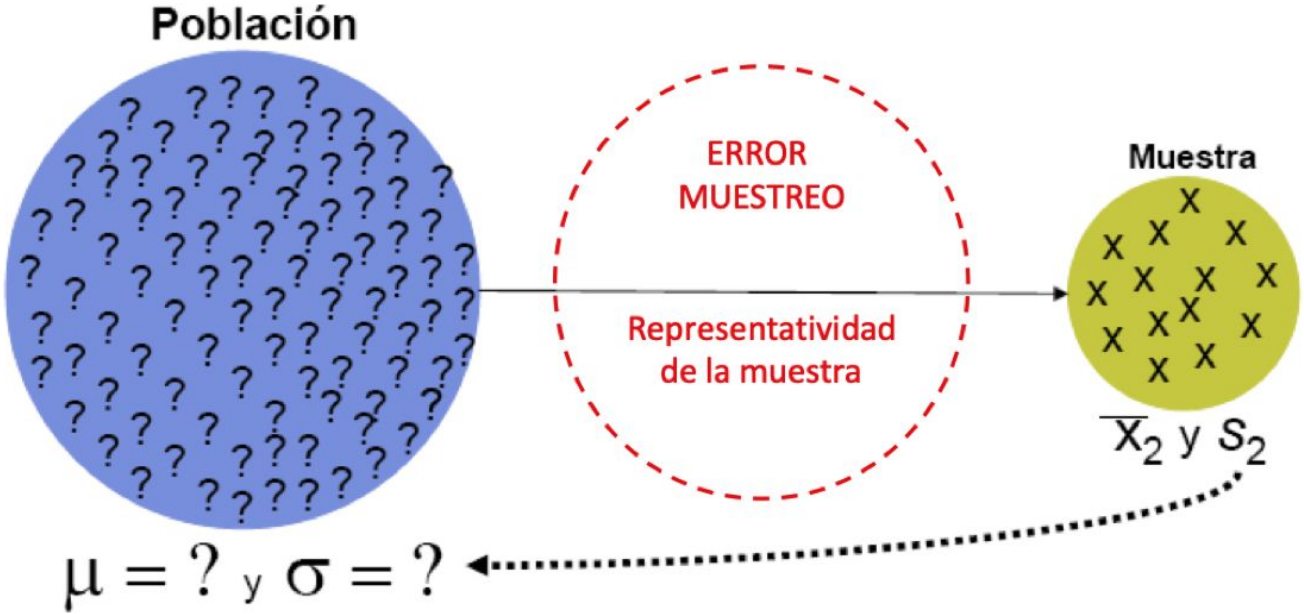
Repaso IC y
Test de hipótesis

¿Qué es la inferencia estadística?

La estadística inferencial busca elaborar conclusiones de la población a partir de una muestra:



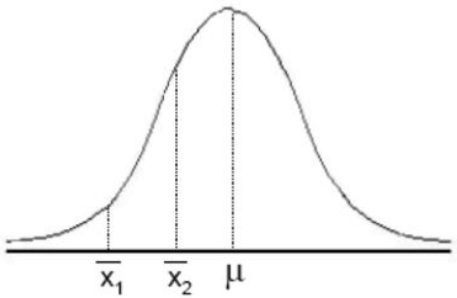
> La muestra no es un fiel espejo de la población (hay un error por el hecho de trabajar con muestras).
>
> Las distribuciones muestrales contienen probabilidades de que el estadístico muestral sea diferente al parámetro poblacional.

Inferencia Estadística: Estimación de parámetros

- **Estimación puntual**: es la estimación que utiliza el valor del estadístico muestral como indicador del valor del parámetro.

¿Cuál es el problema de la estimación puntual?

La estimación puntual no incorpora la variabilidad del estadístico entre distintas muestras: esta variabilidad es la probabilidad de lo que sucede en mi muestra no sea lo que ocurre en la población.



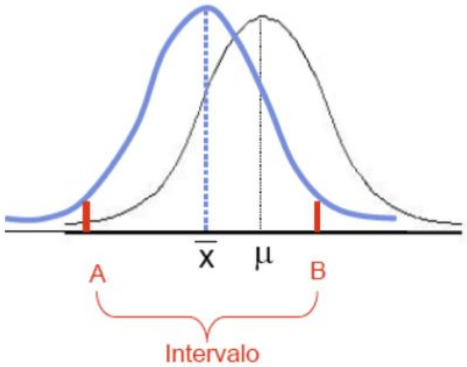
De todas las muestras posibles, puede ser que la media sea: $\bar{x}_1$ ó $\bar{x}_2$.

Ambas medias son diferentes al parámetro, pero hay una que está más próxima.

Para poder estimar qué pasa en la población, se debe incorporar la variabilidad del estadístico muestral.

Inferencia Estadística: Estimación por intervalo

La estimación por intervalo del parámetro poblacional es aquella que identifica un rango de valores entre los que **puede** situarse el parámetro poblacional. Estos valores se determinan a partir de la muestra.



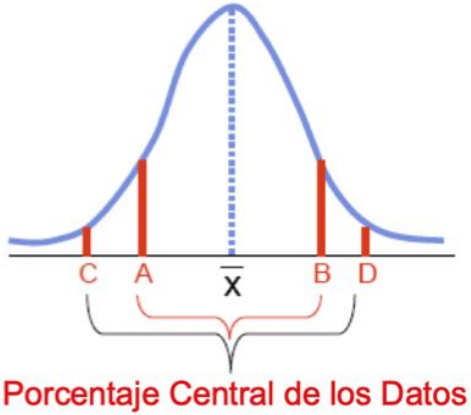
El estadístico muestral **puede estar en cualquier punto de la distribución muestral**.

El intervalo es un rango determinado a partir de valores de la muestra (tamaño, desviación estándar y media) que tiene una **probabilidad conocida de comprender el parámetro**.

La probabilidad de que el intervalo contenga el parámetro se calcula por el área bajo la curva.

Inferencia Estadística: Nivel de confianza y significancia

¿Cómo se determina entre qué rangos de valores está el parámetro?. La respuesta depende de cuán confiados deseemos estar de que el parámetro está en ese rango. De esta forma:

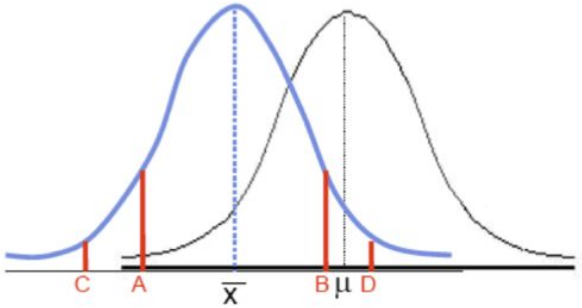


Nivel de confianza ($1-\alpha$): Es la probabilidad de que la muestra elegida produzca un intervalo que incluya el parámetro que se está estimando.

Nivel de significancia (α): Probabilidad de que el intervalo calculado a partir de la muestra no comprenda el parámetro.

Inferencia Estadística: Nivel de confianza y significancia

Mientras más grande el nivel de confianza, el rango se amplía y por lo tanto aumenta la certeza (la probabilidad de incluir el parámetro).



Entre A y B está el 95% de los datos.

Entre C y D está el 99,7% de los datos.

Inferencia Estadística: Intervalos de confianza

Estimador puntual es una valor posible para un parámetro poblacional

¿Y si usamos un rango de valores?

El intervalo de confianza es una estimación intervalar que busca capturar un paramétro poblacional (desconocido) bajo una cierta probabilidad en muestras repetidas o nivel de confianza. Si consideramos un nivel de confianza ($1-\alpha$) del 95% podemos obtener una medida de los posibles valores para el parámetro poblacional de la siguiente manera:

$$Estimador\ Puntual\ \pm\ 1.96 * EE$$

Donde EP es el estimador puntual y EE el error estándar. En este caso el estimador está a $\pm$ 1.96 desviaciones estándar del parámetro en un 95% de las veces.

SUPUESTO:

1. Los datos provienen de una muestra aleatoria de la población.
2. Normalidad en la distribución o tamaños muestrales grandes (por Teorema del Límite Central)

Inferencia Estadística: Intervalos de confianza

Error estándar nos da una guía del tamaño del intervalo de confianza

Podemos tener (aprox.) un 95% de confianza en que hemos encontrado el parámetro poblacional si:

$$Estimador\ Puntual\ \pm\ 1.96 * EE$$

Notas:

> Aproximadamente verdad, se cumple muy bien para distribuciones normales y muestras grandes
>
> El estimador está a menos de 1.96 desviaciones estándar del parámetro poblacional un 95% de las veces

Intervalo de confianza para medias

Intervalo de medias con desviación poblacional conocida (σ) y muestra grande

$$IC = \bar{x} \pm z_{(\alpha/2)} * \frac{\sigma}{\sqrt{n}}$$

Estimación puntual

Error máximo de estimación

$$IC_{\text{inferior}} = \bar{x} - z_{(\alpha/2)} * \frac{\sigma}{\sqrt{n}}$$

Límite inferior

$$IC_{\text{superior}} = \bar{x} + z_{(\alpha/2)} * \frac{\sigma}{\sqrt{n}}$$

Límite superior

$$\text{Error máximo de estimación} = z_{(\alpha/2)} * \frac{\sigma}{\sqrt{n}}$$

Coeficiente de Confianza

Error Estándar

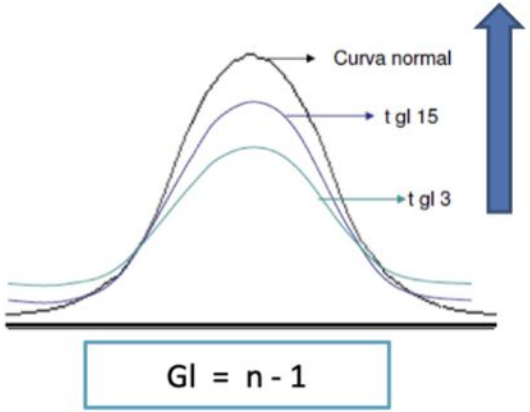
Intervalo de confianza para medias:

Intervalo de medias con desviación poblacional desconocida (σ) y muestra pequeña o grande

Recordemos algunas propiedades de la distribución t de student:

- t está distribuida con media 0 y simétricamente alrededor de la media.
- El área bajo la curva es igual a 1.
- t forma una familia de distribuciones (una distribución diferente para cada gl cuando estos son mayores o iguales a 1)
- La distribución t es menos puntiaguda en la media y más ancha en las colas que una distribución normal.
- La distribución t se aproxima a la normal estándar a medida que aumentan los grados de libertad (Gl=121).

Distribución normal estándar y Distribución t de student



Intervalo de confianza para medias:

Intervalo de medias con desviación poblacional desconocida (σ) y muestra pequeña o grande

$$IC_{inferior} = \bar{x} - t_{(gl,\propto/2)} * \frac{S}{\sqrt{n}}$$

Límite inferior

$$IC_{superior} = \bar{x} + t_{(gl,\propto/2)} * \frac{S}{\sqrt{n}}$$

Límite superior

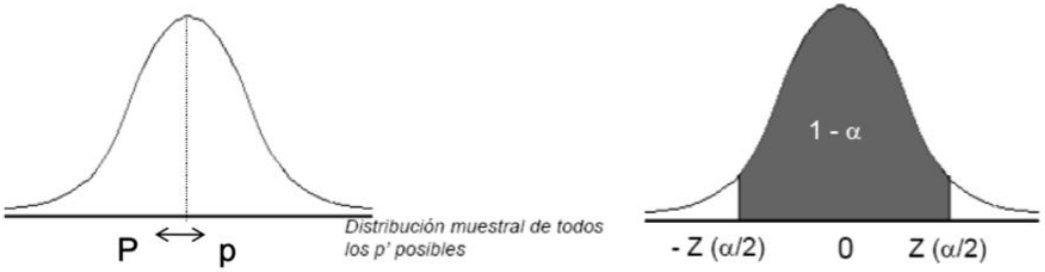
$$\text{Error máximo de estimación} = t_{(gl,\ \alpha/2)} * \frac{s}{\sqrt{n}}$$

Nivel de Confianza

Error Estándar

Intervalos de confianza para proporciones

En el caso de variables nominales u ordinales (categóricas), la estimación de intervalos de confianza se basa en el cálculo de proporciones.



Hay que considerar, además, que la distribución de probabilidades en una distribución binomial se asume normal cuando:

$$n > 20;\ n * p > 5;\ n * q > 5$$

Para localizar el intervalo:

1. Se localiza el coeficiente de confianza $z_{(\alpha/2)}$.
2. Se posicionan $-z_{(\alpha/2)}$ y $+z_{(\alpha/2)}$ que representan los valores críticos entre los que se encuentra una probabilidad de ($1-\alpha$) de incluir el parámetro.

Intervalos de confianza para proporciones

El intervalo de confianza $1-\alpha$ para la estimación de la proporción poblacional "P", se elabora del siguiente modo:

$$IC_{\text{inferior}} = p - z_{(\alpha/2)} * \sqrt{\frac{p*q}{n}}$$

Límite inferior

$$IC_{\text{inferior}} = p + z_{(\alpha/2)} * \sqrt{\frac{p*q}{n}}$$

Límite superior

$$\text{Error máximo de estimación} = z_{(\alpha/2)} * \sqrt{\frac{p*q}{n}}$$

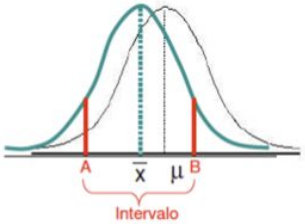
Coeficiente de Confianza

Error Estándar

PH: Conceptos fundamentales

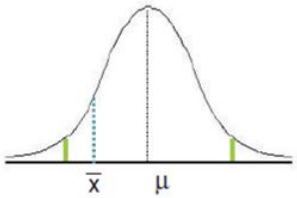
Intervalo de confianza

Los IC responden a la pregunta: **Si el resultado de mi muestra es** $\overline{x}$ ***¿Cuál es la probabilidad de que a partir de mi muestra elabore un intervalo que comprenda el valor del parámetro poblacional?***



Prueba de hipótesis

En las pruebas de hipótesis la pregunta es: **Si la hipótesis nula es verdadera** ***¿Cuán probable es obtener un promedio muestral como el mío?***



Las **pruebas de hipótesis** son otro método de estimación de parámetros. A través de ellas, hay dos tipos de problemas que se intenta resolver:

1. Determinar si es probable que un valor obtenido a partir de una muestra (media o proporción), pertenece realmente a una población.

Responde a la pregunta: **¿Cómo podemos determinar si el valor obtenido en una variable en una muestra se debe o no al azar (error muestral)?**

1. Determinar, en términos de probabilidad, si las diferencias observadas entre dos muestras significa que las poblaciones de las que se han obtenido las muestras son realmente diferentes.

Responde a la pregunta: **¿Cómo podemos determinar que las diferencias observadas entre dos muestras (grupos o variables) se deben o no al azar (error muestral)?**

PH: Fases de la prueba de hipótesis

1. Formular una hipótesis estadística

Hipótesis nula (Ho)

- Es aquella hipótesis que se formula y que por tanto se quiere contrastar.
- Es la hipótesis que considera la "igualdad".
- A partir de esta igualdad se puede determinar una distribución muestral.
- Esto se logra asignando un valor específico al *parámetro poblacional*.
- La prueba parte del supuesto que la hipótesis nula *es verdadera.*

Hipótesis alternativa (Ha, H_1)

- Corresponde a cualquier otra hipótesis que sea diferente; contraria a Ho.
- Es nuestra conclusión en los casos en que rechacemos la Ho.

Ambas hipótesis (nula y alternativa) deben abarcar todos los escenarios posibles

Siempre hay que redactar estas hipótesis haciendo referencia a la "población de donde proviene la muestra"

2. Elección de prueba estadística para rechazar hipótesis nula:

		Variable a comparar		
		Escalar	Dicotómica	Nominal
Grupos a comparar	Una Muestra (representatividad muestral – comparación con valor objetivo)	Prueba de hipótesis de medias para una muestra (distribución t de Student o normal)	Prueba de hipótesis de proporciones para una muestra (distribución binomial o normal)	Prueba chi cuadrado para una muestra
	Dos Muestras (relaciones entre dos variables)	Prueba de hipótesis de medias para dos muestras (distribución t de Student o normal, ANOVA – para más de dos grupos -)	Prueba de hipótesis de proporciones para dos muestras (distribución binomial o normal)	Prueba chi cuadrado para dos muestras

PH: Fases de la prueba de hipótesis

3. Especificar el nivel de significancia

Nivel de significancia: probabilidad de que una prueba estadística producirá un valor bajo el cual se rechazará $H0$, cuando de hecho es verdadera (probabilidad de cometer Error Tipo I)

p varía entre 0 y 1 → Es la probabilidad de que un suceso ocurra por azar

$\alpha = 0,05$

→ En el caso de ser verdadera la $H0$, sólo en un 5% de las veces (o menos) las diferencias observadas serán lo suficientemente grandes como para conducirnos erróneamente a la conclusión de $H1$.

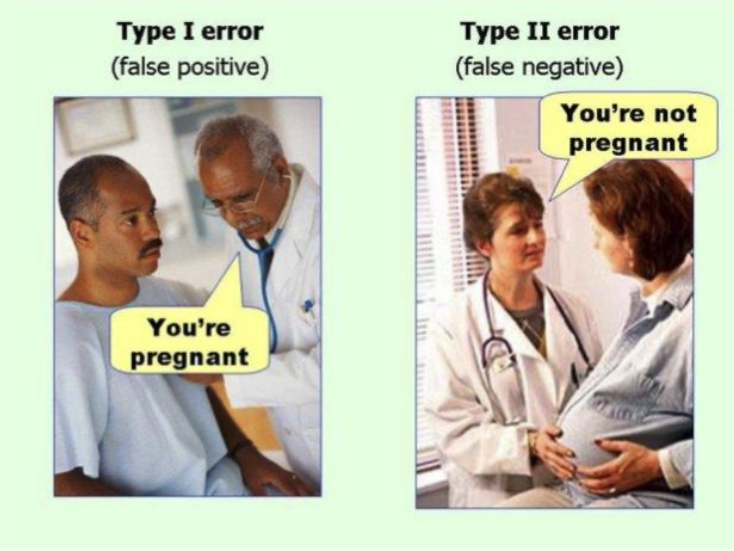
Resultados posibles de las decisiones de rechazo:

	La verdad desconocida sobre los parámetros	
Decisión del investigador:	Cuando la hipótesis nula es en realidad verdadera	Cuando la hipótesis nula es en realidad falsa
Rechazar la hipótesis nula	Error Tipo I (α)	Decisión correcta
No rechazar la hipótesis nula	Decisión correcta	Error Tipo II (β)

PH: Fases de la prueba de hipótesis

3. Especificar el nivel de significancia

Nivel de significancia: probabilidad de que una prueba estadística producirá un valor bajo el cual se rechazará $H0$, cuando de hecho es verdadera (probabilidad de cometer Error Tipo I)



¿Cuál es la hipótesis nula?

PH: Fases de la prueba de hipótesis

4. Determinar la distribución muetral

> Distribución normal
>
> Distribución binomial
>
> Distribución t de Student
>
> Distribución x^2
>
> Distribución F de Fisher

5. Determinar la región de rechazo

Consiste en un conjunto de valores posibles que son tan extremos que, cuando Ho es verdadera, la probabilidad de que la muestra se ubique ahí es muy pequeña (α). Se debe decidir si la prueba es unidireccional o bidireccional.

6. Calcular el valor de la prueba estadística

Calcular valores observados y compararlos con los valores teóricos, es distinto para cada prueba.

7. Concluir estadística y sustantivamente

Señalar el resultado de la prueba estadística e interpretar sustantivamente sus resultados.

PH para una muestra

Busca diferencias significativas entre la media $(\bar{x})$ o proporción (p) de una muestra y un valor de **referencia poblacional** $(\mu \; o \; P)$.

1. Medir representatividad
2. Comparar con otro valor de referencia (Ej: otros estudios en el área)

Los estadísticos de prueba para medias pueden provenir de distribuciones muestrales normales o distribuciones muestrales t de Student

Para distribuciones normales:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Para distribuciones t de Student:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Para pruebas de proporciones, si se cumple que n>20; n*p>5 y n*q>5, la distribución binomial se comporta como normal:

Para distribuciones binomiales con n>20, n*p>5 y n*q>5:

$$Z = \frac{p - P}{\sqrt{\frac{P * Q}{n}}}$$

Ejemplo: Proyecto STAR

- Usando diferencias de medias

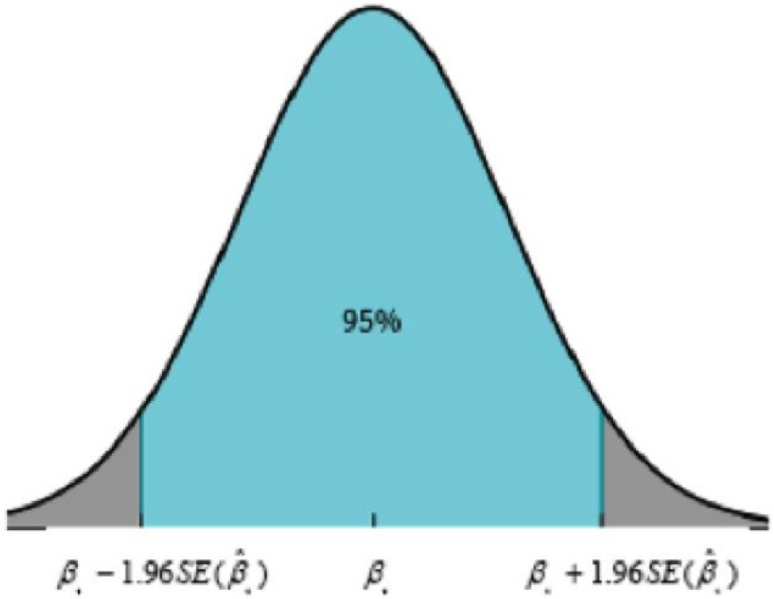
```
t.test(tscorek ~ str, alternative = c("two.sided"),
       data=df,
       paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

```
    Welch Two Sample t-test

data:  tscorek by str
t = -6.3768, df = 3129.1, p-value = 2.075e-10
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -17.965473  -9.515634
sample estimates:
mean in group 0 mean in group 1
       918.2013        931.9419
```

Distribución muestral de los estimadores MCO

Para muestras grandes, la distribución de $\hat{\beta}_1$ es normal con media igual al valor poblacional (verdadero) de la pndiente de la FRP, β_1. Esto se mantiene bajo ciertos supuestos que hablaremos en breve.



Distribución muestral de los estimadores MCO

- Tal como se calcula la varianza y error estándar de un promedio para armar los IC, ahora calculamos el error estándar de $\hat{\beta}_1$:

$$SE(\hat{\beta}_1) = s_{\hat{\beta}_1} = \sqrt{\frac{n}{n-2} \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 (\hat{\mu}_i)^2}{[\sum_{i=1}^{n} (X_i - \bar{X})^2]^2}}$$

- Intervalo de confianza para el 95% de significancia estadística: $\hat{\beta}_1 \pm 1.96 \cdot s_{\hat{\beta}_1}$
- Testeamos la hipótesis nula de que: $H_0 : \beta_1 = a$

$$t = \frac{\hat{\beta}_1 - a}{s_{\hat{\beta}_1}}$$

  - al 95% rechazamos la hipótesis nula si $t > 1.96$ para un test de dos colas
  - p-value: Si la hipótesis nula es verdad, es la probabilidad de obtener un test estadístico tan extremo o más extremo de lo que observamos

Ejemplo

```
Call:
lm_robust(formula = tscorek ~ ratio, data = df2, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
(Intercept)   698.93    10.3644  67.436 9.487e-227  678.560  719.306 418
ratio          -2.28     0.5195  -4.389  1.447e-05   -3.301   -1.259 418

Multiple R-squared:  0.05124 ,    Adjusted R-squared:  0.04897
F-statistic: 19.26 on 1 and 418 DF,  p-value: 1.447e-05
```

Trabajo en grupos

```
summary(lm_robust(bwght ~ monpre_bin, data = df, se_type = "stata"), digits=4)
```

```
Call:
lm_robust(formula = bwght ~ monpre_bin, data = df, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
(Intercept)   3378.8       26.7  126.59    0.000     3326   3431.2 1825
monpre_bin      29.7       30.9    0.96    0.337      -31     90.3 1825

Multiple R-squared:  0.000473 ,    Adjusted R-squared:  -7.42e-05
F-statistic: 0.921 on 1 and 1825 DF,  p-value: 0.337
```

Trabajo en grupos

```
m1 ← lm_robust(bwght ~ monpre_bin + cigs, data = df, se_type = "stata")
summary(m1)
```

```
Call:
lm_robust(formula = bwght ~ monpre_bin + cigs, data = df, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
(Intercept)   3397.8      26.97  126.00 0.000000   3344.9  3450.69 1715
monpre_bin      32.3      31.09    1.04 0.299036    -28.7    93.27 1715
cigs           -11.2       3.14   -3.58 0.000352    -17.4    -5.09 1715

Multiple R-squared:  0.00788 ,    Adjusted R-squared:  0.00672
F-statistic: 6.86 on 2 and 1715 DF,  p-value: 0.00108
```