

Herramientas de Diseño Estadístico e Interpretación de las Mediciones

Usando R

Carlos J. Gómez
Magíster en Estadística
Bioquímico
Químico



FACULTAD DE
CIENCIAS QUÍMICAS
Y FARMACÉUTICAS
UNIVERSIDAD DE CHILE

Objetivo

El alumno será capaz de aplicar métodos estadísticos *modernos* en distintos problemas de la química analítica implementando las técnicas en lenguaje R. Desarrollará el pensamiento estadístico crítico para evaluar información cuantitativa en ciencias de las mediciones.



¿Qué veremos en la clase online?

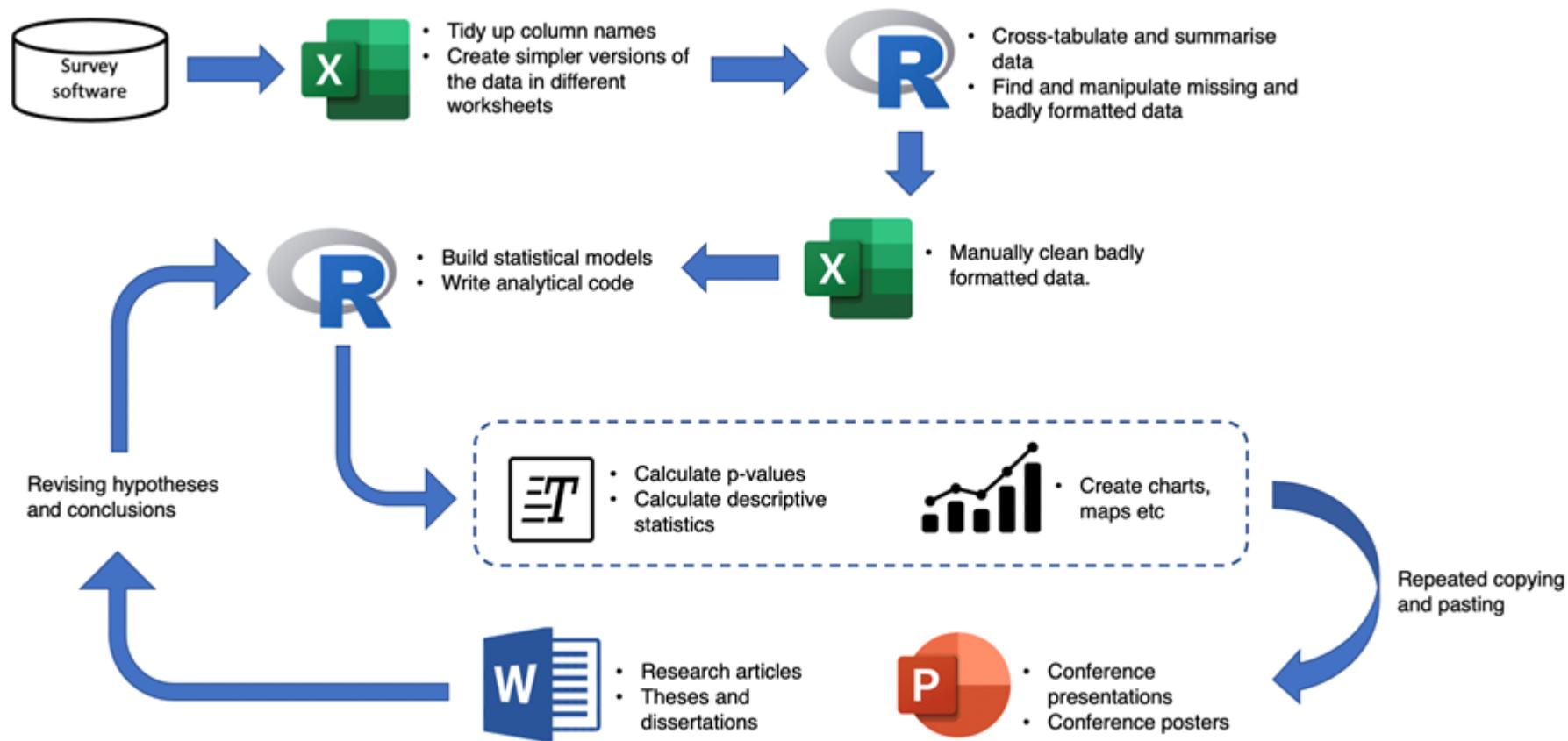
1. Introducción al lenguaje de programación **R**
2. Métodos de simulación
3. Investigación reproducible usando **rmarkdown**
4. Análisis exploratorio de datos
5. Análisis de varianza (Estudios r&R)
6. Calibración lineal

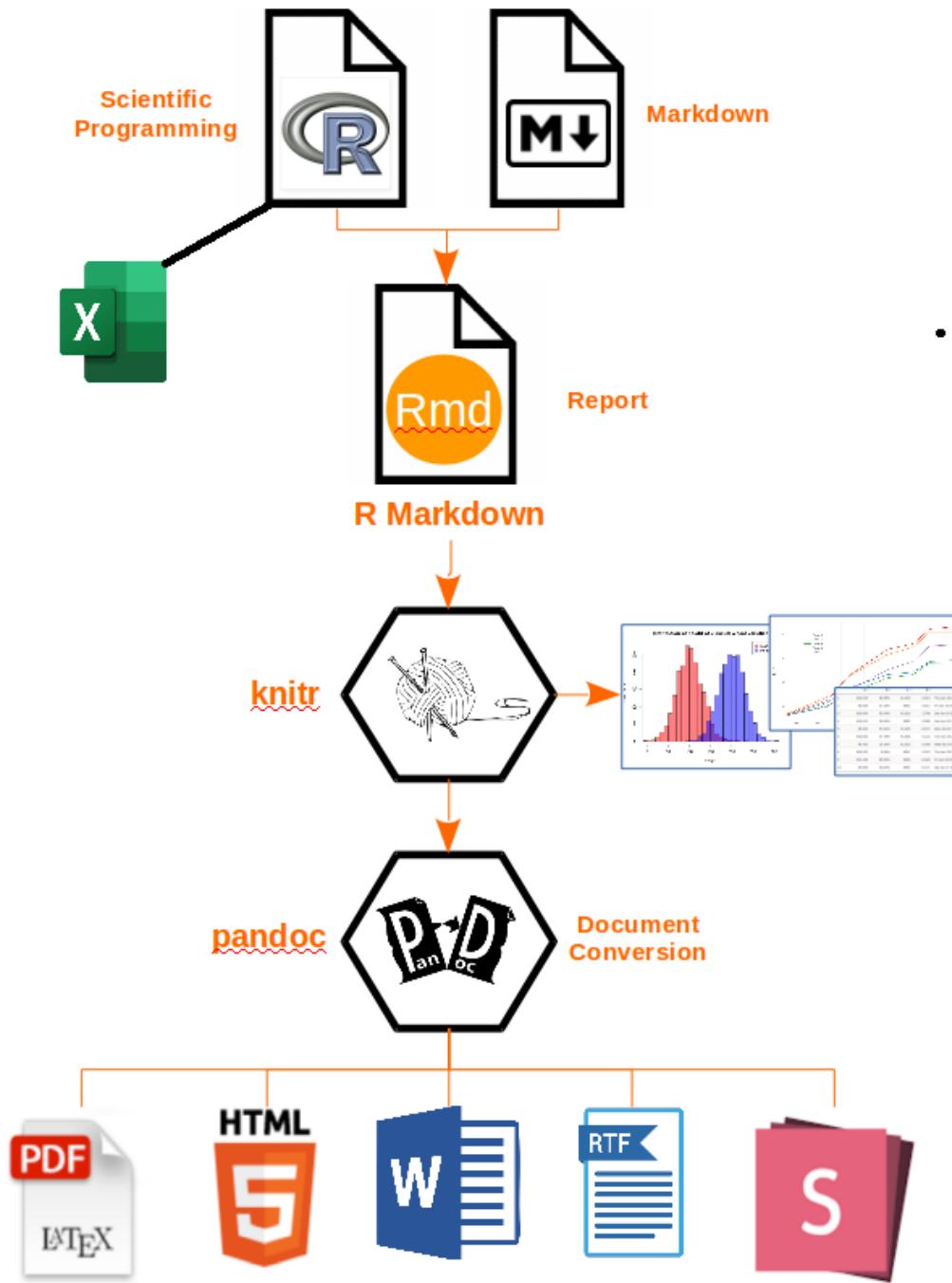


Introducción a R

¿y por qué usar R ?... si con Excel en suficiente (¿o no?)

El ciclo del *copiar y pegar*.





Se asume que...

- Maneja Excel a nivel usuario (ingresar una fórmula, *linkear celdas...*)
- Tiene un conocimiento básico de estadística
- Tiene alguna experiencia en estimar incertidumbre de medición



Funciones o comandos en R: Ejemplo `round()`

round (x, digits)

Vector de datos

Número de decimales que se redondeará el vector/dato x

```
x <- 23.678446  
round(x, digits = 2)
```

```
[1] 23.68
```



R packages (+18,000)

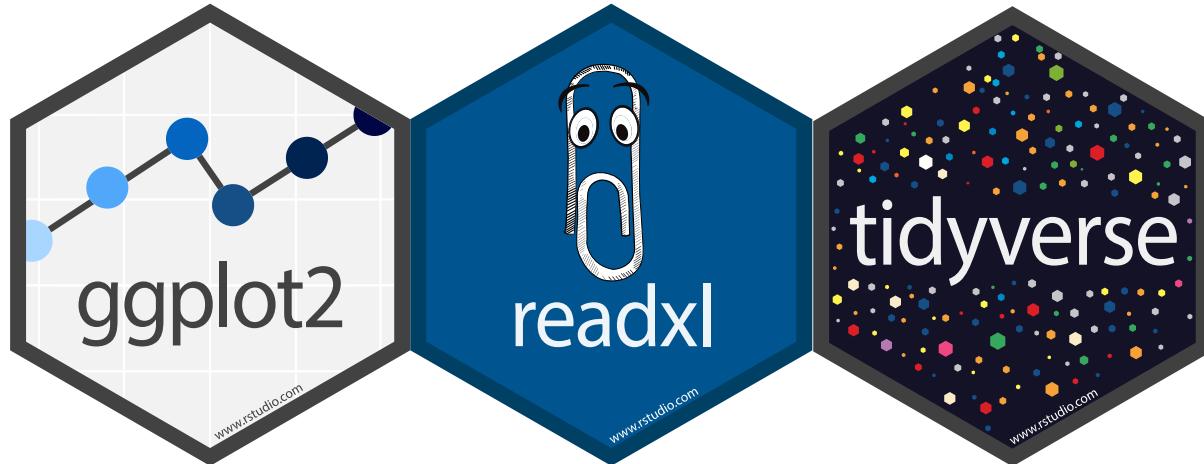
"An R package is a collection of functions, data, and documentation that extends the capabilities of base R. Using packages is key to the successful use of R."

<https://cran.r-project.org/>

R packages (+ 18,000)



En este curso usaremos algunos:

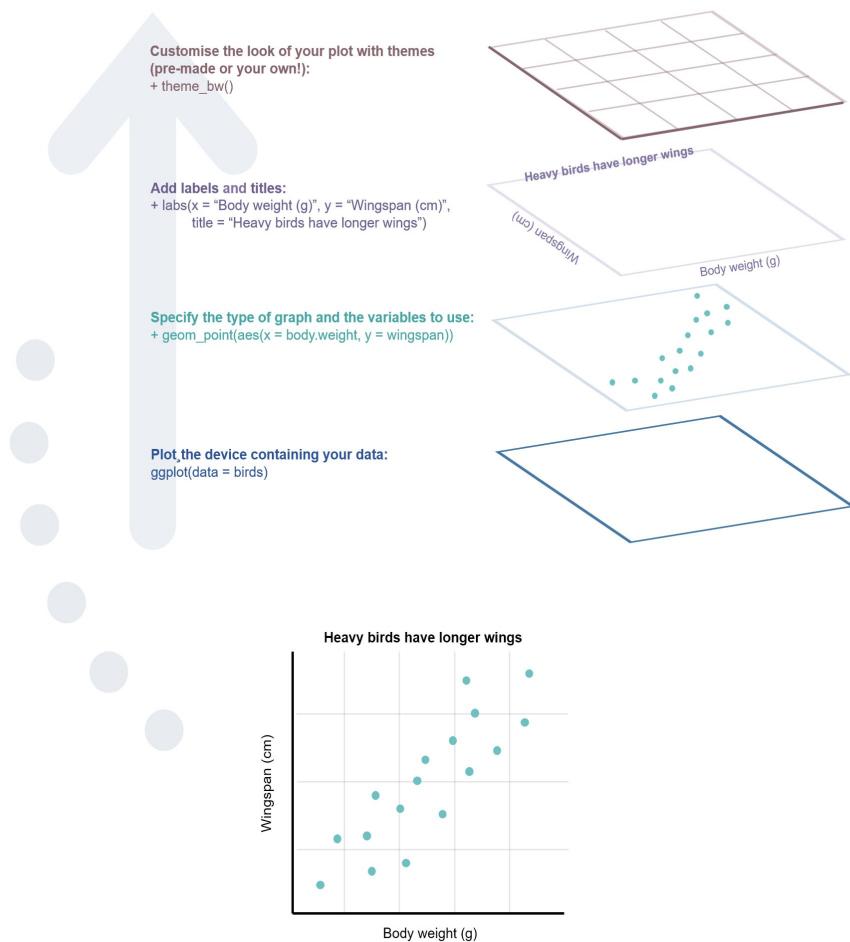


R posee varias herramientas para simulación aleatoria. Usaremos el comando `rnorm` para simular datos normales:

```
# Guardaremos en x 100 datos aleatorios normales con media 120 y desviación  
# estándar 10  
  
x <- rnorm(100, 120, 10)
```



MAKING A GRAPH WITH GGPLOT2



"Grammar of Graphics"

- Concepto desarrollado por Leland Wilkinson (1999)
- **ggplot2** desarrollado por Hadley Wickham (2005)





All models are wrong, but some are useful.

— *George E. P. Box* —

Si todos los modelos son incorrectos,
entonces...

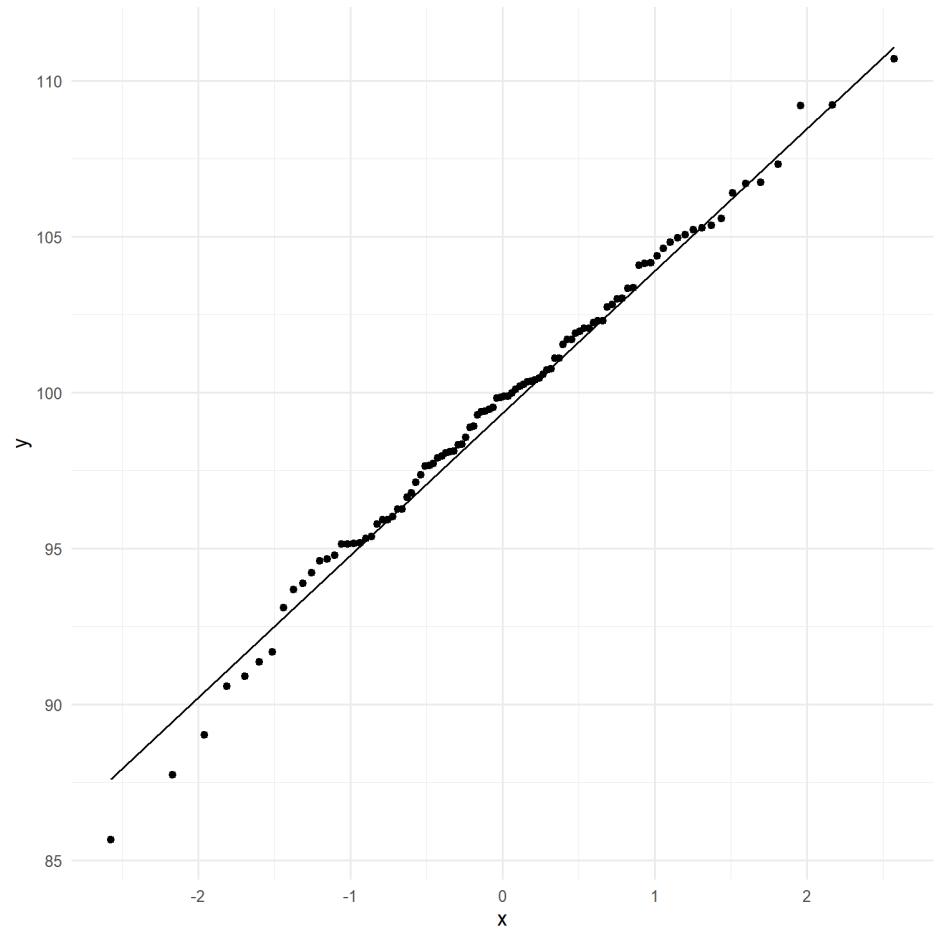
¡No existen los datos experimentales normales!

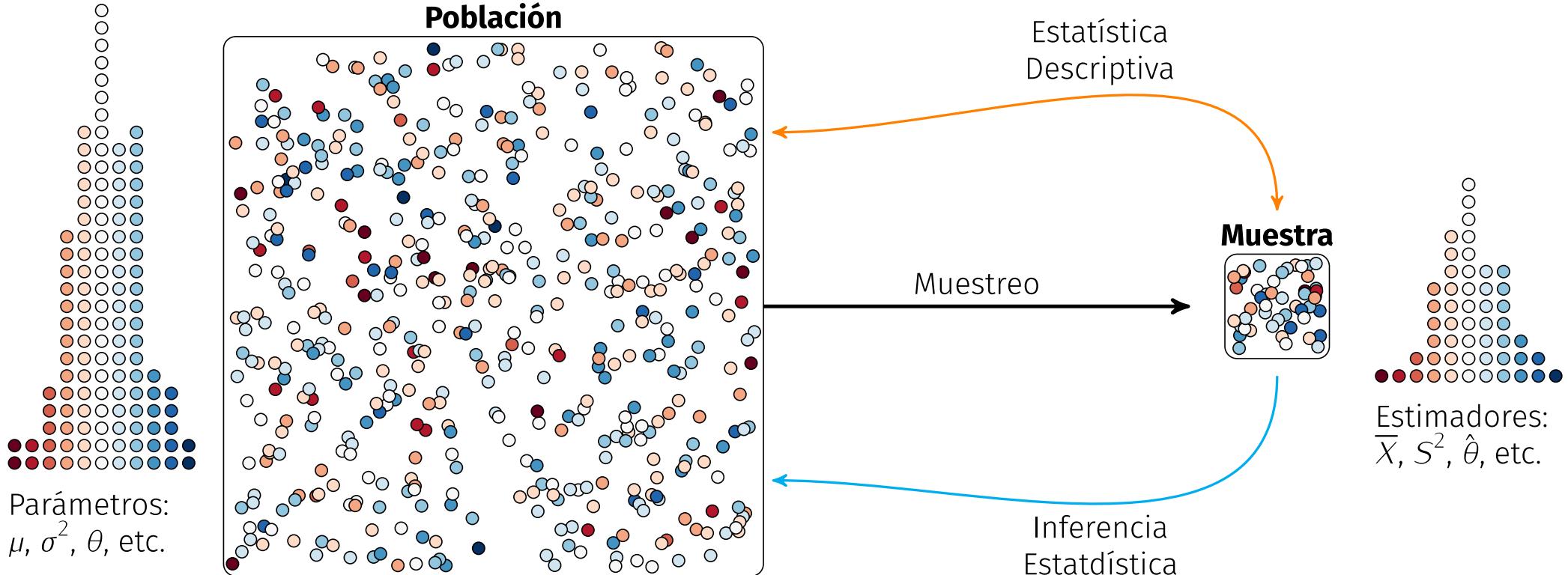


Test de Normalidad de Shapiro

Shapiro-Wilk normality test

```
data: x  
W = 0.9915, p-value = 0.7838
```





Cortesía de <https://github.com/walmes/Tikz>

Cómo se interpreta el IC [93.1 -- 106.7]

Sólo para incertidumbres tipo A

NO significa que exista un 95% de probabilidad de que la verdadera media esté dentro del intervalo [93.1 - 106.7]

Tu Turno

</>

05 : 00

Objetivo: utilizar el método bootstrap para estimar la incertidumbre estándar de un parámetro de datos no normales

Estimar la incertidumbre estándar de la **mediana** de los datos de la variable **Temperatura** del archivo **datos.xlsx** hoja **temperatura**, utilizando el método bootstrap con el package **simpleboot**. Guarde los datos en **R** con el nombre **datos.temp**. Haga **R = 2000** bootstraps y guarde todo el análisis como **mediana.boot**

Pistas:

```
library(readxl) # para importar los datos desde Excel a R  
datos.temp <- read_excel('datos.xlsx', sheet = 'aquí la hoja')  
  
library(simpleboot) # hará los cálculos de bootstrap  
# ¿Cómo extraigo el vector de datos de temperatura?$$$$  
  
mediana.boot <- one.boot(aquí va el vector, FUN = m..., R = 2000)
```



Incertidumbre de calibración lineal

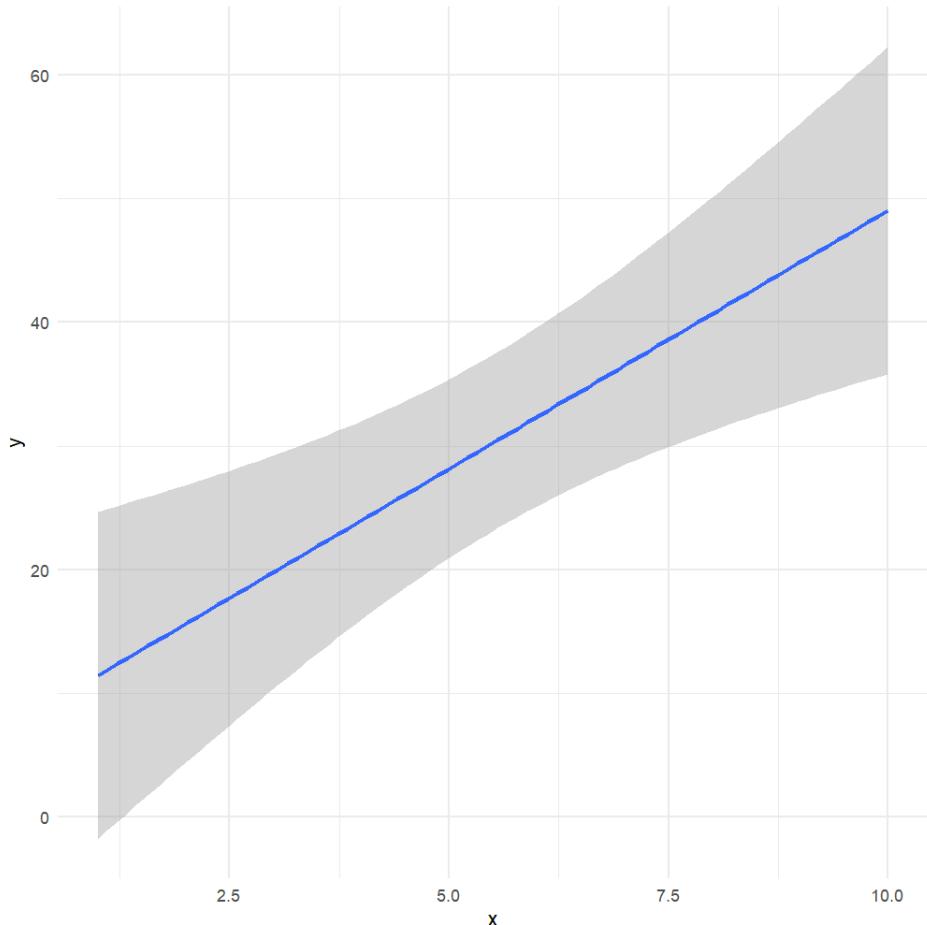
$$u(x_0) = \frac{\sigma_{y/x}}{\beta_1} \sqrt{\frac{1}{n} + \frac{1}{m_0} + \frac{(x_0 - \bar{x})^2}{\sum_i^n (x_i - \bar{x})^2}}$$

donde:

- $\sigma_{y/x}$ es la desviación estándar del error aleatorio ϵ o error de calibración
- n es el número de calibrantes independientes
- m_0 es el número de replicados independientes de la muestra problema
- \bar{x} es el promedio de las concentraciones de los calibrantes

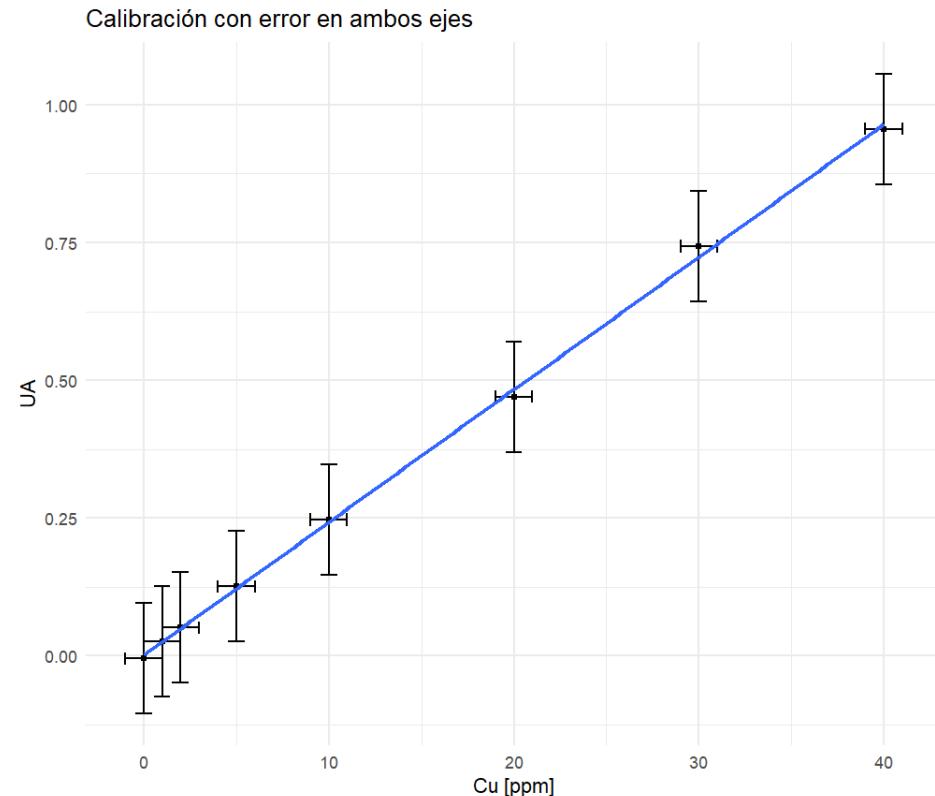
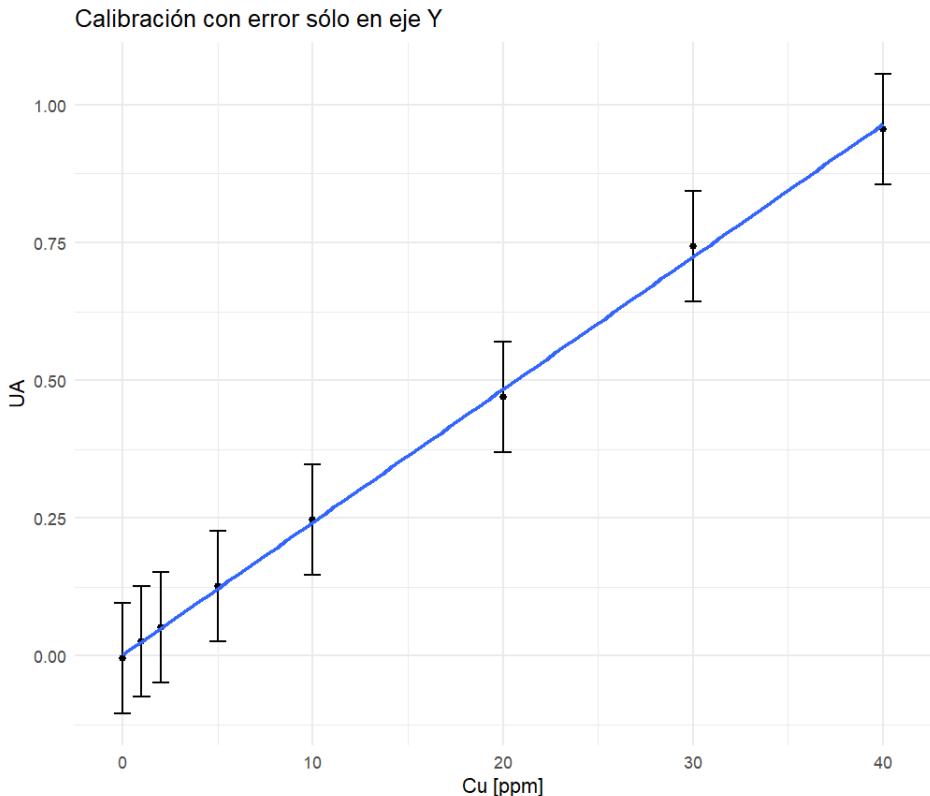


Incertidumbre de calibración lineal



- Esta incertidumbre **solo** refleja el error instrumental
- **NO** incluye la incertidumbre de los calibrantes
- Si se desea incluir la incertidumbre de los calibrantes se debe emplear el modelo "Regresión con error en ambos ejes"

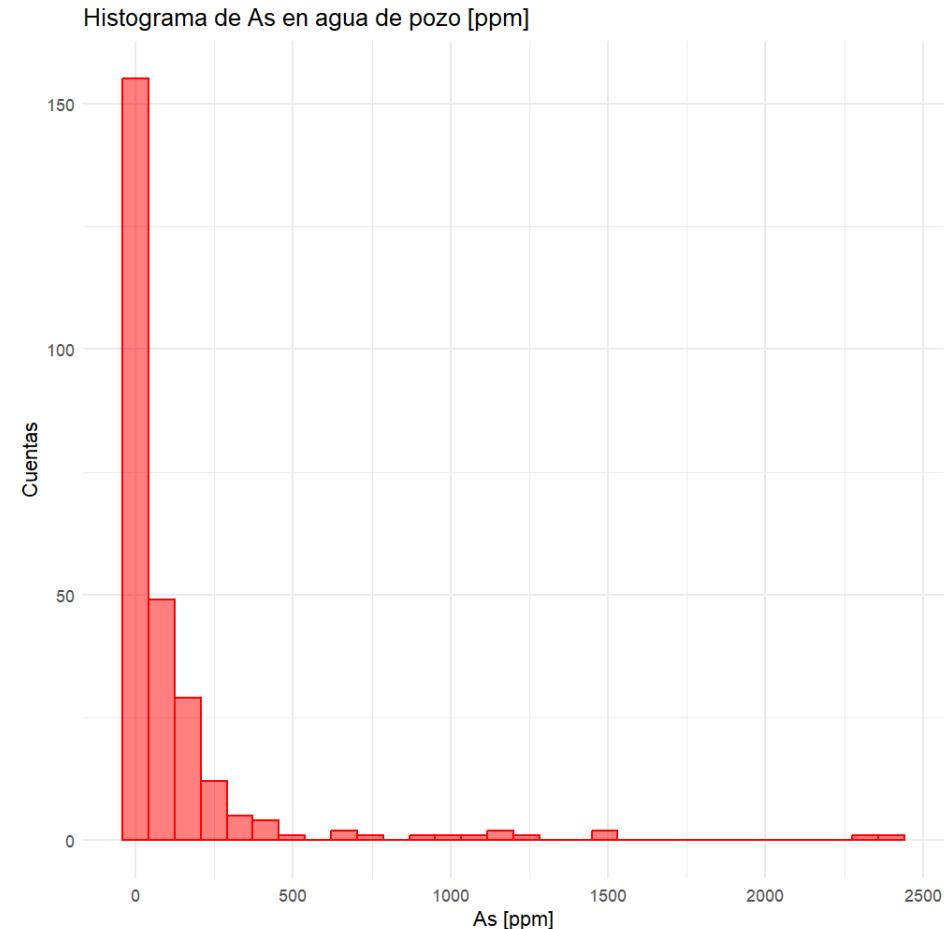
Esta incertidumbre sólo considera la incertidumbre del instrumento, no incorpora la incertidumbre de los calibrantes. Se debe utilizar el método de *Regresión con error en ambos ejes*



Análisis de datos NO Normales

Histograma

```
ggplot(datos.no.normales, aes(x = As)) +  
  geom_histogram(col = 'red', fill = 'red')  
  ggttitle('Histograma de As en agua de pozo')  
  xlab('As [ppm]') +  
  ylab('Cuentas') +  
  theme_minimal()
```

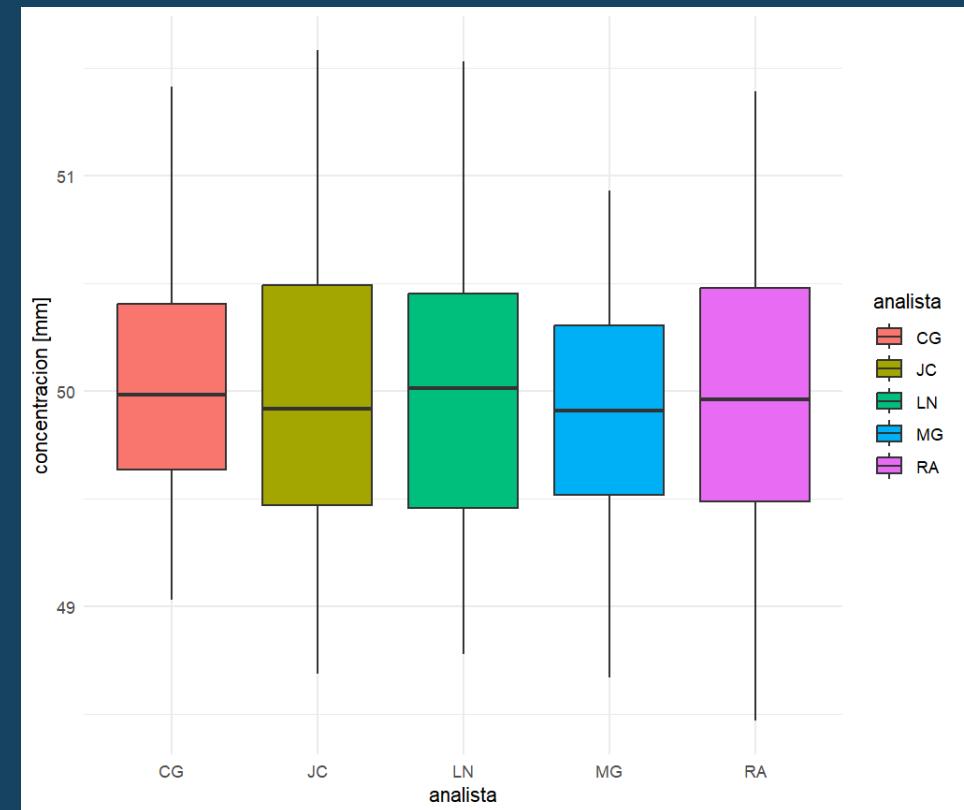
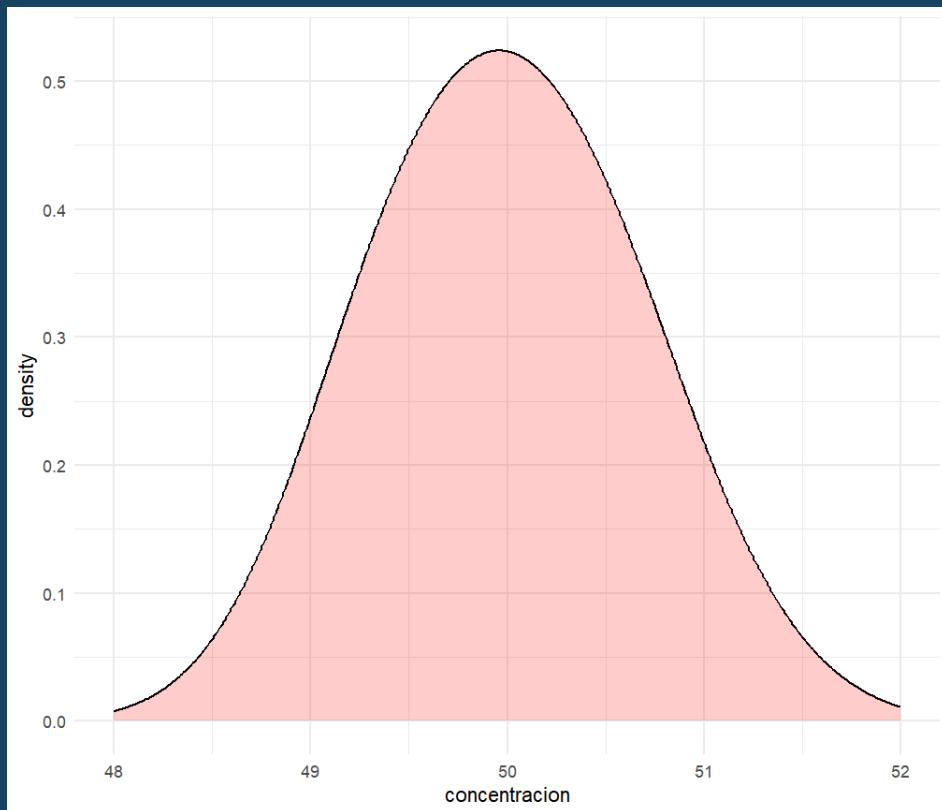


Si dividimos estas sumas de cuadrado por sus g.l obtenemos las varianzas

$$\underbrace{\sum \sum (y_{ij} - \bar{y})^2}_{\text{variabilidad total}} = \underbrace{\sum \sum (\bar{y}_i - \bar{y})^2}_{\text{entre-grupos}} + \underbrace{\sum \sum (y_{ij} - \bar{y}_i)^2}_{\text{dentro-grupo}}$$
$$\sigma_{\text{total}}^2 = \sigma_{\text{entre-grupos}}^2 + \sigma_{\text{dentro-grupo}}^2$$



Si NO hubiesen diferencias entre los analistas (grupos), entonces:
¿Cómo serían $\sigma^2_{\text{entre-grupos}}$ y $\sigma^2_{\text{dentro-grupos}}$?



Si no hubiesen diferencias entre los analistas (grupos) los $\tau_i = 0$

$$\sigma_{\text{total}}^2 = \sigma_{\text{entre-grupos}}^2 + \sigma_{\text{dentro-grupo}}^2$$

$$F_{\text{obs}} = \frac{\sigma_{\text{entre-grupos}}^2}{\sigma_{\text{dentro-grupo}}^2} = 1$$

