

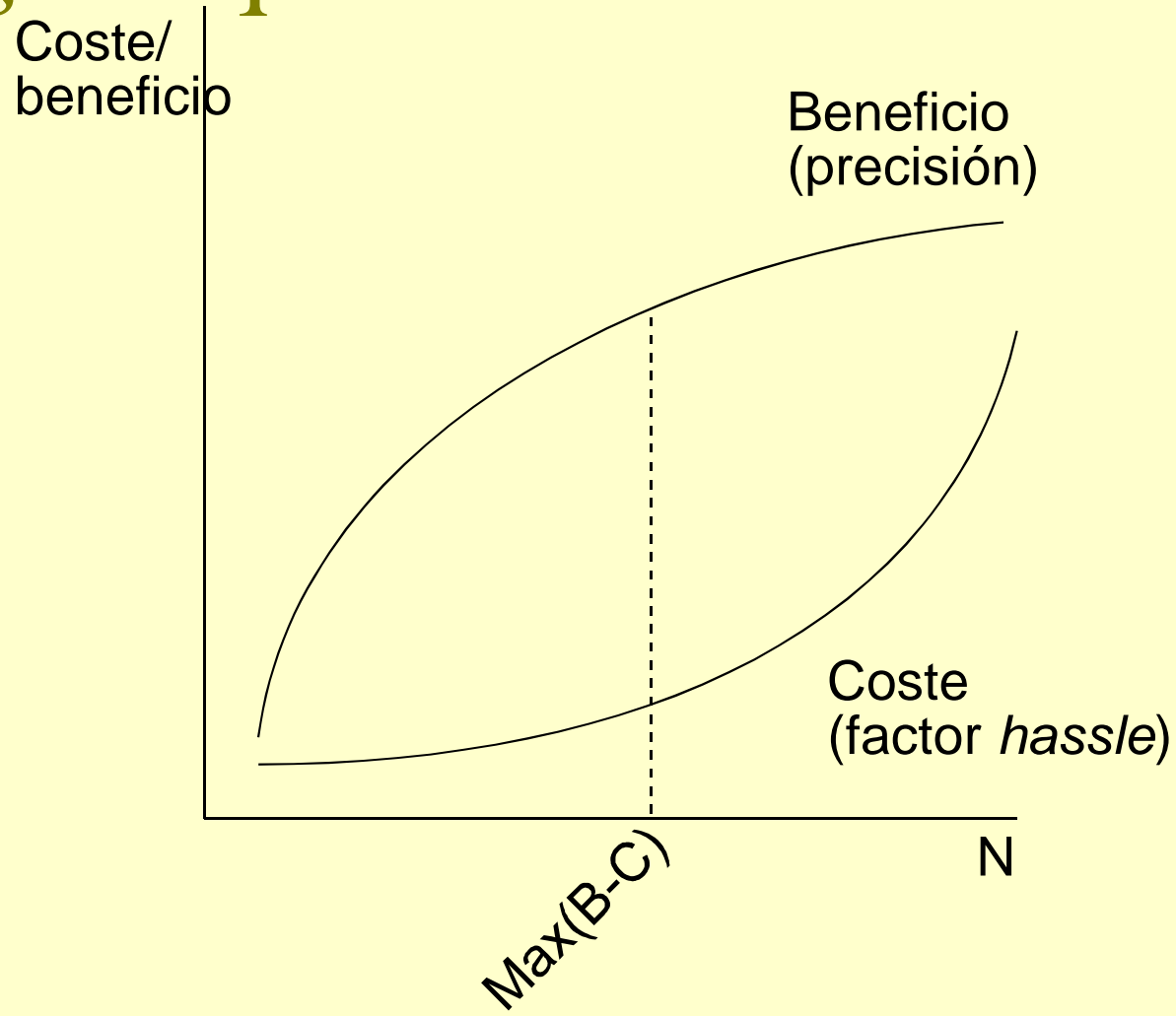
# Muestreo e inferencia

Calidad de los datos y las mediciones

# Razones para hablar de muestreo

- Formación académica de la población
- Comprender los datos que se van a utilizar
- Saber cómo obtener una muestra, si es necesario
- Realizar inferencias estadísticas

# ¿Por qué realizamos muestreos?



# ¿Cómo realizamos muestreos?

- Muestreo aleatorio simple
  - Variante: muestreo sistemático con inicio aleatorio
- Estratificado
- Conglomerado

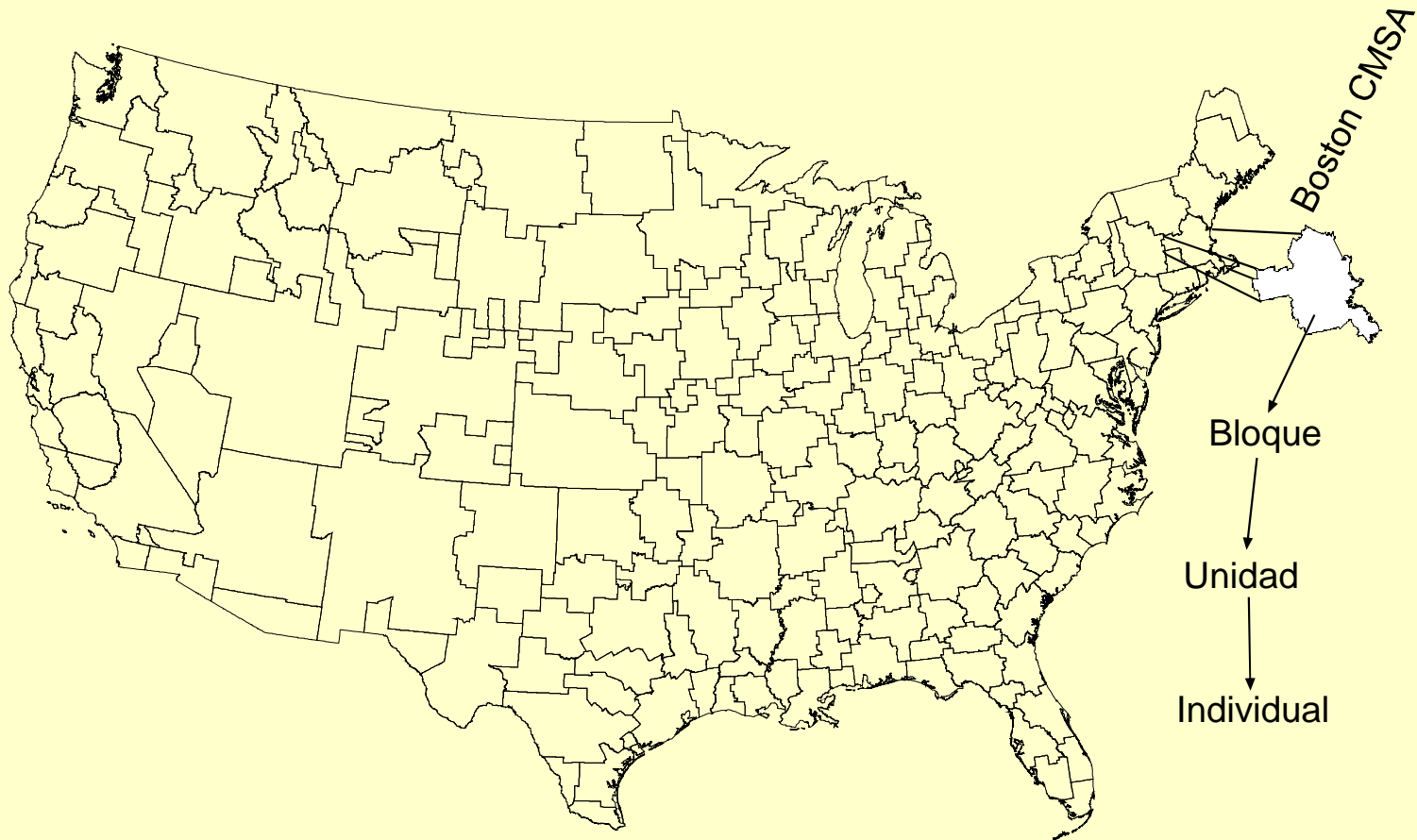
# Estratificación

- Dividir la muestra en subconjuntos o estratos, basados en características *conocidas* (raza, sexo, religión, continente, departamento)
- Beneficio: preservar o aumentar la variabilidad

# Ejemplo de estratificación

	NES		Ejemplo hipotético	
	N	s.e. @ 50%	N	s.e. @ 50%
Cristianos blancos	1.215	0,7%	350	1,3%
Cristianos negros	187	1,8%	350	1,3%
Judíos blancos	30	4,6%	350	1,3%
Judíos negros	2	17,7%	350	1,3%
Otra raza/religión	53	3,4%	87	2,7%
Ausentes	227	n.a.		
Total	1.714	0,6% (sobre 1.487 obs. válidas)		

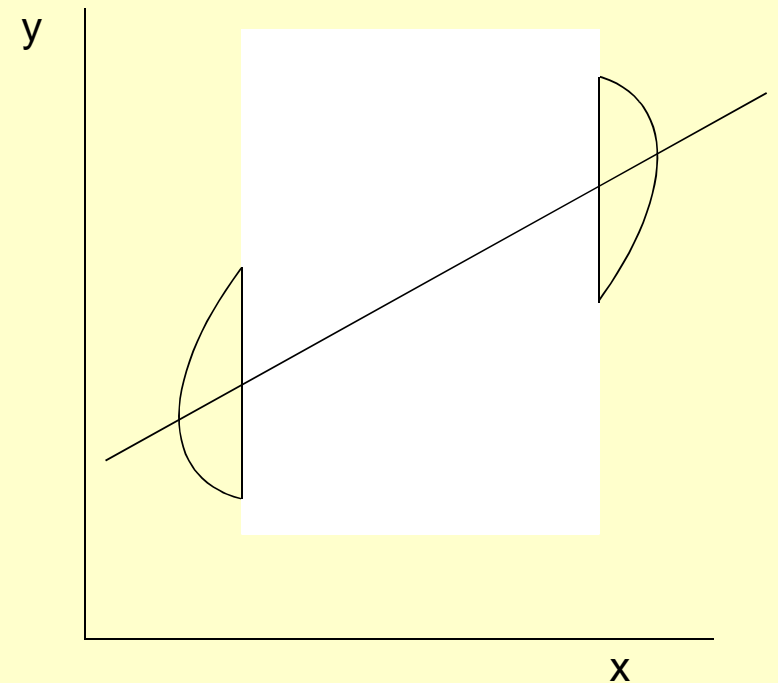
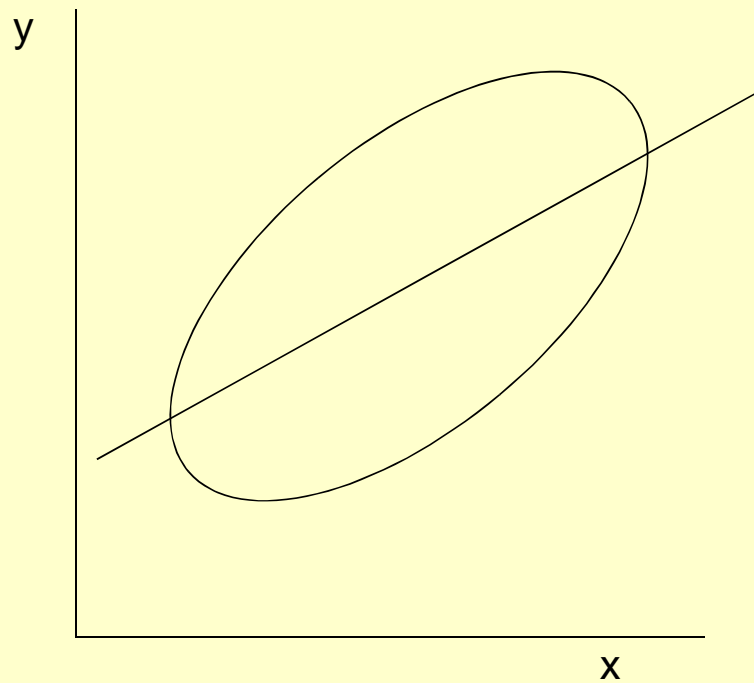
# Muestreo conglomerado



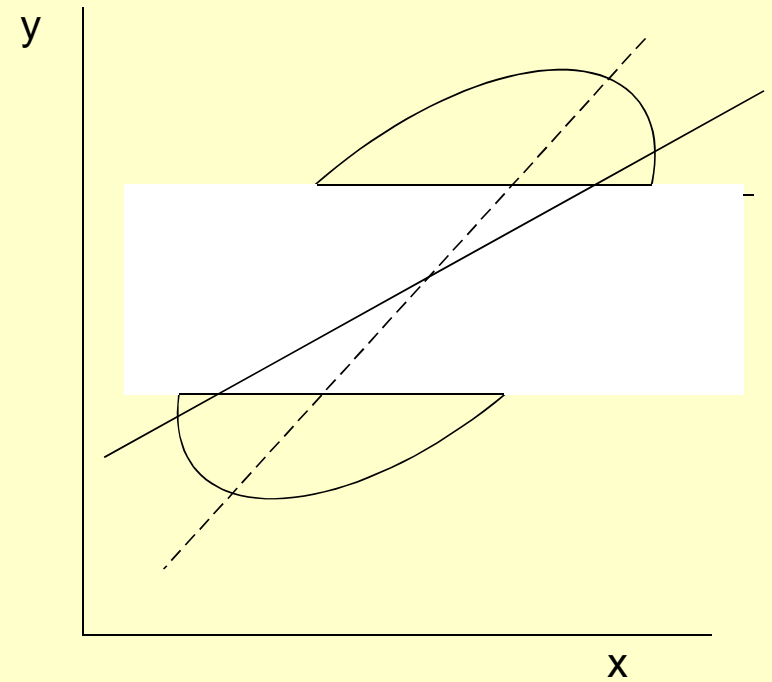
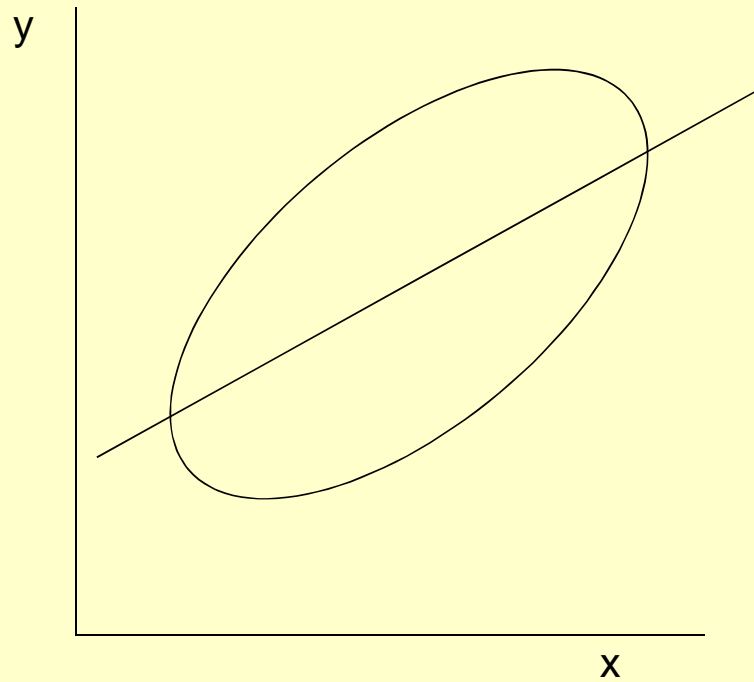
# Efectos de las muestras

- Obvio: influye en las marginales
- Menos obvio:
  - Permite el uso eficaz del tiempo y el trabajo
  - Efecto sobre las técnicas multivariadas
    - Muestreo de variable independiente: mayor precisión en cálculos de regresión
    - Muestreo sobre variable dependiente: sesgo

# Muestreo sobre variable independiente



# Muestreo sobre variable dependiente



# Muestreo

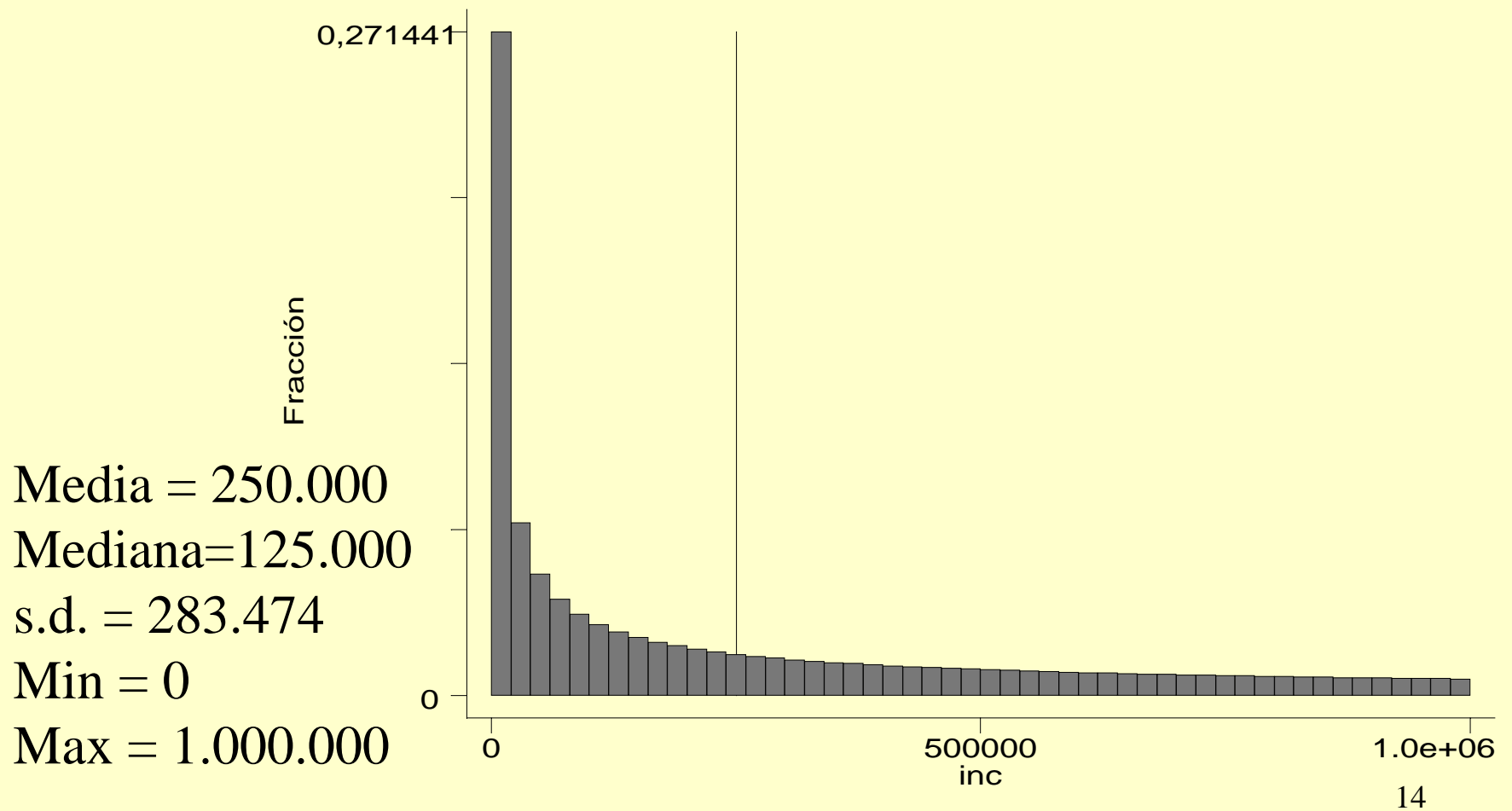
Consecuencias para la inferencia  
estadística

Inferencia estadística:  
aprender sobre lo desconocido a partir de lo  
conocido

- Razonamiento hacia adelante: distribuciones de medias de muestra, cuando se conoce la media de la población, la desviación estándar (*s.d.*), y  $n$
- Racionamiento hacia atrás: aprender sobre la media de la población cuando sólo se conoce la muestra, la desviación estándar, y  $n$ .

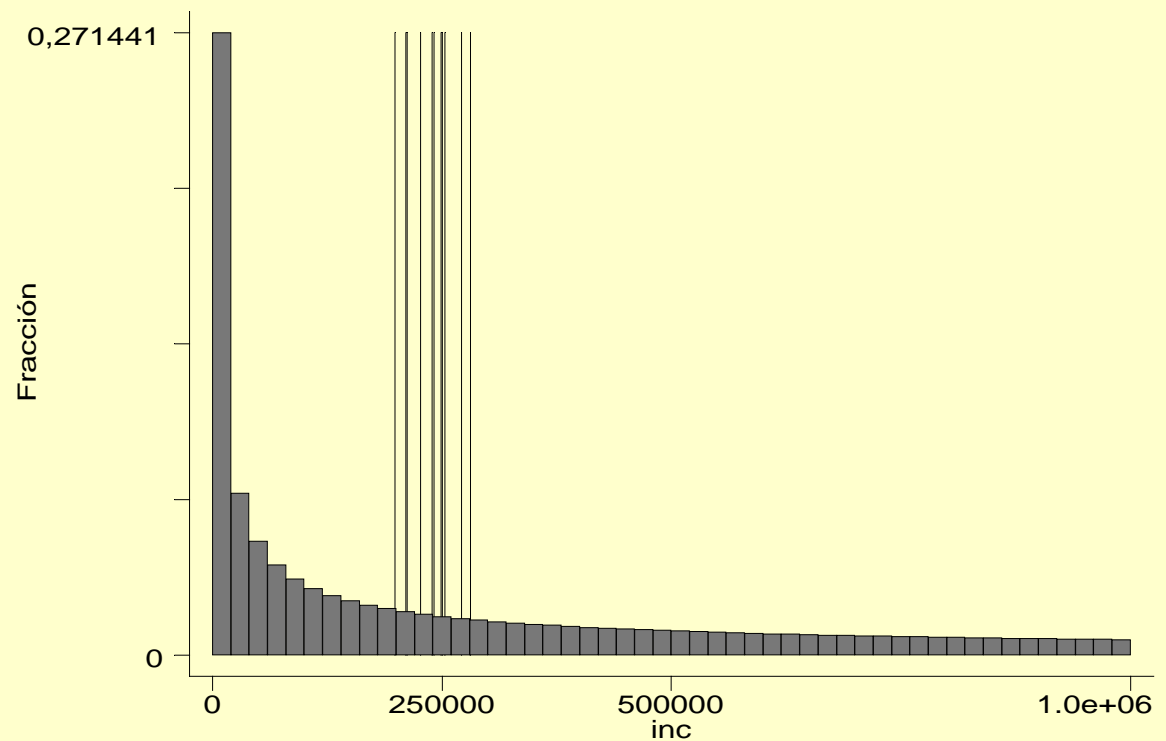
# Razonamiento hacia adelante

# Ejemplo de distribución exponencial



Considere 10 muestras aleatorias  
de  $n = 100$  cada una

Muestra	Media
1	253.396,9
2	198.789,6
3	271.074,2
4	238.928,7
5	280.657,3
6	241.369,8
7	249.036,7
8	226.422,7
9	210.593,4
10	212.137,3



# Considere 10.000 muestras de $n = 100$

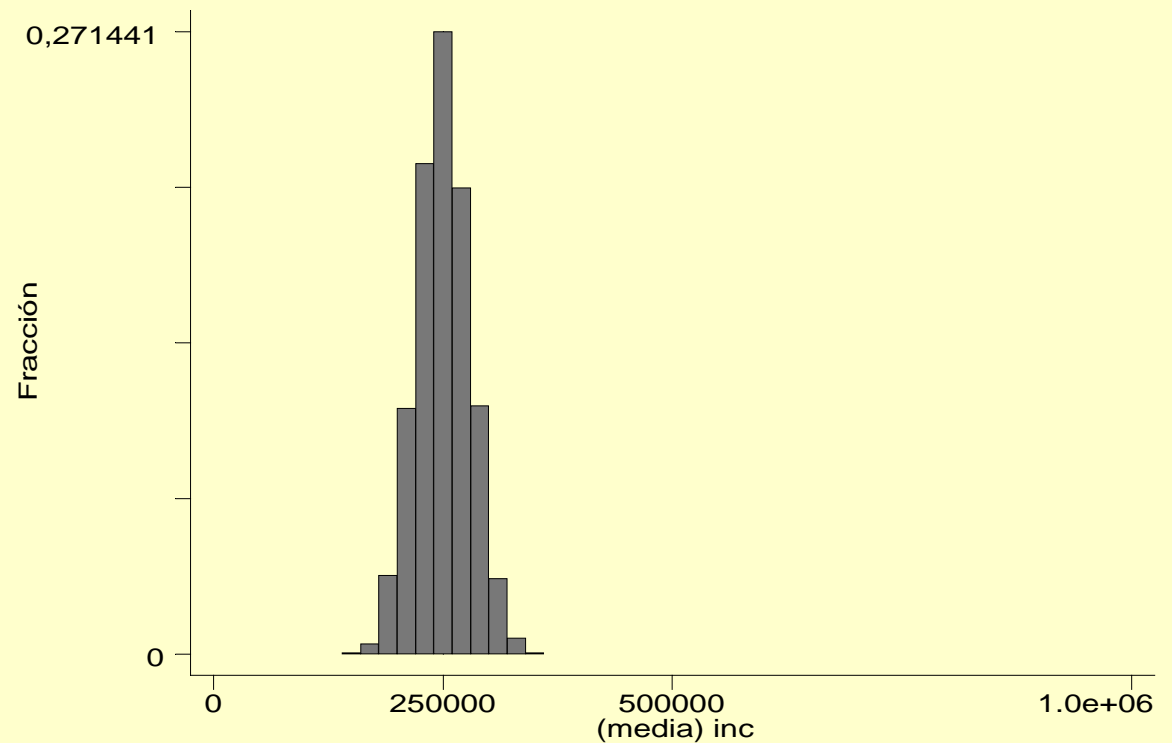
$N = 10.000$

Media = 249.993

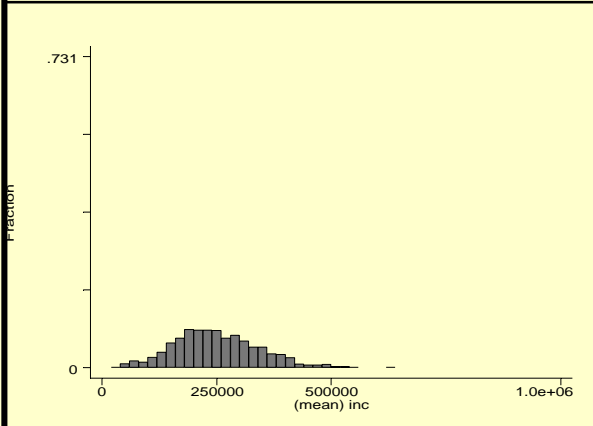
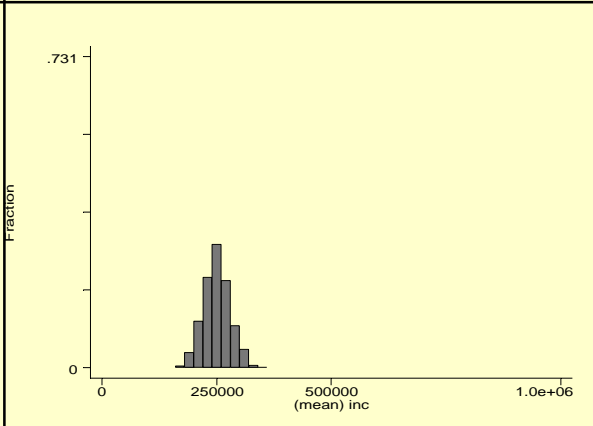
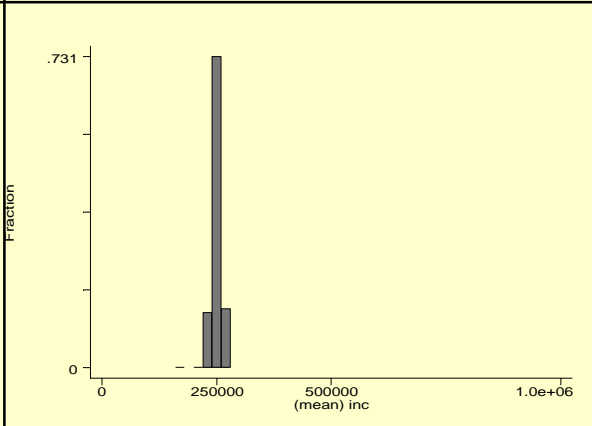
s.d. = 28.559

Asimetría = 0,060

Curtosis = 2,92



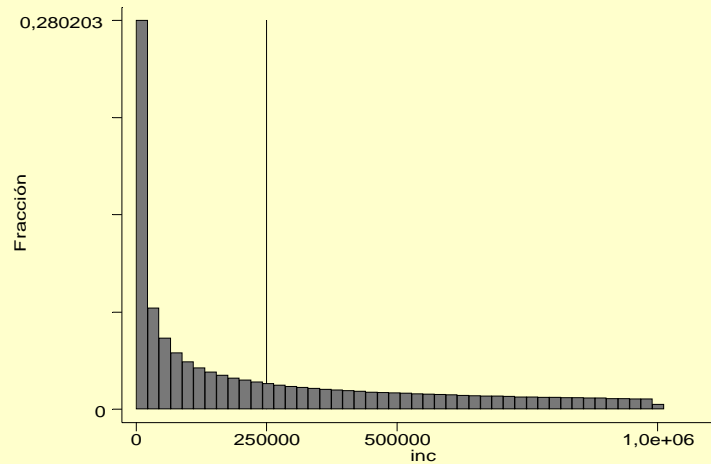
# Considere 1.000 muestras de varios tamaños

10	100	1000
		
<p>Media = 250.105 s.d. = 90.891 Asimetría = 0,38 Curtosis = 3,13</p>	<p>Media = 250.498 s.d. = 28.297 Asimetría = 0,02 Curtosis = 2,90</p>	<p>Media = 249.938 s.d. = 9.376 Asimetría = -0,50 Curtosis = 6,80<sup>17</sup></p>

# Ejemplo de diferencia de medias

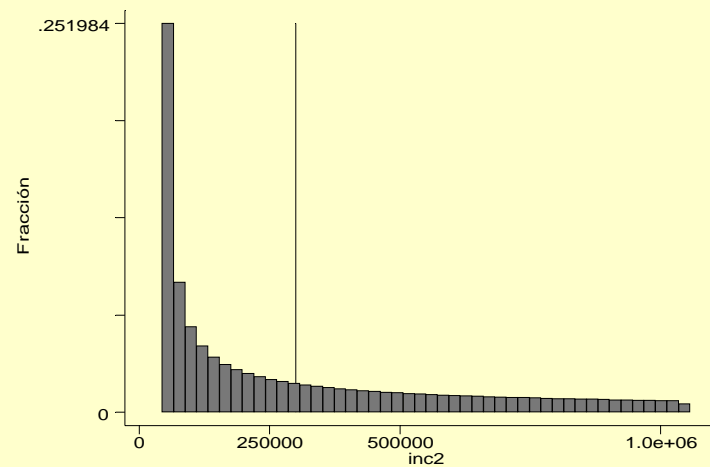
## Estado 1

Media = 250.000



## Estado 2

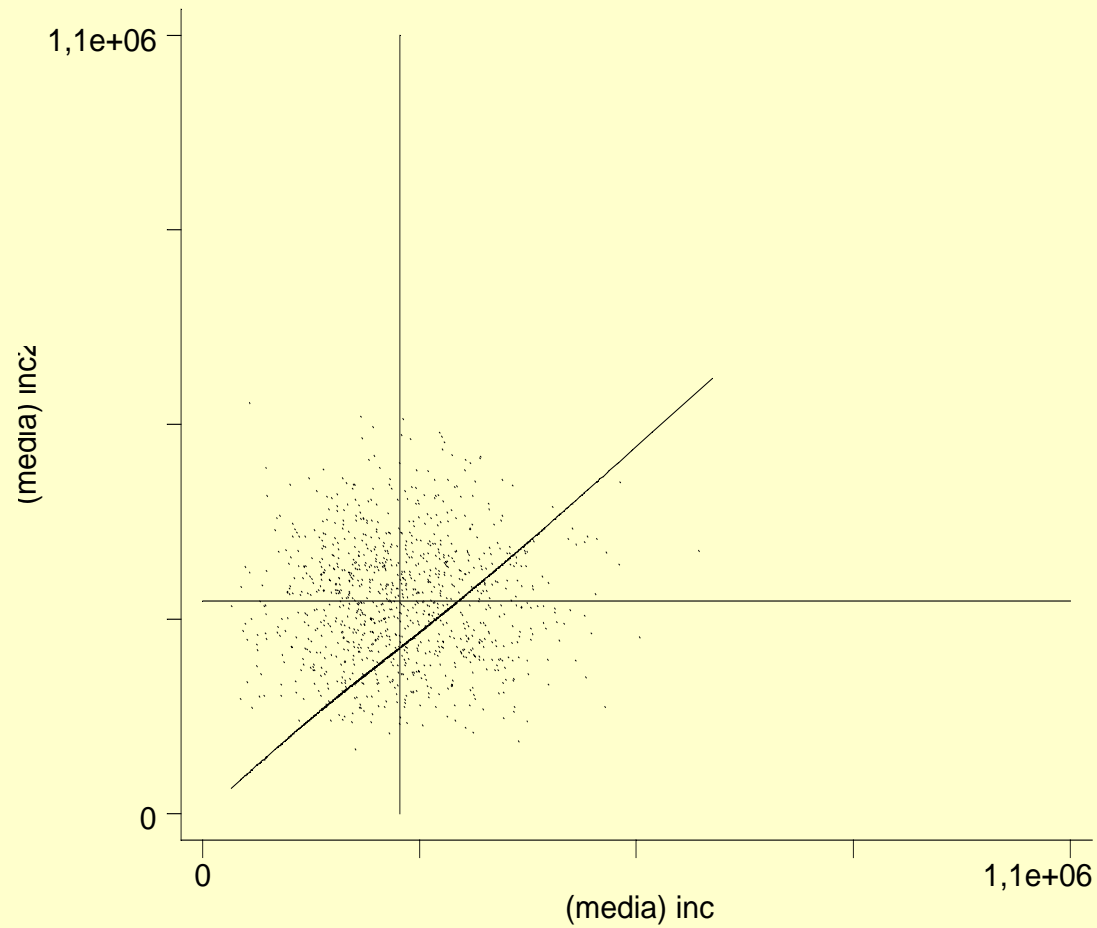
Media = 300.000



Tome 1.000 muestras de 10 de cada estado y compárelas

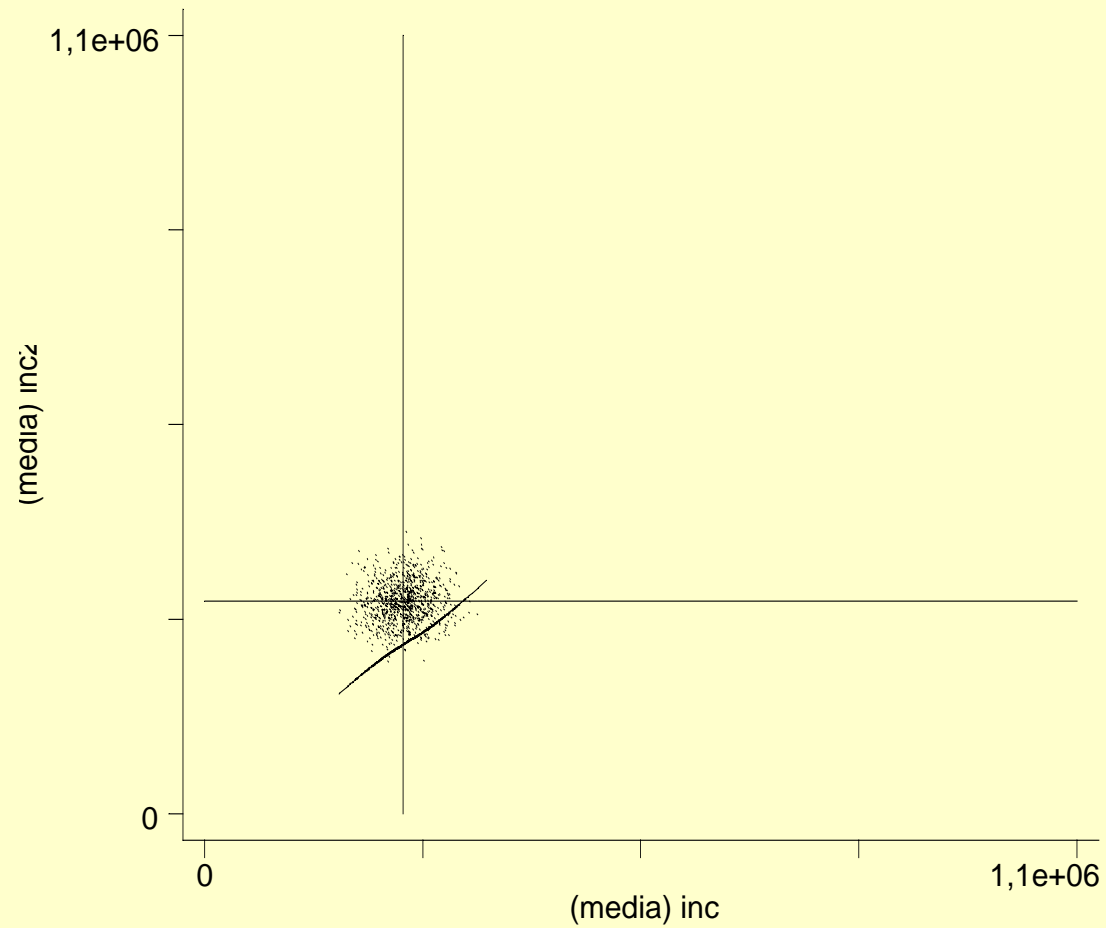
Primeras 10 muestras			
Muestra	Estado 1		Estado 2
1	311.410	<	365.224
2	184.571	<	243.062
3	468.574	>	438.336
4	253.374	<	557.909
5	220.934	>	189.674
6	270.400	<	284.309
7	127.115	<	210.970
8	253.885	<	333.208
9	152.678	<	314.882
10	222.725	>	152.312

# 1.000 muestras de 10



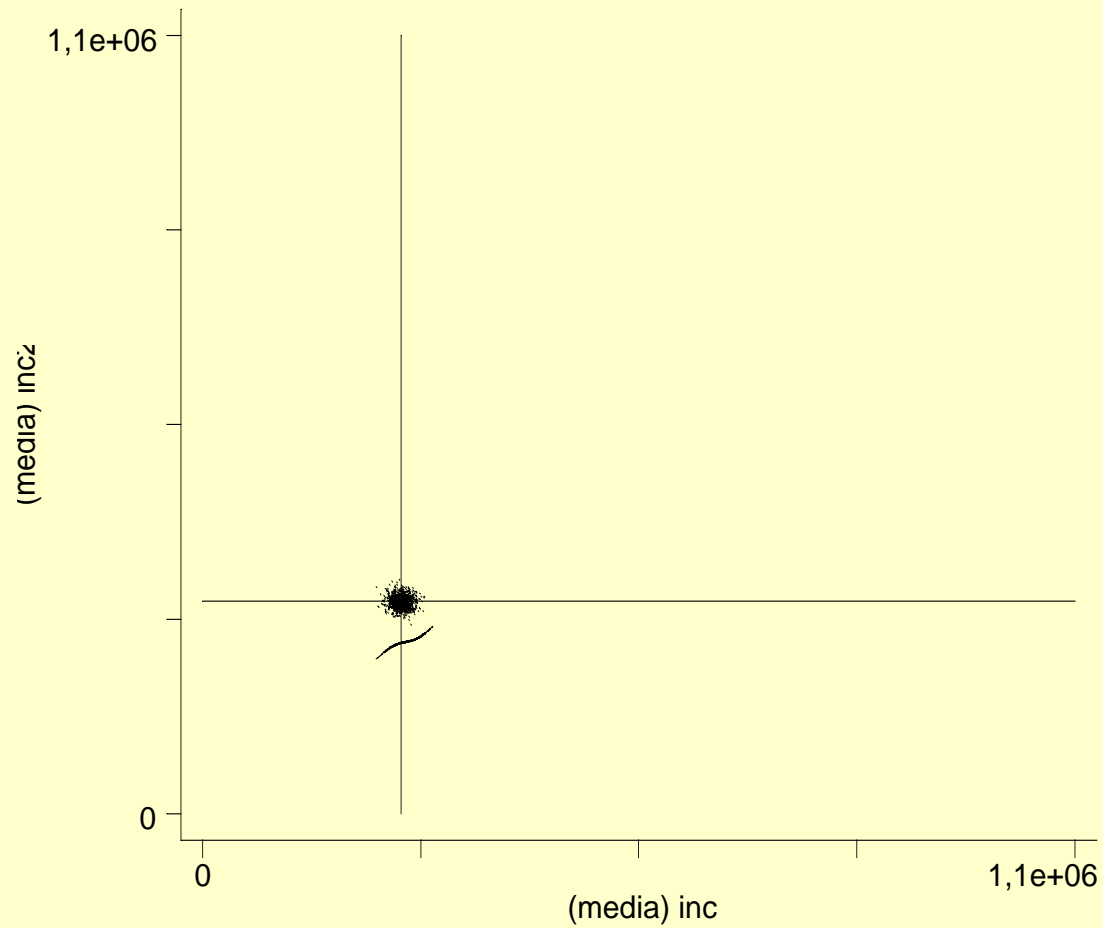
Estado2 > Estado1: 673 veces

# 1.000 muestras de 100



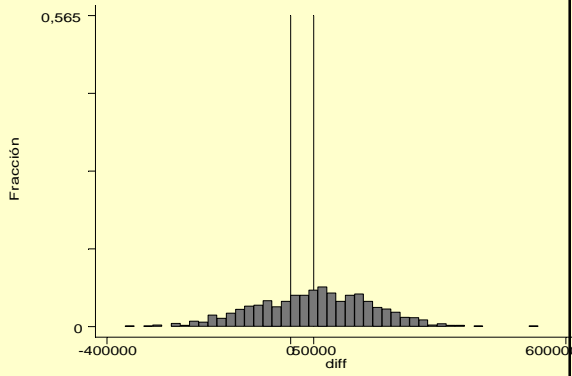
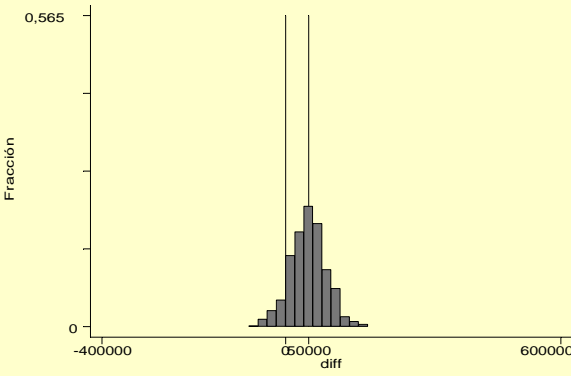
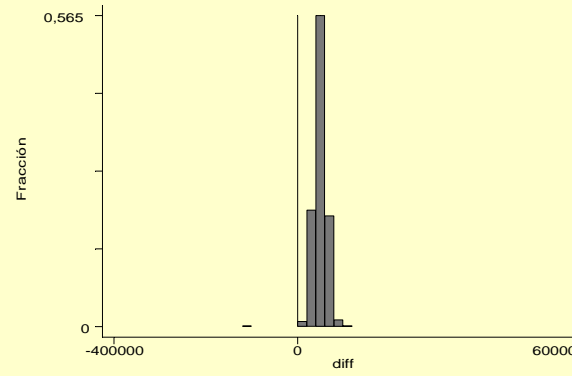
Estado 2 > Estado 1: 909 veces

# 1.000 muestras de 1.000



Estado 2 > Estado 1: 1.000 veces

# Otro modo de ver la cosa: La distribución de $Inc_2 - Inc_1$

$n = 10$	$n = 100$	$n = 1.000$
		
<p>Media = 51.845 s.d. = 124.815</p>	<p>Media = 49.704 s.d. = 38.774</p>	<p>Media = 49.816 s.d. = 13.932</p>

# Razonamiento hacia atrás

ya conoce  $n$ ,  $\bar{X}$ , y  $s$   
pero desea saber algo de  $\mu$

# Teorema del límite central

A medida que se amplía el tamaño de la muestra  $n$ , la distribución de la media  $\bar{X}$  de una muestra aleatoria tomada de **prácticamente cualquier población** se acerca a una distribución *normal*, de media  $\mu$  y desviación estándar de

$$\frac{\sigma}{\sqrt{n}}$$

# Cálculo de errores estándar

En general:

$$\text{s.e.} = \frac{s}{\sqrt{n}}$$

# Errores estándar (*s.e.*) más importantes

Media	$\frac{s}{\sqrt{n}}$
Proporción	$\sqrt{\frac{p(1-p)}{n}}$
Dif. de 2 medias	$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Coef. de regresión (pendiente)	$\frac{s.e.r.}{\sqrt{n}} \times \frac{1}{s_x}$

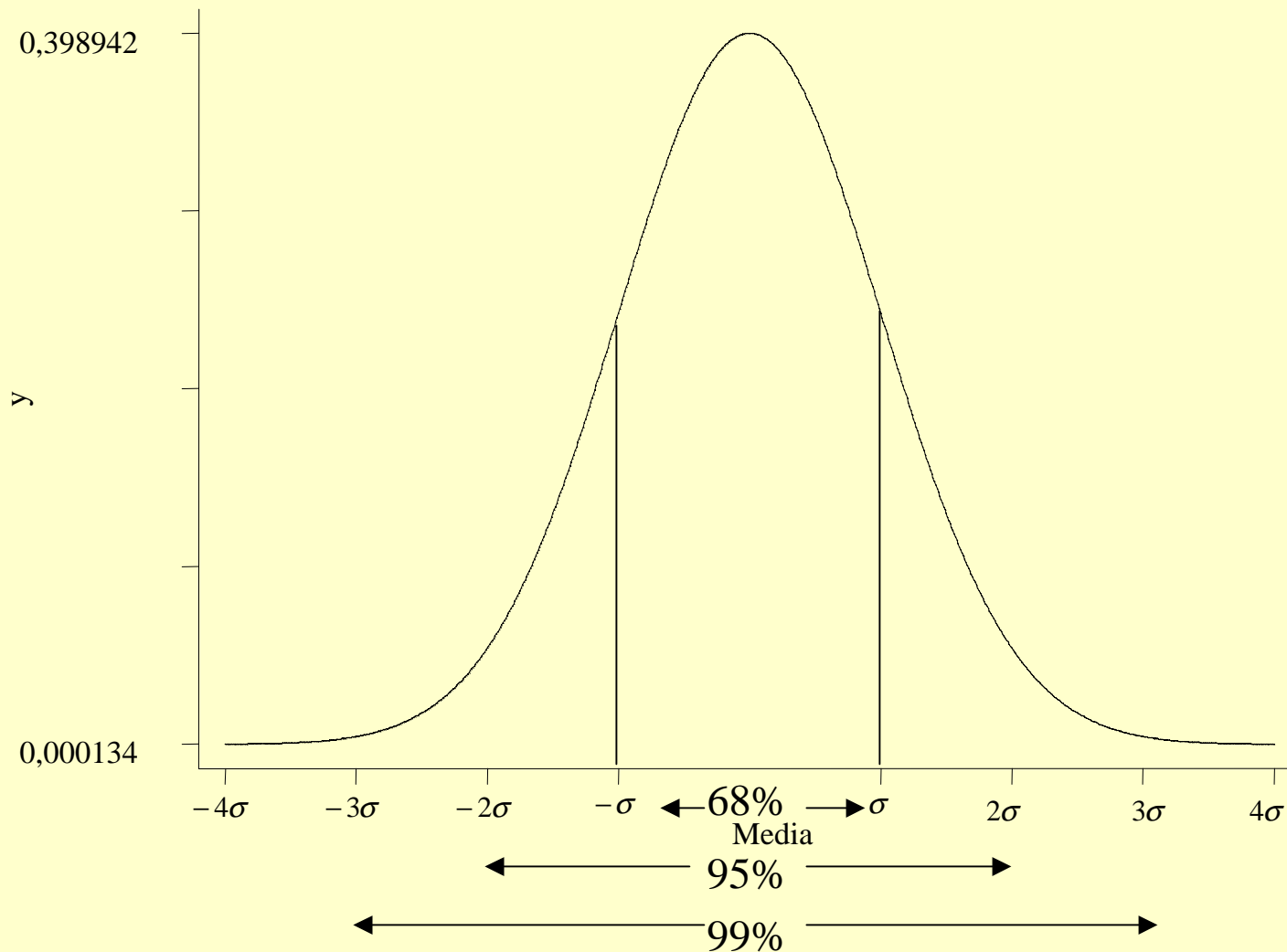
Si conoce la media de la muestra, la desviación estándar y  $n$ , ¿qué puede decir sobre la media de la población?

En general,

media de la población =

media de la muestra  $\pm$  intervalo arbitrario  $\times$  error estándar

Si  $n$  es suficientemente grande, elija el intervalo utilizando la curva normal



# Media de la población utilizando el ejemplo original ( $n = 10$ )

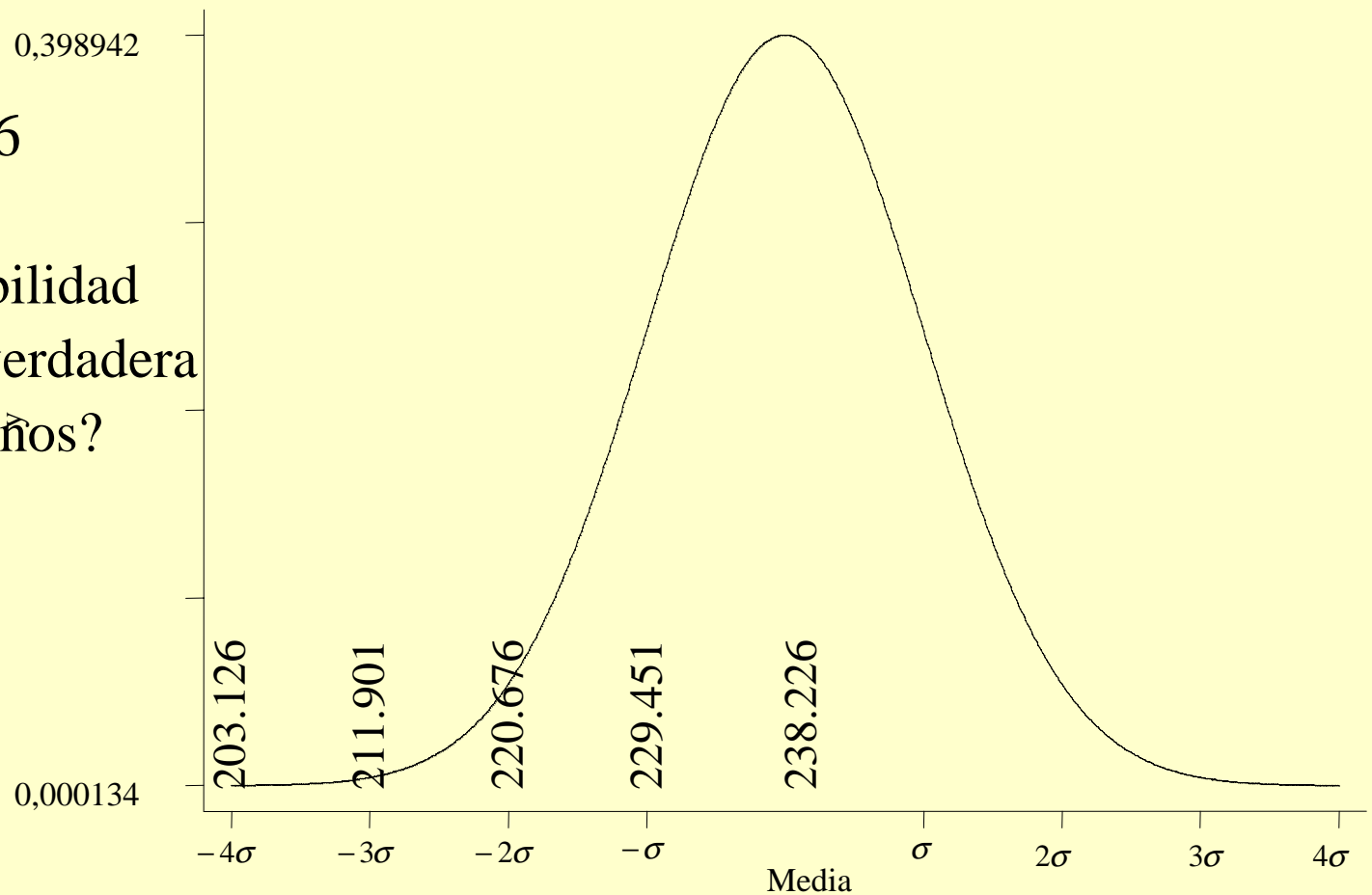
Muestra	Media	s.d.	s.e.	68%		95%		99%	
				inferior	superior	inferior	superior	inferior	superior
1	311.410	241.392	76.335	<b>235.075</b>	<b>387.744</b>	<b>158.740</b>	<b>464.079</b>	<b>82.405</b>	<b>540.414</b>
2	184.571	215.655	68.196	<b>116.375</b>	<b>252.767</b>	<b>48.179</b>	<b>320.963</b>	<b>-20.017</b>	<b>389.159</b>
3	468.574	348.908	110.334	358.240	578.909	<b>247.905</b>	<b>689.243</b>	<b>137.571</b>	<b>799.578</b>
4	253.574	321.599	101.699	<b>151.875</b>	<b>355.272</b>	<b>50.177</b>	<b>456.971</b>	<b>-51.522</b>	<b>558.669</b>
5	220.934	273.256	86.411	<b>134.522</b>	<b>307.345</b>	<b>48.111</b>	<b>393.756</b>	<b>-38.300</b>	<b>480.167</b>
6	270.400	346.008	109.417	<b>160.983</b>	<b>379.817</b>	<b>51.565</b>	<b>489.235</b>	<b>-57.852</b>	<b>598.652</b>
7	127.115	197.071	62.319	64.796	189.435	<b>2.477</b>	<b>251.754</b>	<b>-59.842</b>	<b>314.073</b>
8	253.885	127.711	40.386	<b>213.500</b>	<b>294.271</b>	<b>173.114</b>	<b>334.657</b>	<b>132.728</b>	<b>375.043</b>
9	152.678	201.009	63.564	89.113	216.242	<b>25.549</b>	<b>279.806</b>	<b>-38.016</b>	<b>343.371</b>
10	222.725	264.339	83.591	<b>139.134</b>	<b>306.317</b>	<b>55.543</b>	<b>389.908</b>	<b>-28.048</b>	<b>473.499</b>

# Media de la población utilizando el ejemplo original ( $n = 1000$ )

Muestra	Media	s.d.	s.e.	68%		95%		99%	
				inferior	superior	inferior	superior	inferior	superior
1	238.226	277.492	8.775	229.450	247.001	<b>220.675</b>	<b>255.776</b>	<b>211.900</b>	<b>264.551</b>
2	260.658	290.954	9.201	251.458	269.859	<b>242.257</b>	<b>279.060</b>	<b>233.056</b>	<b>288.261</b>
3	253.374	277.022	8.760	<b>244.614</b>	<b>262.134</b>	<b>235.853</b>	<b>270.894</b>	<b>227.093</b>	<b>279.655</b>
4	242.002	283.772	8.974	<b>233.028</b>	<b>250.975</b>	<b>224.055</b>	<b>259.949</b>	<b>215.081</b>	<b>268.923</b>
5	244.437	279.343	8.834	<b>235.603</b>	<b>253.271</b>	<b>226.770</b>	<b>262.104</b>	<b>217.936</b>	<b>270.938</b>
6	248.896	279.213	8.829	<b>240.067</b>	<b>257.726</b>	<b>231.237</b>	<b>266.555</b>	<b>222.408</b>	<b>275.385</b>
7	267.218	291.150	9.207	258.011	276.425	<b>248.804</b>	<b>285.632</b>	<b>239.597</b>	<b>294.839</b>
8	244.138	276.490	8.743	<b>235.394</b>	<b>252.881</b>	<b>226.651</b>	<b>261.624</b>	<b>217.908</b>	<b>270.368</b>
9	247.996	275.994	8.728	<b>239.268</b>	<b>256.723</b>	<b>230.540</b>	<b>265.451</b>	<b>221.813</b>	<b>274.179</b>
10	255.023	287.118	9.079	<b>245.944</b>	<b>264.103</b>	<b>236.864</b>	<b>273.182</b>	<b>227.785</b>	<b>282.262</b>

# Otra manera de abordar la pregunta: el ratio z

Con  
media = 238.226  
s.e. = 8.775,  
¿cuál es la probabilidad  
de que la media verdadera  
sea 200.000 o menos?



Z

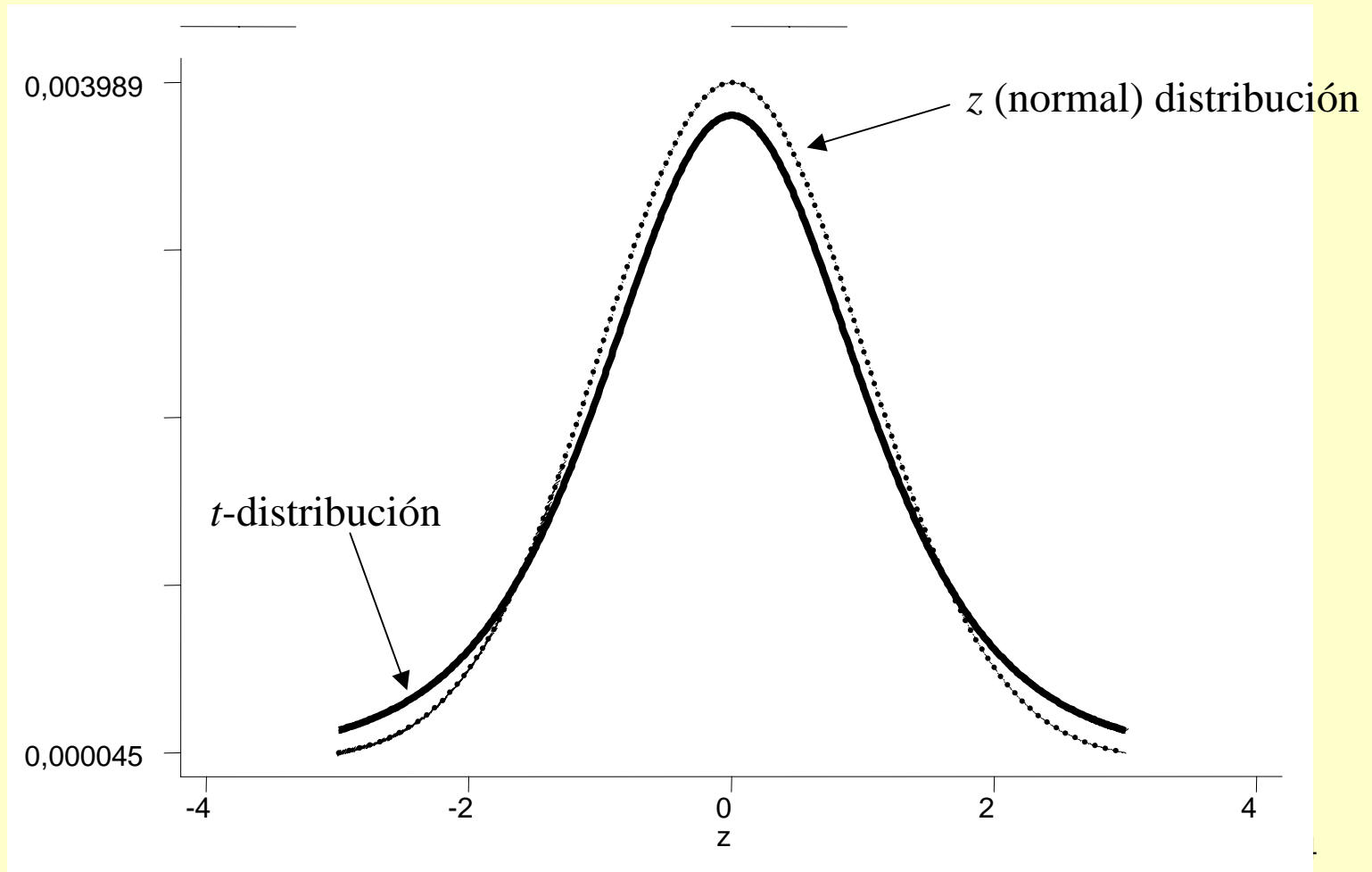
$$z = \frac{(\text{media de muestra-valor test})}{\text{error estándar}},$$

en este caso,

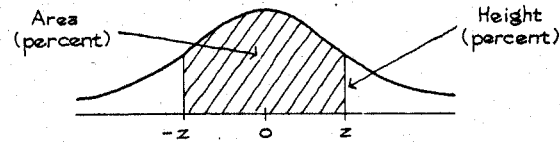
$$z = \frac{(238.226 - 200.000)}{8.775} = 4,37$$

$t$

(cuando la muestra es pequeña)



# Lectura de una tabla z

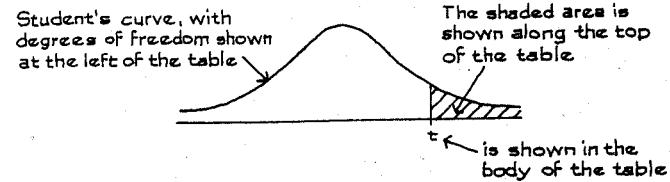


A NORMAL TABLE

<i>z</i>	<i>Height</i>	<i>Area</i>	<i>z</i>	<i>Height</i>	<i>Area</i>	<i>z</i>	<i>Height</i>	<i>Area</i>
0.00	39.89	0	1.50	12.95	86.64	3.00	0.443	99.730
0.05	39.84	3.99	1.55	12.00	87.89	3.05	0.381	99.771
0.10	39.69	7.97	1.60	11.09	89.04	3.10	0.327	99.806
0.15	39.45	11.92	1.65	10.23	90.11	3.15	0.279	99.837
0.20	39.10	15.85	1.70	9.40	91.09	3.20	0.238	99.863
0.25	38.67	19.74	1.75	8.63	91.99	3.25	0.203	99.885
0.30	38.14	23.58	1.80	7.90	92.81	3.30	0.172	99.903
0.35	37.52	27.37	1.85	7.21	93.57	3.35	0.146	99.919
0.40	36.83	31.08	1.90	6.56	94.26	3.40	0.123	99.933
0.45	36.05	34.73	1.95	5.96	94.88	3.45	0.104	99.944
0.50	35.21	38.29	2.00	5.40	95.45	3.50	0.087	99.953
0.55	34.29	41.77	2.05	4.88	95.96	3.55	0.073	99.961
0.60	33.32	45.15	2.10	4.40	96.43	3.60	0.061	99.968
0.65	32.30	48.43	2.15	3.96	96.84	3.65	0.051	99.974
0.70	31.23	51.61	2.20	3.55	97.22	3.70	0.042	99.978
0.75	30.11	54.67	2.25	3.17	97.56	3.75	0.035	99.982
0.80	28.97	57.63	2.30	2.83	97.86	3.80	0.029	99.986
0.85	27.80	60.47	2.35	2.52	98.12	3.85	0.024	99.988
0.90	26.61	63.19	2.40	2.24	98.36	3.90	0.020	99.990
0.95	25.41	65.79	2.45	1.98	98.57	3.95	0.016	99.992
1.00	24.20	68.27	2.50	1.75	98.76	4.00	0.013	99.9937
1.05	22.99	70.63	2.55	1.54	98.92	4.05	0.011	99.9949
1.10	21.79	72.87	2.60	1.36	99.07	4.10	0.009	99.9959
1.15	20.59	74.99	2.65	1.19	99.20	4.15	0.007	99.9967
1.20	19.42	76.99	2.70	1.04	99.31	4.20	0.006	99.9973
1.25	18.26	78.87	2.75	0.91	99.40	4.25	0.005	99.9979
1.30	17.14	80.64	2.80	0.79	99.49	4.30	0.004	99.9983
1.35	16.04	82.30	2.85	0.69	99.56	4.35	0.003	99.9986
1.40	14.97	83.85	2.90	0.60	99.63	4.40	0.002	99.9989
1.45	13.94	85.29	2.95	0.51	99.68	4.45	0.002	99.9991

# Lectura de una tabla $t$

A  $t$ -TABLE



Degrees of freedom	25%	10%	5%	2.5%	1%	0.5%
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17
11	0.70	1.36	1.80	2.20	2.72	3.11
12	0.70	1.36	1.78	2.18	2.68	3.05
13	0.69	1.35	1.77	2.16	2.65	3.01
14	0.69	1.35	1.76	2.14	2.62	2.98
15	0.69	1.34	1.75	2.13	2.60	2.95
16	0.69	1.34	1.75	2.12	2.58	2.92
17	0.69	1.33	1.74	2.11	2.57	2.90
18	0.69	1.33	1.73	2.10	2.55	2.88
19	0.69	1.33	1.73	2.09	2.54	2.86
20	0.69	1.33	1.72	2.09	2.53	2.85
21	0.69	1.32	1.72	2.08	2.52	2.83
22	0.69	1.32	1.72	2.07	2.51	2.82
23	0.69	1.32	1.71	2.07	2.50	2.81
24	0.68	1.32	1.71	2.06	2.49	2.80
25	0.68	1.32	1.71	2.06	2.49	2.79

# Hacer una prueba $t$

P: ¿Cuál es la posibilidad de que la tasa de voto residual en 1996 fuese 2,5% o menos?

Media: 0,02618

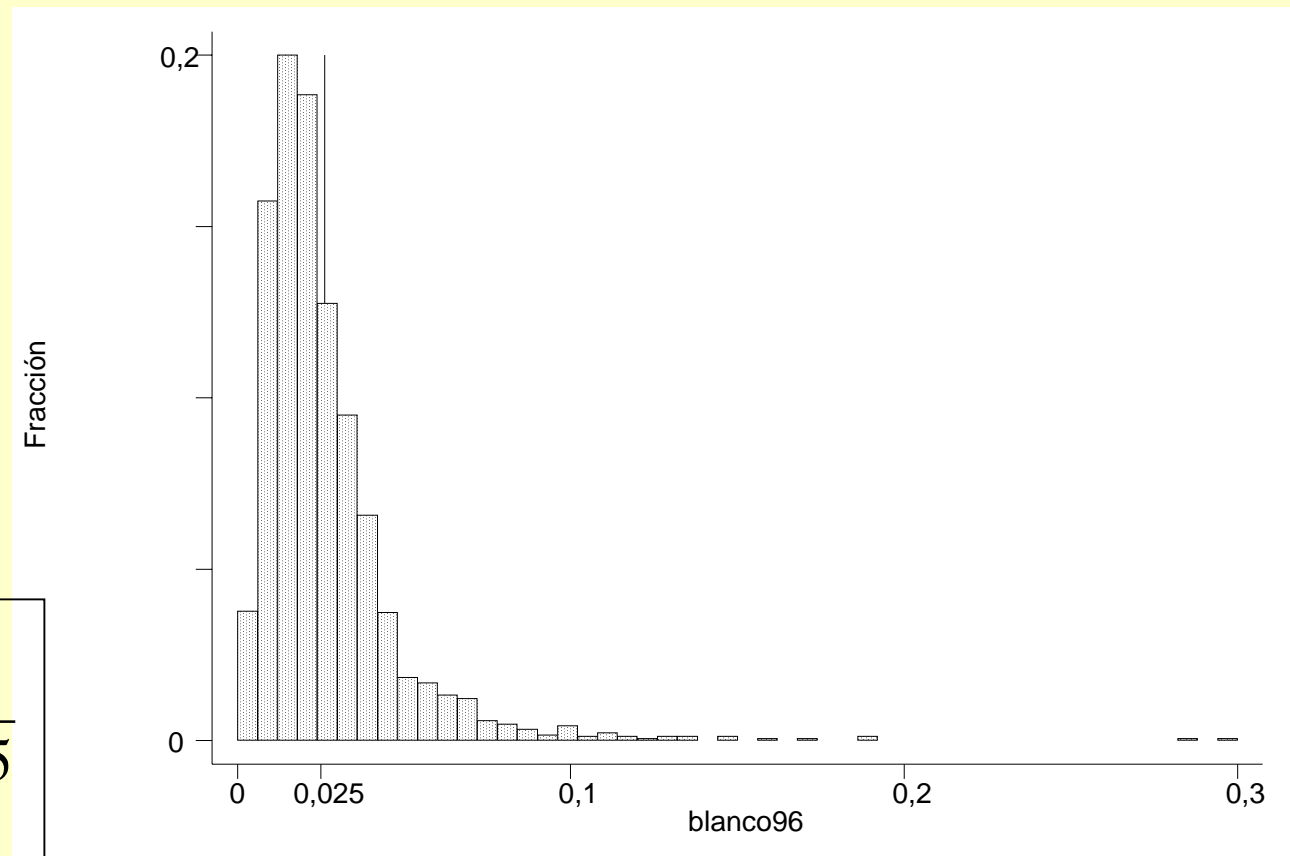
s.d.: 0,02140

N: 1905

$$s.e. = s / \sqrt{n}$$

$$= 0,02140 / \sqrt{1905}$$

$$= 0,00049$$



## El cuadro

Media: 0,02618

s.d.: 0,02140

N: 1905

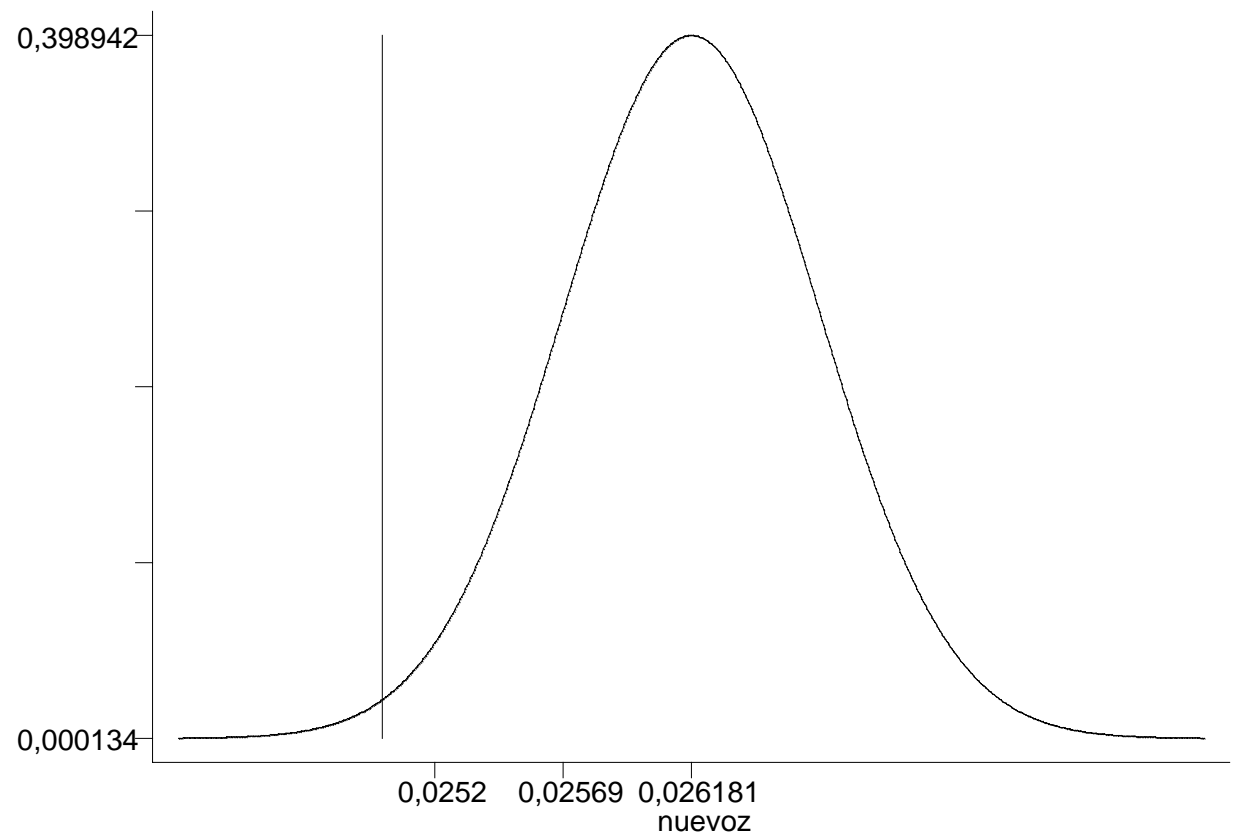
$$s.e. = s / \sqrt{n}$$

$$= 0,02140 / \sqrt{1905} >$$

$$= 0,00049$$

$$t = \frac{0,026181 - 0,025}{0,00049}$$

$$= 2,408$$



# El resultado de *STATA*

```
. ttest blank96=.025
```

One-sample t test

```
-----  
Variable |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]  
-----+-----  
blank96 |    1905   .0261806   .0004903   .0213979   .0252191   .0271421  
-----
```

Degrees of freedom: 1904

Ho: mean(blank96) = .025

Ha: mean < .025

t = 2.4082

P < t = 0.9919

Ha: mean ~= .025

t = 2.4082

P > |t| = 0.0161

Ha: mean > .025

t = 2.4082

P > t = 0.0081

## Otra prueba $t$

P: ¿Cuál es la posibilidad de que la tasa de voto residual en 1996 fuese igual a la de 1992 (esto es,  $\text{blank}_{96} - \text{blank}_{92} = 0$ )?

Media: 0,003069

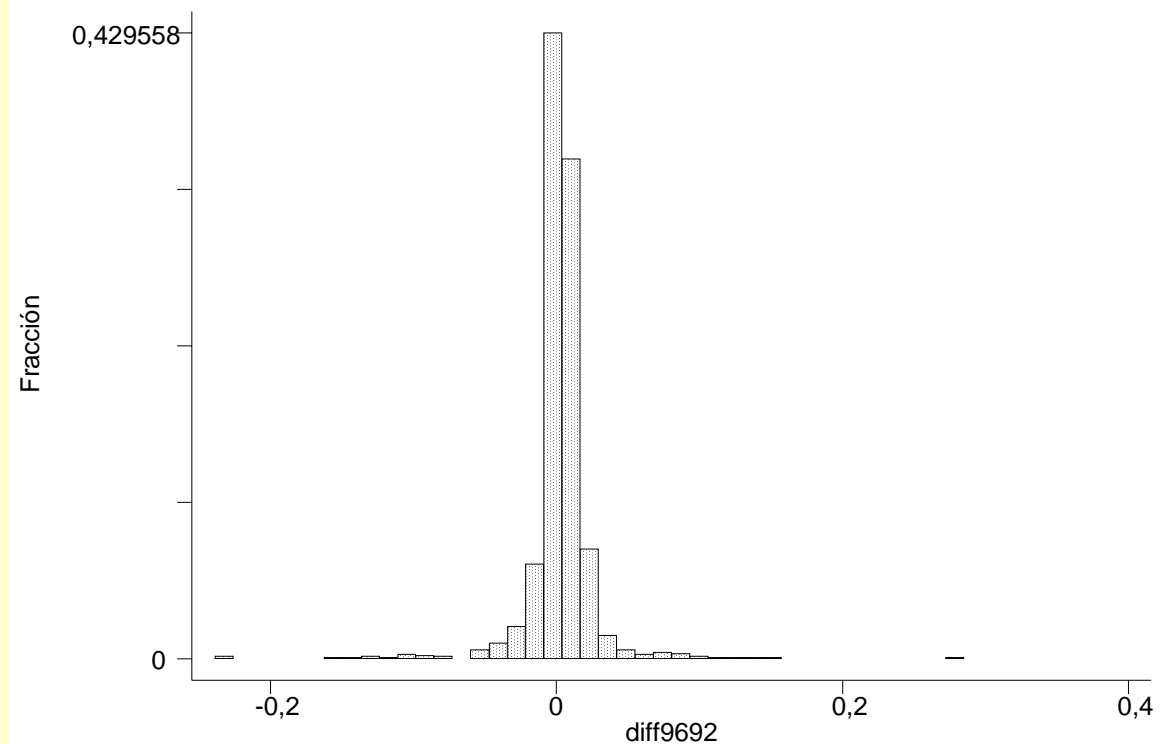
s.d.: 0,02323

N: 1448

$$s.e. = s / \sqrt{n}$$

$$= 0,02323 / \sqrt{1448}$$

$$= 0,00061$$



## El cuadro

Media: 0,003069

s.d.: 0,02323

N: 1448

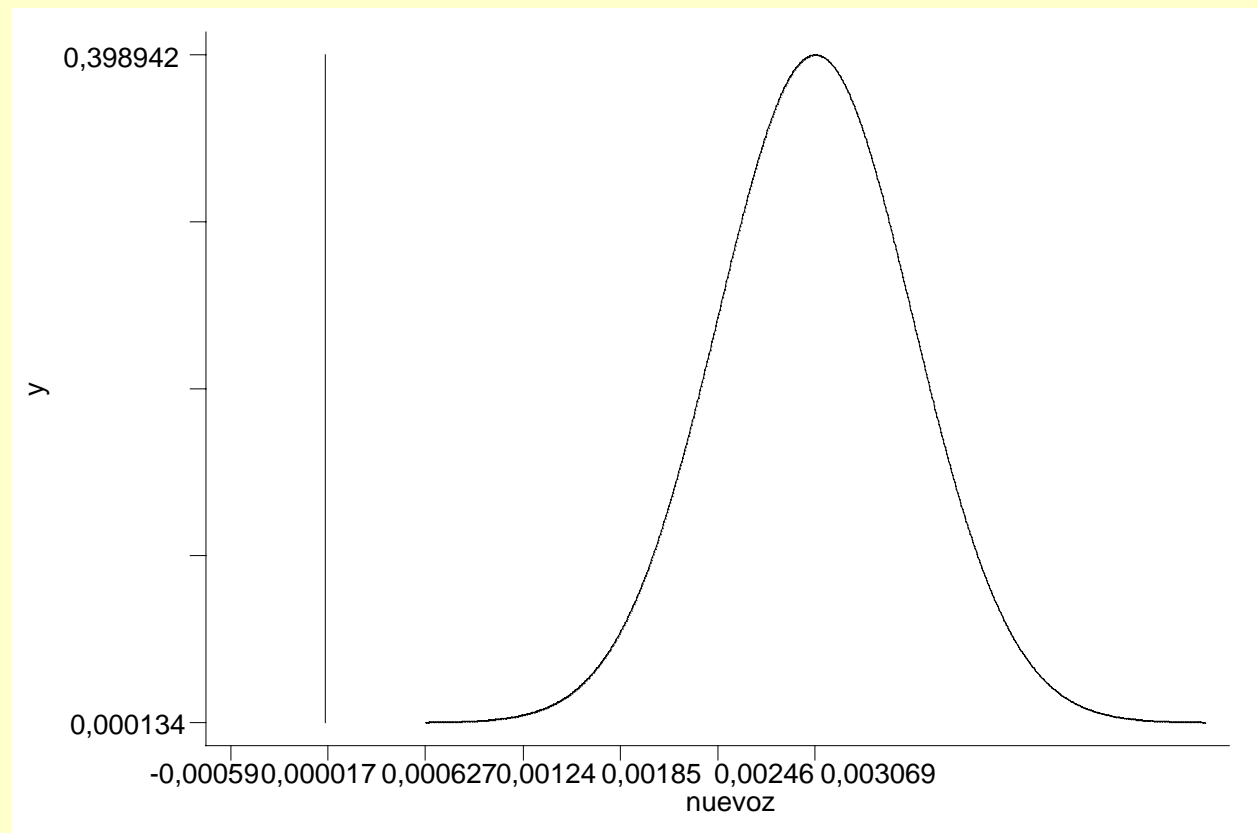
$$s.e. = s / \sqrt{n}$$

$$= 0,02323 / \sqrt{1448}$$

$$= 0,00061$$

$$t = \frac{0,003069 - 0}{0,00061}$$

$$= 5,028$$



# El resultado de *STATA*

```
. ttest blank96=blank92
```

```
Paired t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
blank96	1448	.0242941	.0005116	.0194689	.0232904	.0252977
blank92	1448	.021225	.0005382	.0204813	.0201692	.0222808
diff	1448	.003069	.0006104	.0232279	.0018717	.0042664

```
Ho: mean(blank96 - blank92) = mean(diff) = 0
```

```
Ha: mean(diff) < 0
```

```
t = 5.0278  
P < t = 1.0000
```

```
Ha: mean(diff) ~= 0
```

```
t = 5.0278  
P > |t| = 0.0000
```

```
Ha: mean(diff) > 0
```

```
t = 5.0278  
P > t = 0.0000
```

```
. ttest diff9692=0
```

```
One-sample t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
diff9692	1448	.003069	.0006104	.0232279	.0018717	.0042664

```
Degrees of freedom: 1447
```

```
Ho: mean(diff9692) = 0
```

```
Ha: mean < 0
```

```
t = 5.0278  
P < t = 1.0000
```

```
Ha: mean ~= 0
```

```
t = 5.0278  
P > |t| = 0.0000
```

```
Ha: mean > 0
```

```
t = 5.0278  
P > t = 0.0000
```

# Prueba $t$ final

P: ¿Había alguna relación entre el voto residual y el tamaño del condado en 1996?

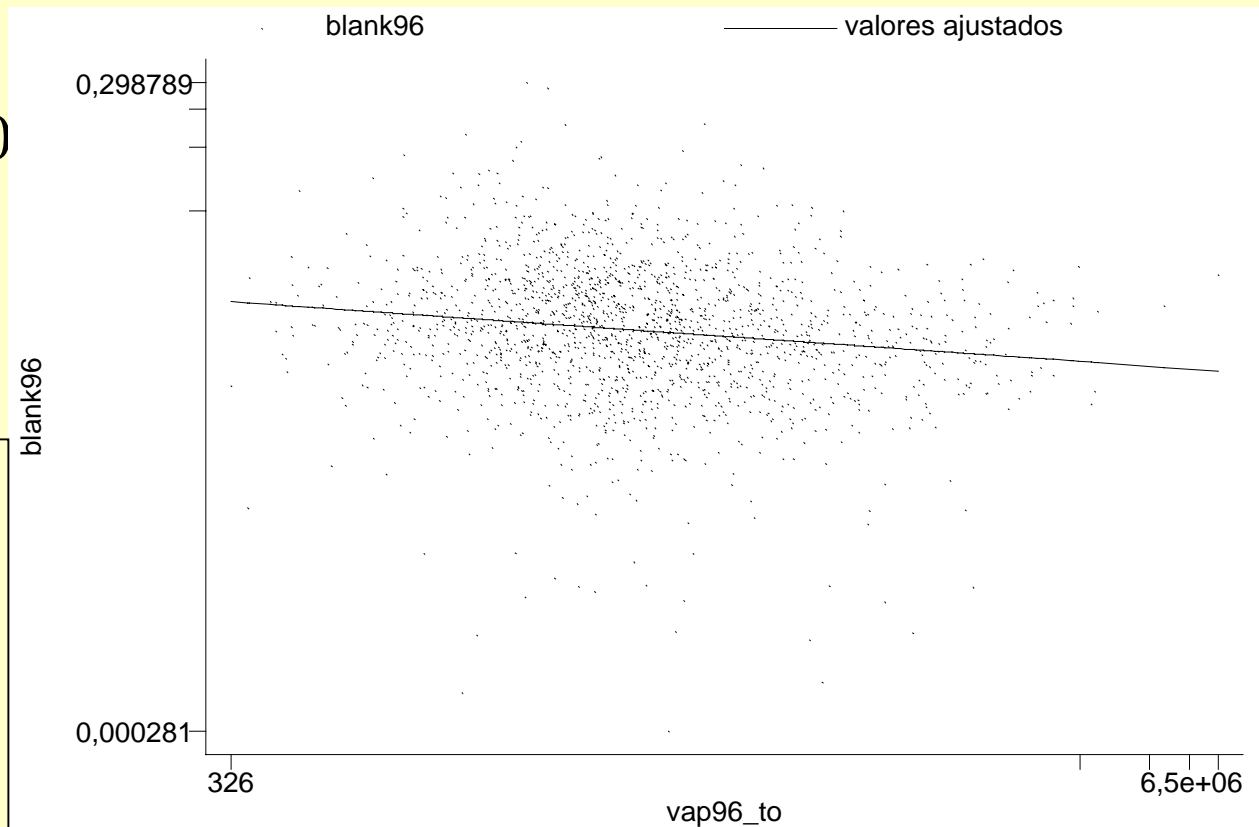
Coef. pend.: -0,07510

s.e.r: 0,7115

N: 1861

$S_x$ : 1,4788

$$\begin{aligned} s.e. &= \frac{s.e.r}{\sqrt{n}} \times \frac{1}{s_x} \\ &= \frac{0,7115}{\sqrt{1861}} \times \frac{1}{1,4788} \\ &= 0,01649 \times 0,6762 \\ &= 0,01115 \end{aligned}$$



## Cálculo de $t$

$$t = \frac{-0,07510}{0,01115}$$
$$= -6,7319$$

# El resultado de *STATA*

```
. reg lblank96 lvap96
```

Source	SS	df	MS	Number of obs = 1861		
Model	22.941515	1	22.941515	F( 1, 1859)	=	45.32
Residual	941.080329	1859	.506229332	Prob > F	=	0.0000
Total	964.021844	1860	.518291314	R-squared	=	0.0238
				Adj R-squared	=	0.0233
				Root MSE	=	.7115

lblank96	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lvap96	-.0750985	.0111556	-6.73	0.000	-.0969774	-.0532197
_cons	-3.129858	.1113781	-28.10	0.000	-3.348298	-2.911419