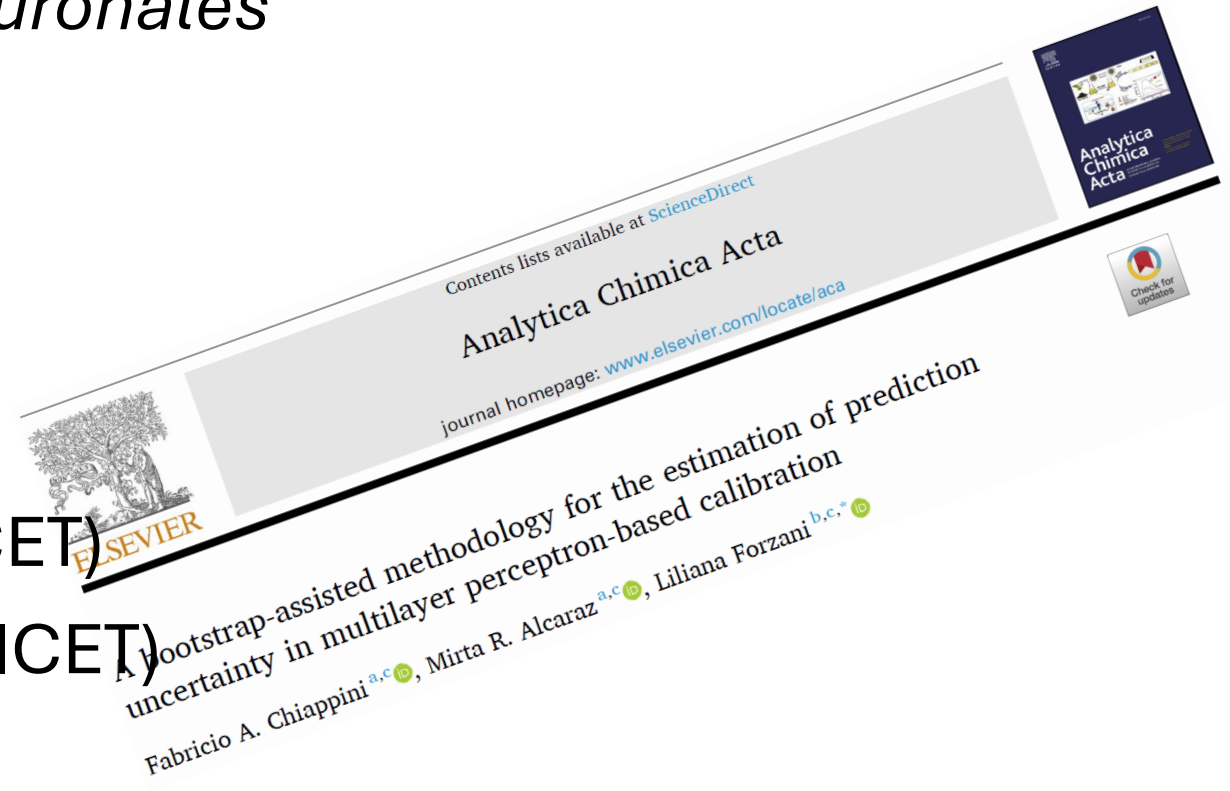
# Un caso que estamos investigando

*Incertidumbre de predicción en problemas de regresión no lineal con datos químicos y redes neuronales*

- Liliana Forzani
- Fabricio Chiappini

Colaboración con

- Mariela Sued (UdeSA, CONICET)
- Alejandro Olivieri (UNR, CONICET)



# El contexto: ¿qué hacemos en quimiometría?

- © Qué muestras nos interesan
- © Cómo son nuestros datos
- © Qué es calibración

# El contexto: ¿qué hacemos en quimiometría?

## © Qué muestras nos interesan

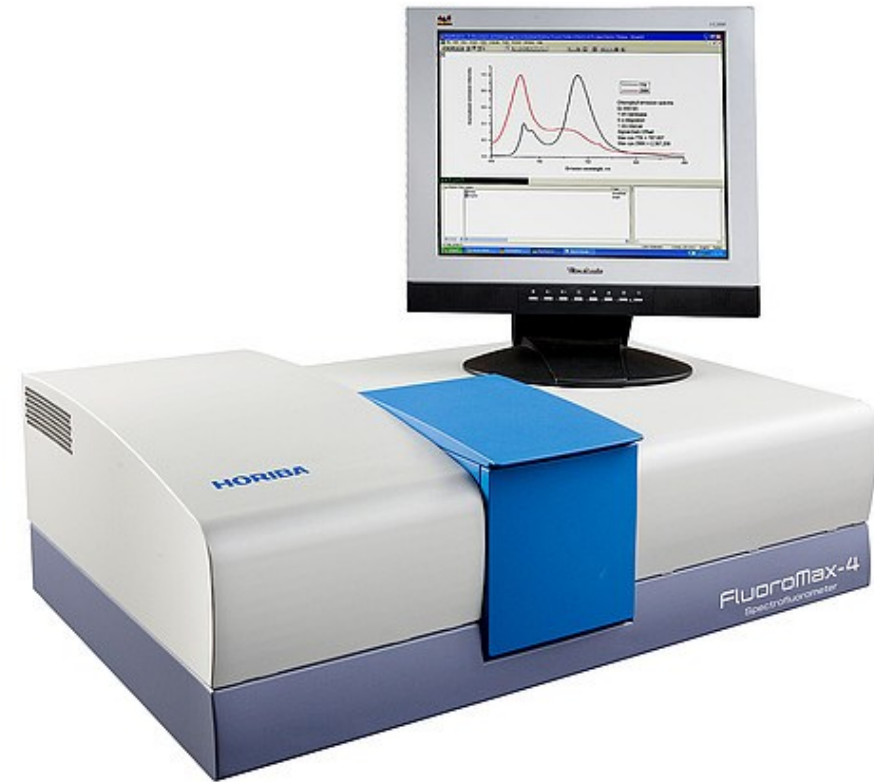
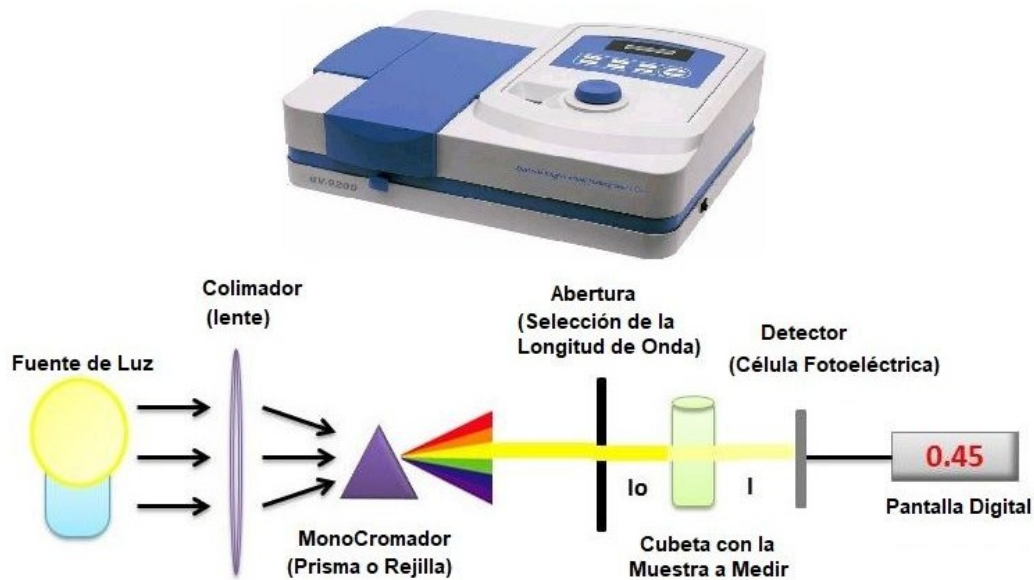


La  $y$  (variable respuesta) es la cantidad de una sustancia disuelta

La expresamos en general como medida relativa (concentración), es decir, cantidad de sustancia por unidad de volumen

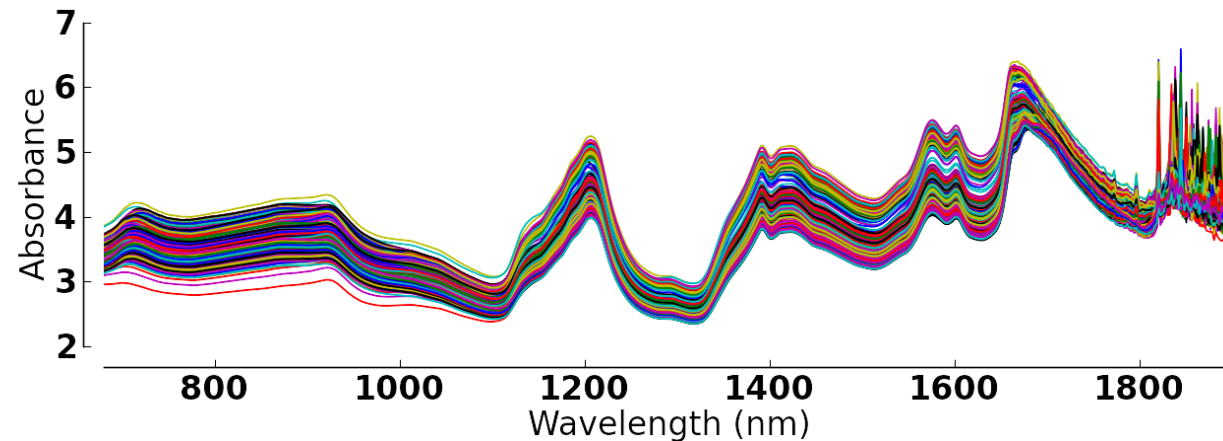
# El contexto: ¿qué hacemos en quimiometría?

© Cómo son nuestros datos

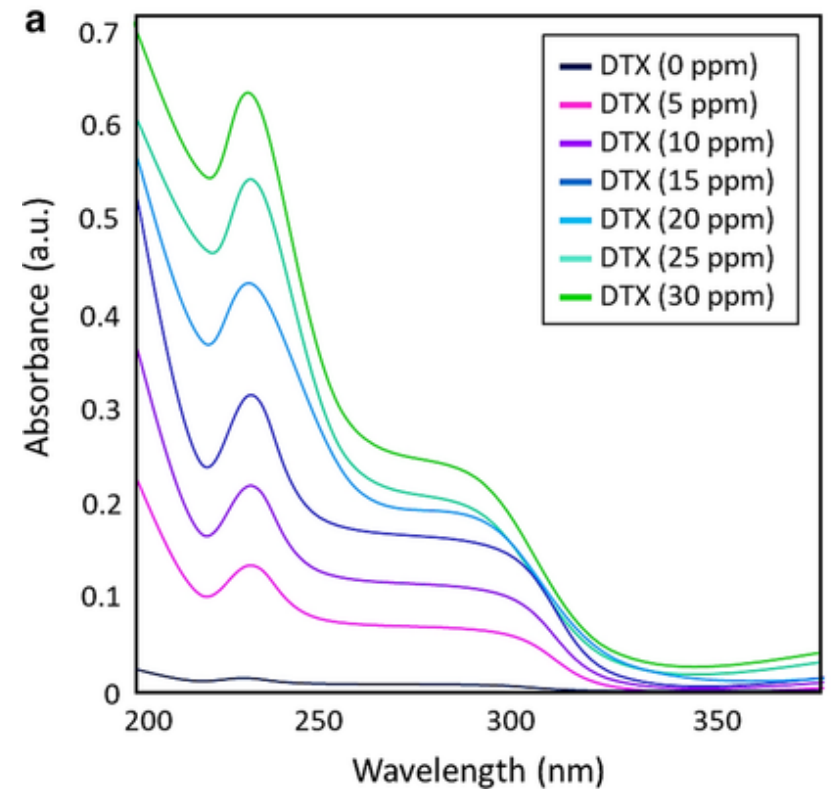


# El contexto: ¿qué hacemos en quimiometría?

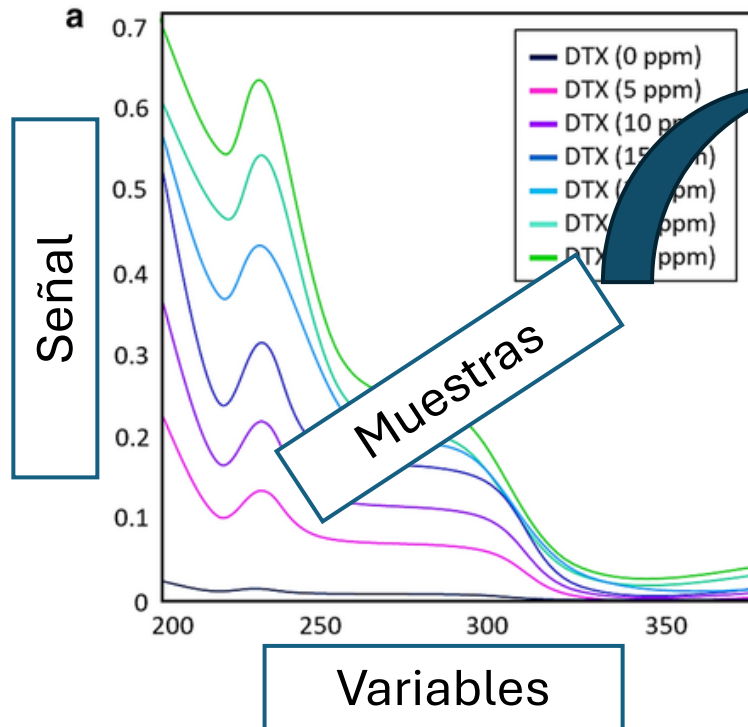
© Cómo son nuestros datos



Estas son nuestras variables  
predictoras (**X**)  
Datos espectrales (espectros)



# El contexto: ¿qué hacemos en quimiometría?



Nuestra matriz de datos

$n$ : 30-100

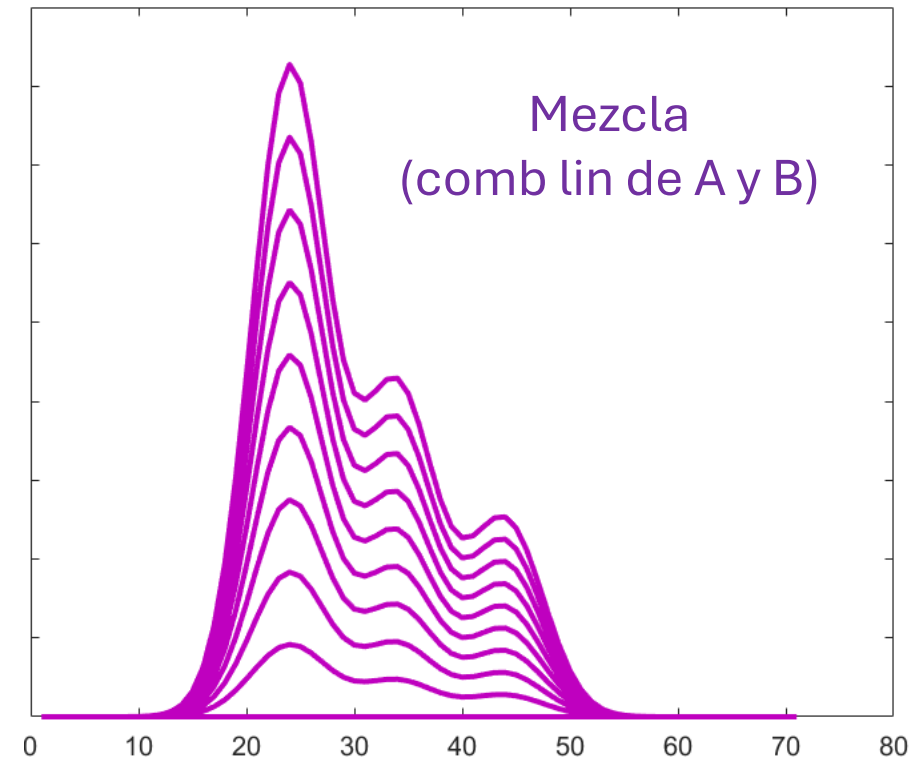
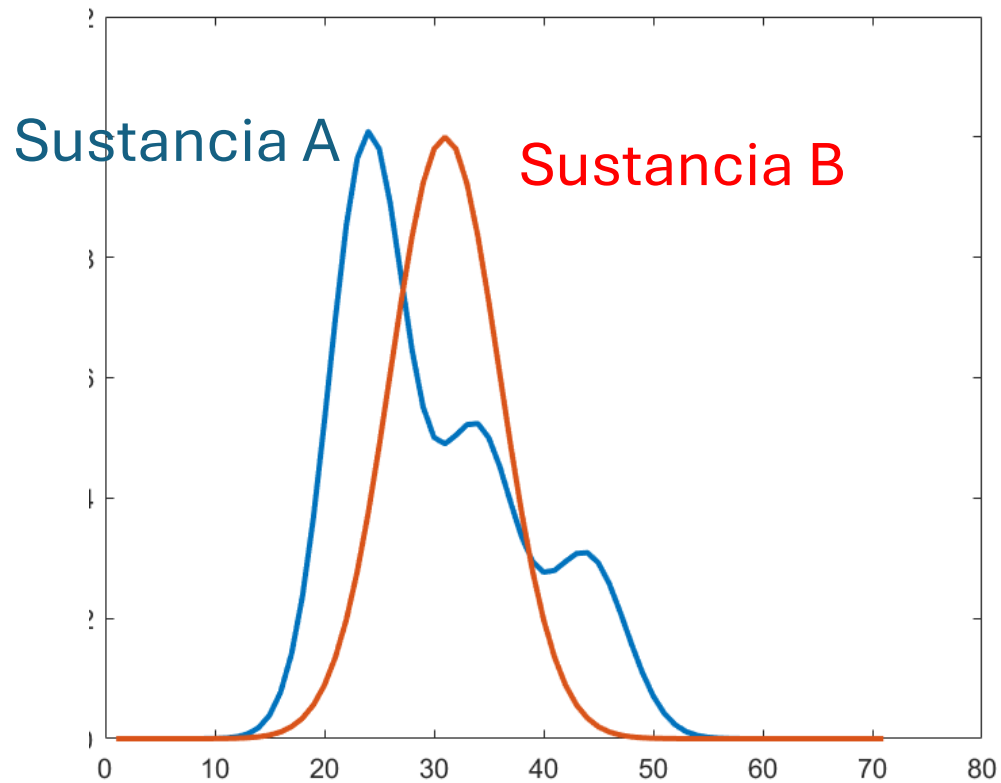
$p$ : 100-10000

$d$ : menor a 10

Muestra	V1	V2	V3	V4	V5	...
1						...
2						...
3						...
...						...

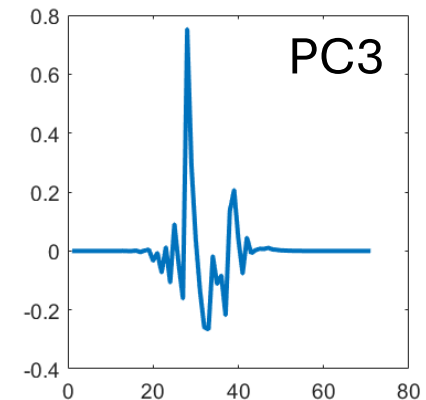
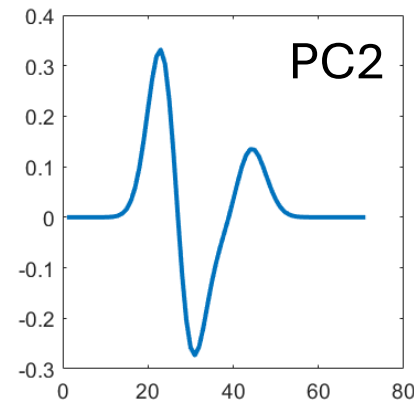
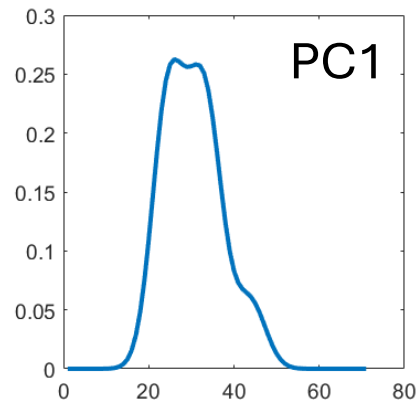
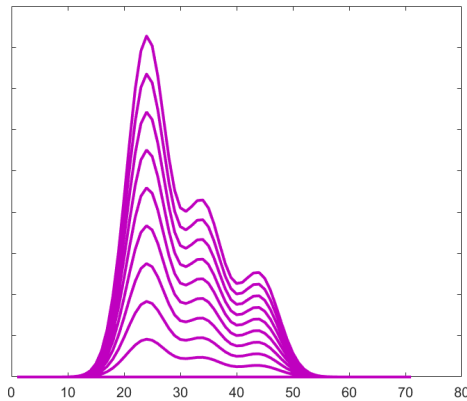
# El contexto: ¿qué hacemos en quimiometría?

Los datos son muy redundantes (multicolinealidad)



# El contexto: ¿qué hacemos en quimiometría?

Aplicamos técnicas de reducción de la dimensión (ej PCA)

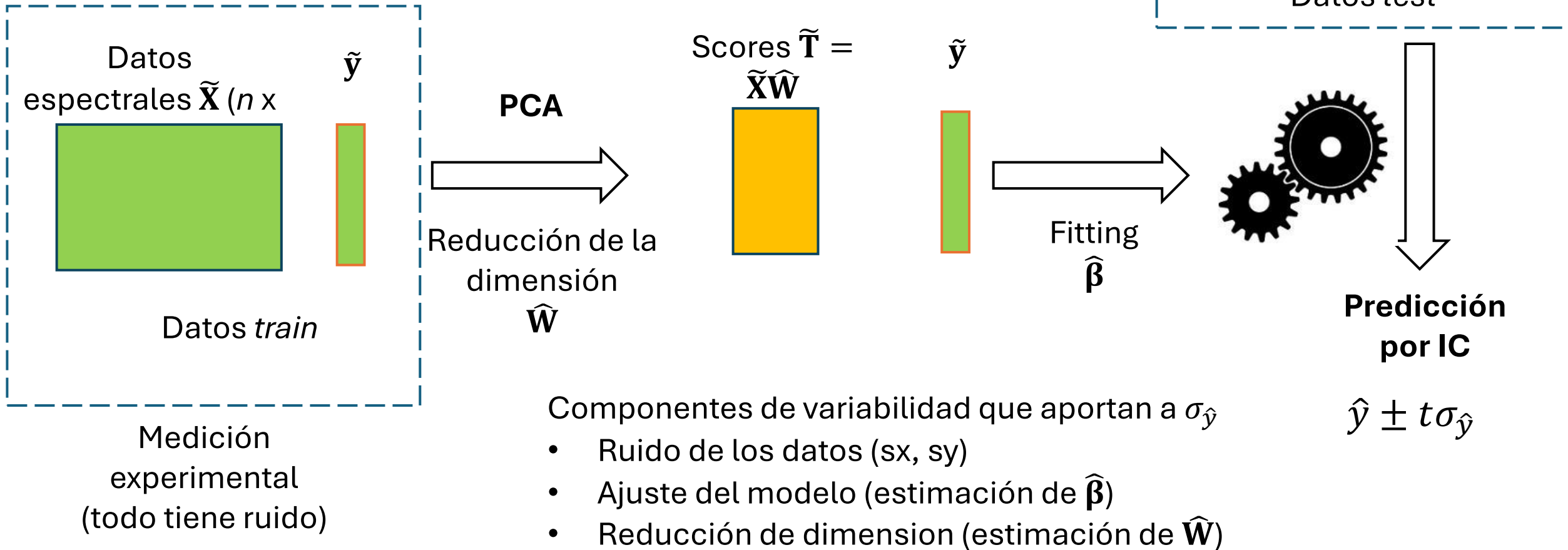


Con los datos, entrenamos modelos de regresión

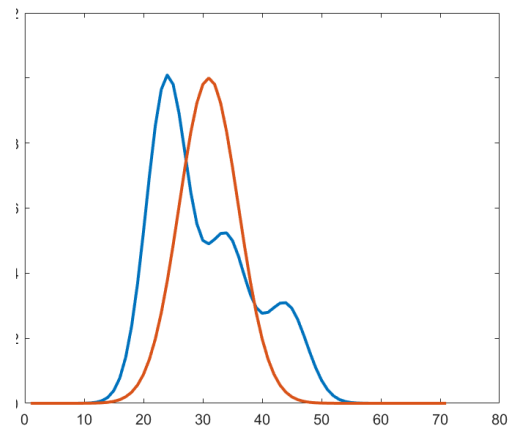
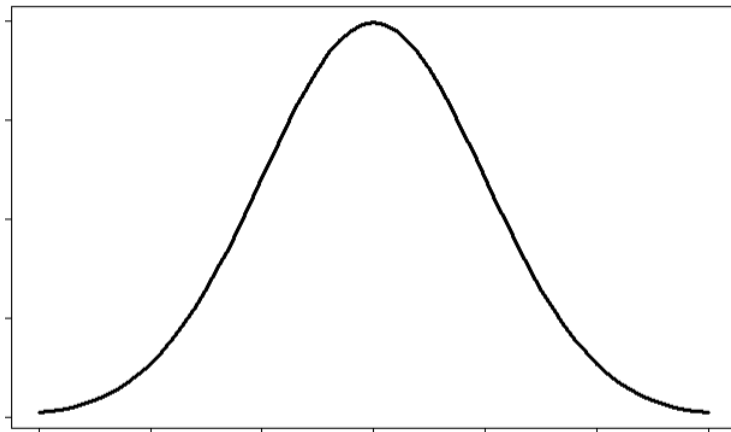
- El modelo base es el mismo:  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} + \varepsilon_i$
- La diferencia es que ahora las  $x$  no son las variables espectrales originales, sino combinaciones lineales de ellas tales que (si la reducción fue buena), puedo resumir la información importante en términos de variabilidad, en un conjunto de variables predictoras mucho menor (variables latentes)



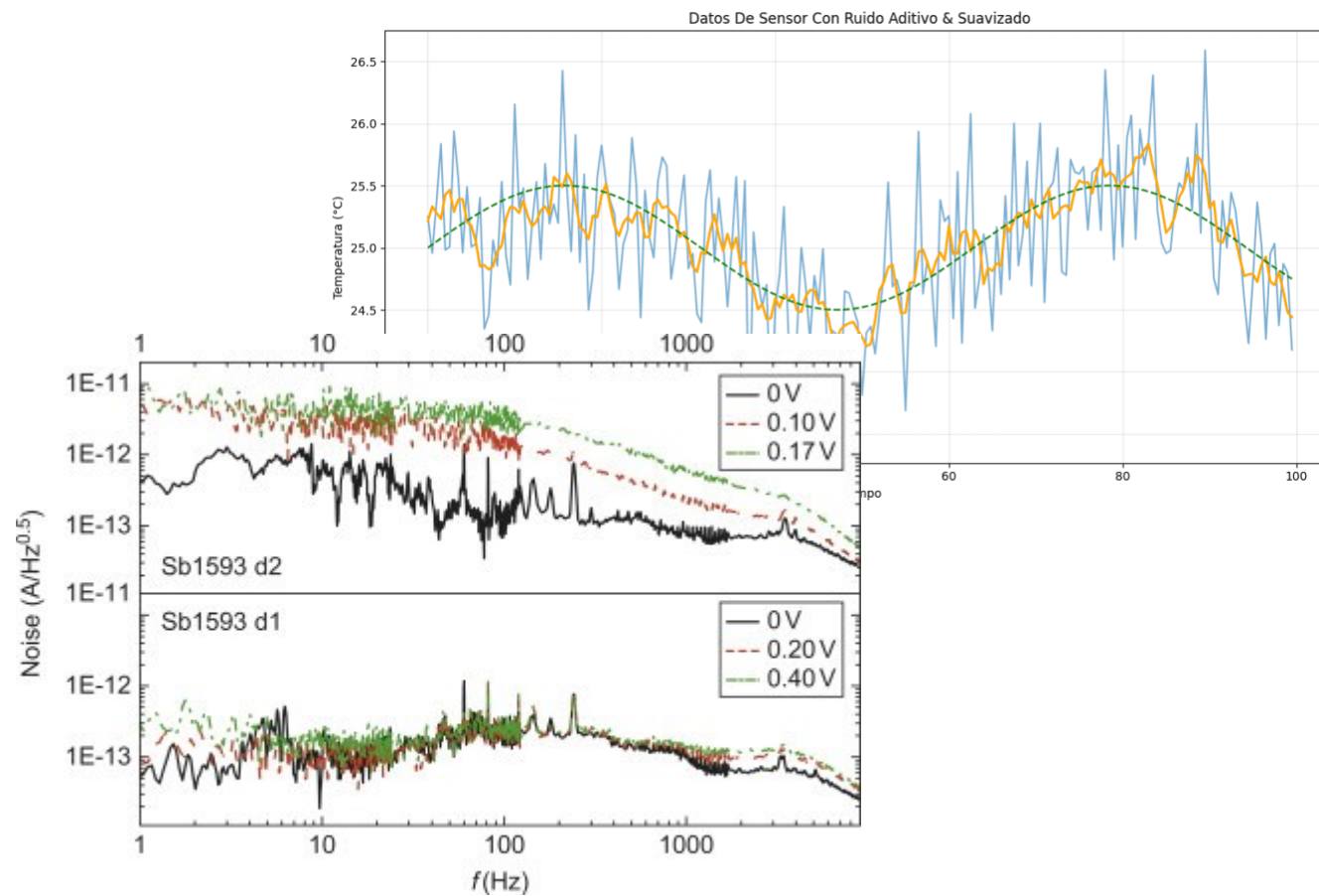
# El problema



Mundo ideal



Mundo **real**



# Modelo lineal con error en las variables (notación)

La verdad  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$

Modelo en términos de la verdad  
y de lo que mido  $\tilde{\mathbf{y}} = (\tilde{\mathbf{X}} + \Delta\mathbf{X})\boldsymbol{\beta} + \Delta\mathbf{y}$   
 $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \Delta\mathbf{X}\boldsymbol{\beta} + \Delta\mathbf{y}$

Un nuevo  $\mathbf{x}$   
(el espectro de una muestra test)  $\tilde{\mathbf{x}}_u = \mathbf{x}_u + \Delta\mathbf{X}$

Ecuación de predicción  $\hat{y}_u = \bar{y} + \tilde{\mathbf{x}}_u^T \hat{\boldsymbol{\beta}} = \bar{y} + \mathbf{x}_u^T \hat{\boldsymbol{\beta}} + \Delta\mathbf{X}\hat{\boldsymbol{\beta}}$

## Varianza de la predicción (Delta Método)

$$\sigma_{\hat{y}_u}^2 \approx \frac{\sigma_{\Delta y}^2 + \|\hat{\boldsymbol{\beta}}\|^2 \sigma_{\Delta \mathbf{x}}^2}{n} + \tilde{\mathbf{x}}_u^T \mathbf{V}(\hat{\boldsymbol{\beta}}) \tilde{\mathbf{x}}_u + \|\hat{\boldsymbol{\beta}}\|^2 \sigma_{\Delta \mathbf{x}}^2$$

# Modelo lineal con error en las variables (notación)

## Varianza de la predicción

$$\sigma_{\hat{y}_u}^2 \approx \frac{\sigma_{\Delta y}^2 + \|\hat{\boldsymbol{\beta}}\|^2 \sigma_{\Delta x}^2}{n} + \tilde{\mathbf{x}}_u^T \mathbf{V}(\hat{\boldsymbol{\beta}}) \tilde{\mathbf{x}}_u + \|\hat{\boldsymbol{\beta}}\|^2 \sigma_{\Delta x}^2$$

Las fuentes de  
variabilidad

## Varianza de la predicción (con reducción, ej PCR)

$$\sigma_{\hat{y}_u}^2 \approx \frac{\sigma_{\Delta y}^2 + \|\hat{\boldsymbol{\beta}}_{\text{PCR}}\|^2 \sigma_{\Delta x}^2}{n} + (\tilde{\mathbf{x}}_u \hat{\mathbf{W}})^T \mathbf{V}(\hat{\boldsymbol{\beta}}_{\text{PCR}}) \tilde{\mathbf{x}}_u \hat{\mathbf{W}} + \|\hat{\boldsymbol{\beta}}_{\text{PCR}}\|^2 \sigma_{\Delta x}^2$$

- No consideramos el error en la estimación de  $\mathbf{W}$
- No tiene en cuenta la corrección del sesgo en la estimación de  $\hat{\boldsymbol{\beta}}_{\text{PCR}}$  (al observar  $\mathbf{x}$  con ruido, hay un problema de consistencia)
- La relación entre las variables podría ser no lineal

# Varianza asintótica H (caso general)

$$H = C_\psi M_{\psi\psi} C_\psi^\top + C_\Theta M_{\Theta\Theta} C_\Theta^\top + C_\Sigma M_{\Sigma\Sigma} C_\Sigma^\top + 2 C_\Theta M_{\Theta\Sigma} C_\Sigma^\top.$$

$$C_\psi = A_{f,\beta} B_{\psi,\beta}^{-1},$$

$$C_\Theta = A_{f,\beta} B_{\psi,\beta}^{-1} B_{\psi,\Theta} - A_{f,\Theta},$$

$$C_\Sigma = A_{f,\beta} B_{\psi,\beta}^{-1} B_{\psi,\Sigma_{\mathbf{u}|\mathbf{x}}},$$

$$M_{\psi\psi} = \mathbb{E} \left[ \psi(\beta, \Theta \tilde{x}_i, \Sigma_{\mathbf{u}|\tilde{x}}) \psi(\beta, \Theta \tilde{x}_i, \Sigma_{\mathbf{u}|\tilde{x}})^\top \right],$$

$$M_{\Theta\Theta} = \mathbb{E} \left[ \text{IF}_\Theta(\tilde{x}_i) \text{IF}_\Theta(\tilde{x}_i)^\top \right],$$

$$M_{\Sigma\Sigma} = \mathbb{E} \left[ \text{IF}_{\Sigma_{\mathbf{u}|\tilde{x}}}(\tilde{x}_i) \text{IF}_{\Sigma_{\mathbf{u}|\tilde{x}}}(\tilde{x}_i)^\top \right],$$

$$M_{\psi\Theta} = \mathbb{E} \left[ \psi(\beta, \Theta \tilde{x}_i, \Sigma_{\mathbf{u}|\tilde{x}}) \text{IF}_\Theta(\tilde{x}_i)^\top \right],$$

$$M_{\psi\Sigma} = \mathbb{E} \left[ \psi(\beta, \Theta \tilde{x}_i, \Sigma_{\mathbf{u}|\tilde{x}}) \text{IF}_{\Sigma_{\mathbf{u}|\tilde{x}}}(\tilde{x}_i)^\top \right],$$

$$M_{\Theta\Sigma} = \mathbb{E} \left[ \text{IF}_\Theta(\tilde{x}_i) \text{IF}_{\Sigma_{\mathbf{u}|\tilde{x}}}(\tilde{x}_i)^\top \right].$$

# Varianza asintótica H (caso lineal)

$$H = \sigma_y^2 \left( 1 + (\mathbf{x}_N - \mu)^T \Theta^T (\Theta L \Theta^T)^{-1} \Theta (\mathbf{x}_N - \mu) \right) + C_\Theta \mathbb{E}[\tilde{r}_i(\tilde{\mathbf{x}}_i) \tilde{r}_i(\tilde{\mathbf{x}}_i)^T] C_\Theta^T,$$

with

$$C_\Theta = -(\beta^T \otimes (\mathbf{x}_N - \mu)^T Q_{\Theta(L)}).$$

	$n$	A_naive	A_corr	B_naive	B_corr	Faber
Setting 1 (weak signal)	500	0.919	0.948	0.913	0.938	0.912
Setting 1 (weak signal)	100	0.901	0.920	0.842	0.857	0.866
Setting 2 (strong signal)	500	0.940	0.947	0.941	0.946	0.932
Setting 2 (strong signal)	100	0.935	0.938	0.935	0.939	0.92

# Algunas ideas finales...

- Usamos Teoría Asintótica para probar existencia de distribuciones límite, es decir,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

- Bootstrap no la reemplaza, es un mecanismo numérico para aproximar una cantidad teórica cuya existencia ya fue demostrada
- Cuando hacemos simulaciones tipo Monte Carlo (cálculo de coverage) hacemos una validación empírica: queremos saber qué tan buena es la aproximación en un contexto práctico (muestras con tamaño finito razonable)



UNL. FACULTAD DE  
INGENIERÍA QUÍMICA



# MUCHAS GRACIAS

---

Curso de Estadística – MADATOP 2026