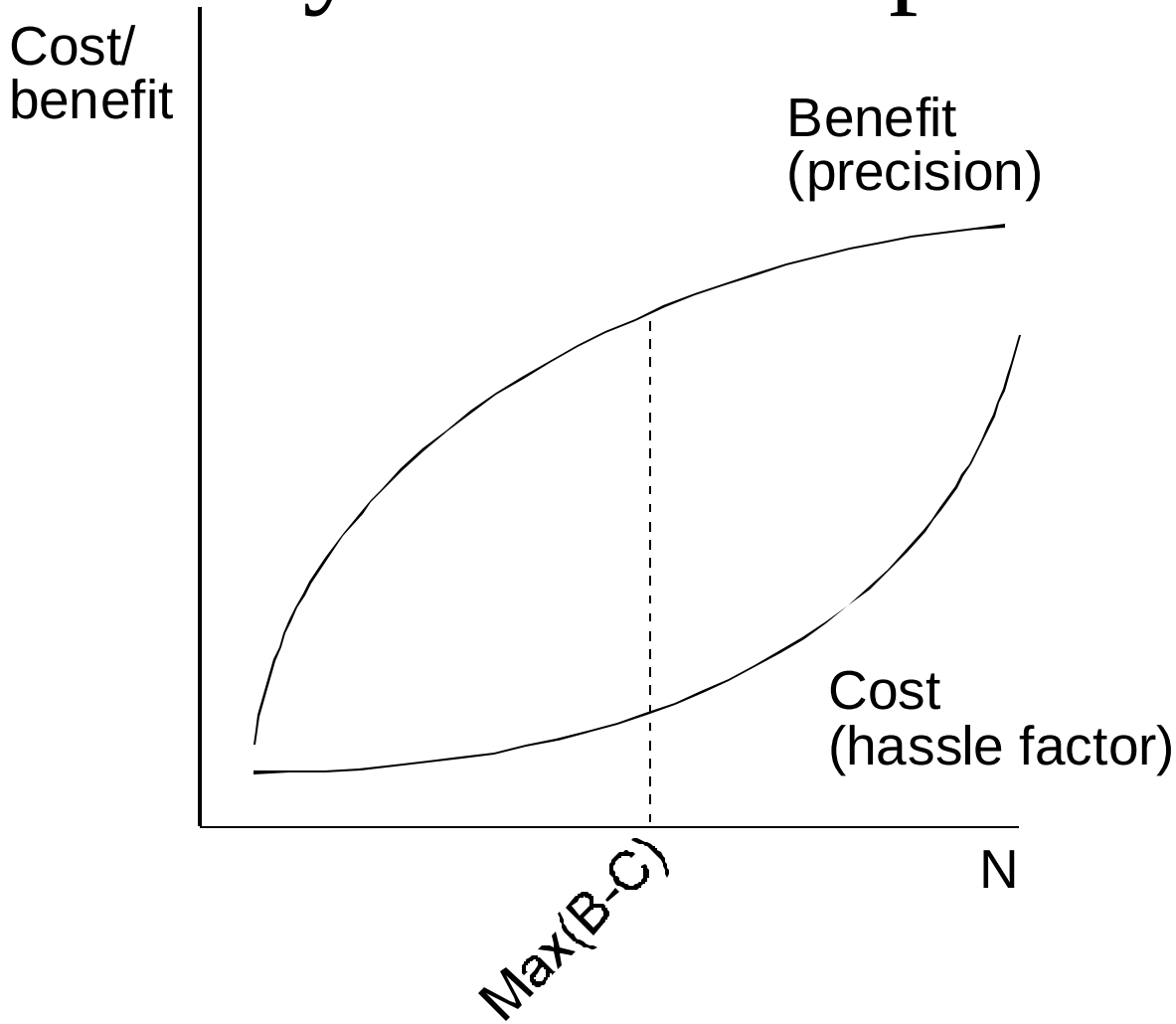
Sampling and Inference

The Quality of Data and Measures

Why we talk about sampling

- General citizen education
- Understand data you'll be using
- Understand how to draw a sample, if you need to
- Make statistical inferences

Why do we sample?



How do we sample?

- Simple random sample
  - Variant: systematic sample with a random start
- Stratified
- Cluster

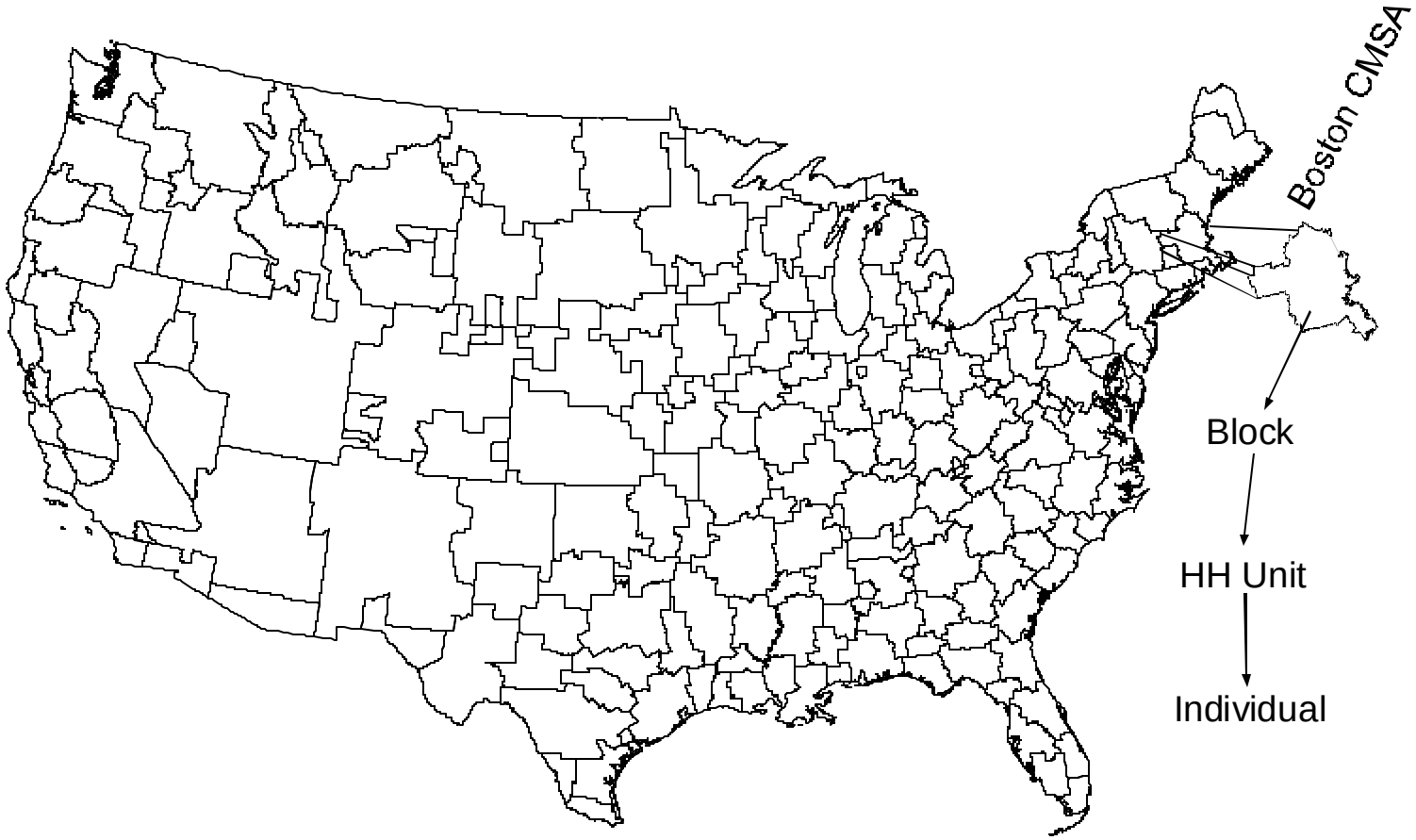
Stratification

- Divide sample into subsamples, based on *known* characteristics (race, sex, religiousity, continent, department)
- Benefit: preserve or enhance variability

Stratification example

	NES		Hypothetical sample	
	N	s.e. @ 50%	N	s.e. @ 50%
White Christians	1,215	0.7%	350	1.3%
Black Christians	187	1.8%	350	1.3%
White Jews	30	4.6%	350	1.3%
Black Jews	2	17.7%	350	1.3%
Other race/religion	53	3.4%	87	2.7%
Missing	227	n.a.		
Total	1,714	0.6% (on 1,487 valid obs.)		

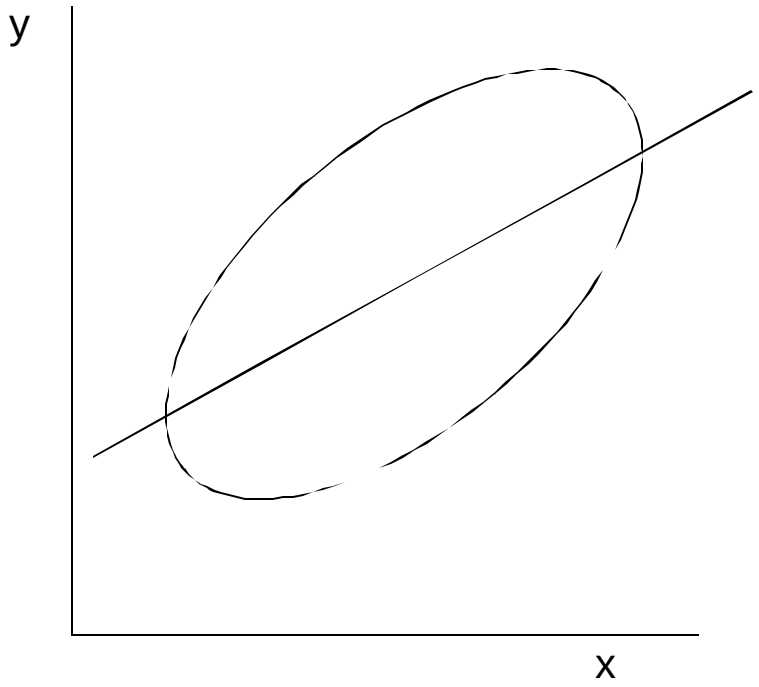
Cluster sampling

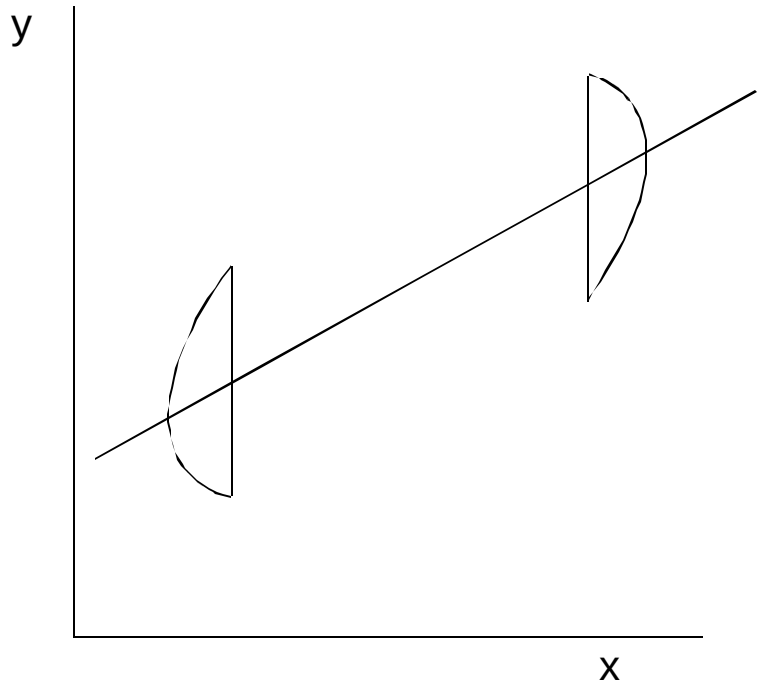


Effects of samples

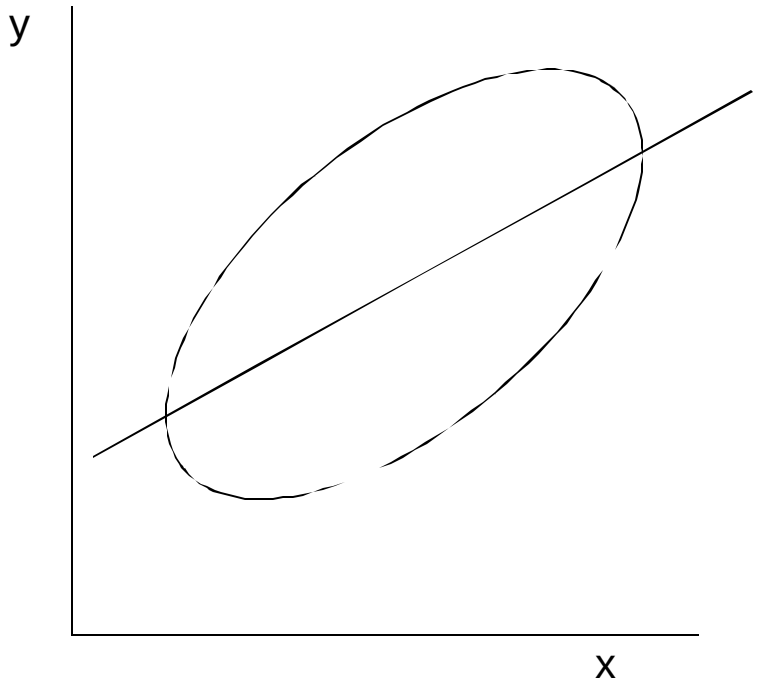
- Obvious: influences marginals
- Less obvious
  - Allows effective use of time and effort
  - Effect on multivariate techniques
    - Sampling of independent variable: greater precision in regression estimates
    - Sampling on dependent variable: bias

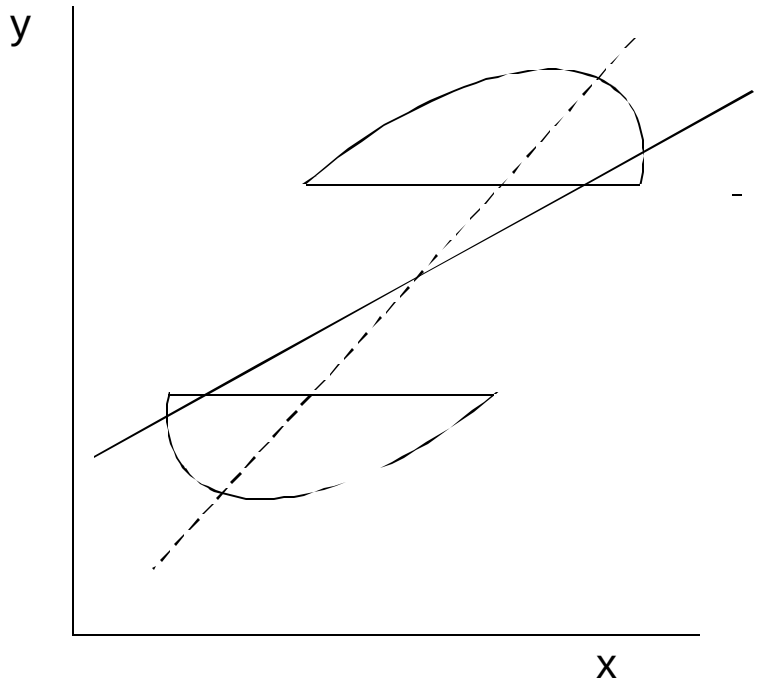
Sampling on Independent Variable





Sampling on Dependent Variable





Sampling

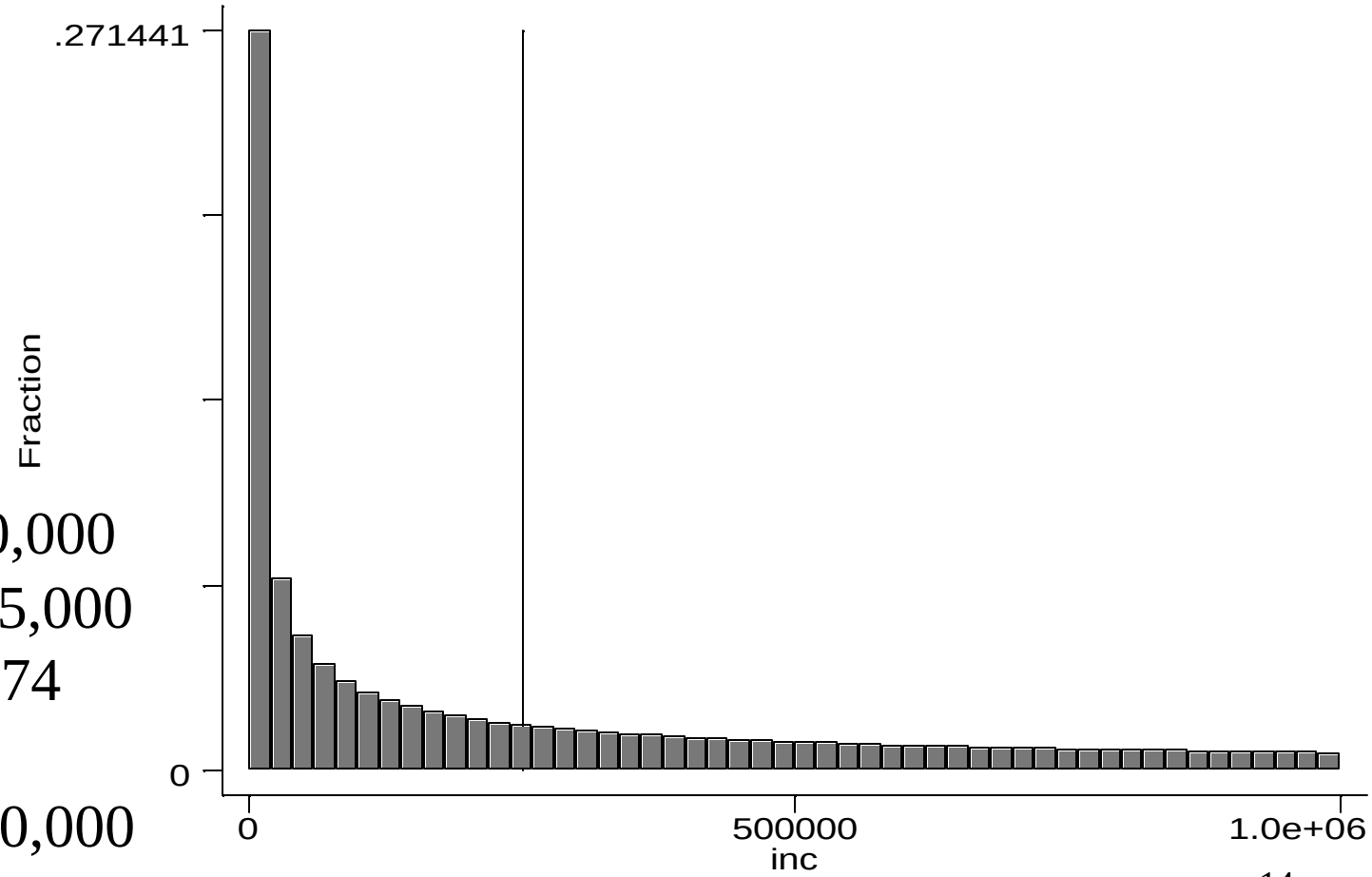
Consequences for Statistical Inference

Statistical Inference: Learning About the Unknown From the Known

- Reasoning forward: distributions of sample means, when the population mean, s.d., and *n* are known.
- Reasoning backward: learning about the population mean when only the sample, s.d., and *n* are known

Reasoning Forward

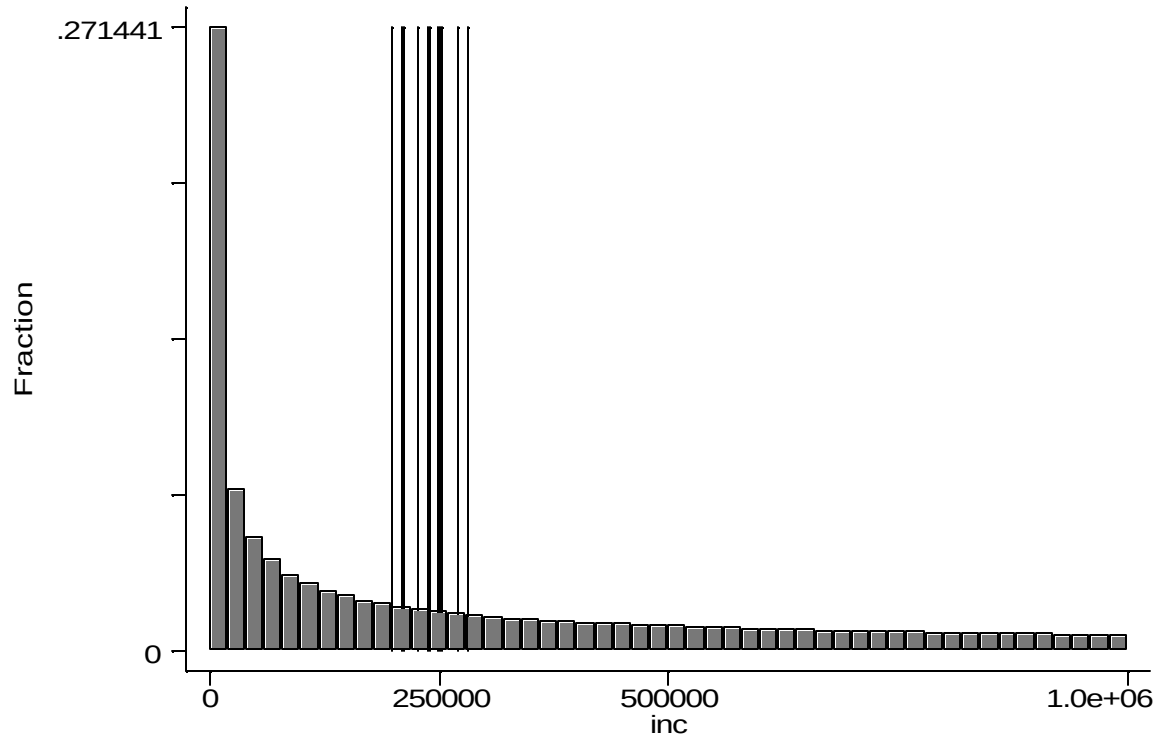
Exponential Distribution Example



Mean = 250,000
Median=125,000
s.d. = 283,474
Min = 0
Max = 1,000,000

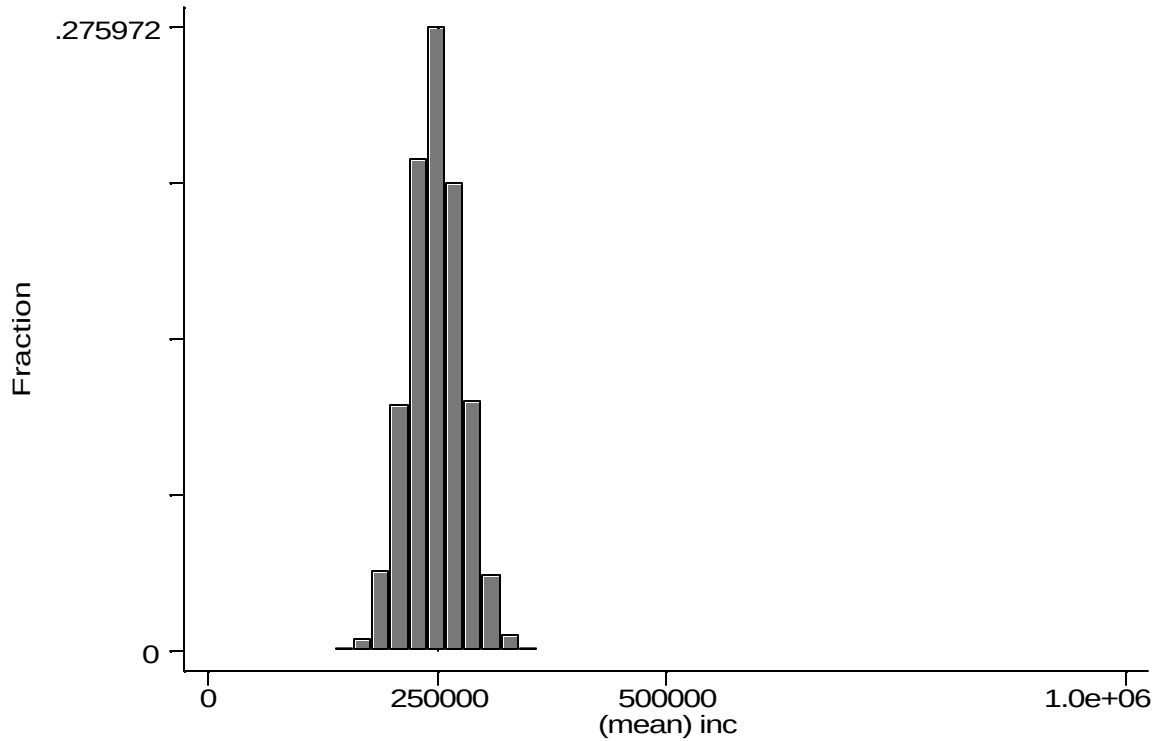
Consider 10 random samples, of n = 100 apiece

Sample	mean
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

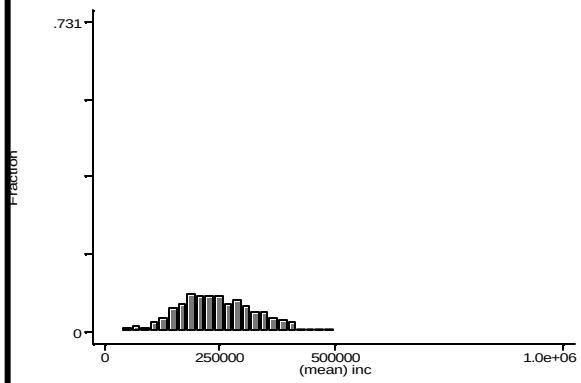
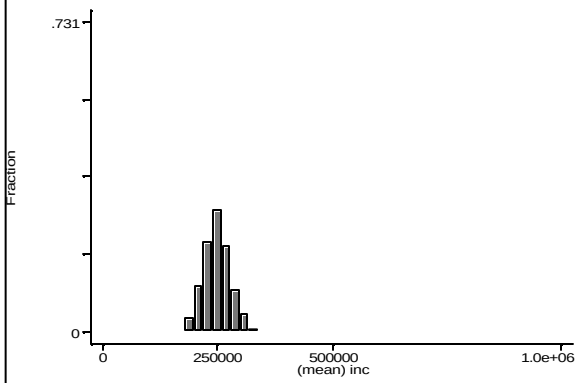
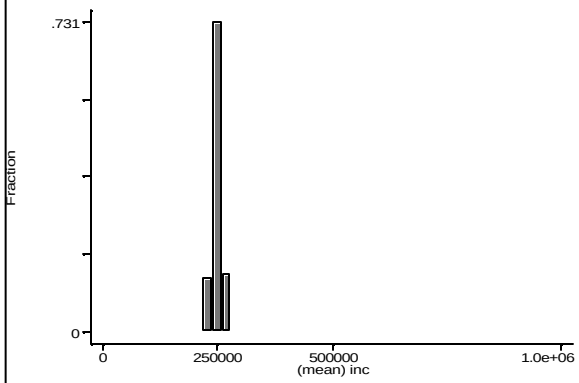


Consider 10,000 samples of n = 100

N = 10,000
Mean = 249,993
s.d. = 28,559
Skewness = 0.060
Kurtosis = 2.92



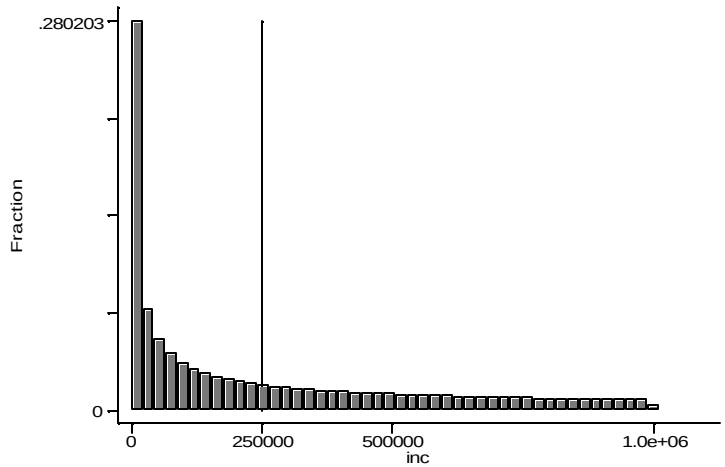
Consider 1,000 samples of various sizes

10	100	1000
		
Mean =250,105 s.d.= 90,891 Skew= 0.38 Kurt= 3.13	Mean = 250,498 s.d.= 28,297 Skew= 0.02 Kurt= 2.90	Mean = 249,938 s.d.= 9,376 Skew= -0.50 Kurt= 6.80

Difference of means example

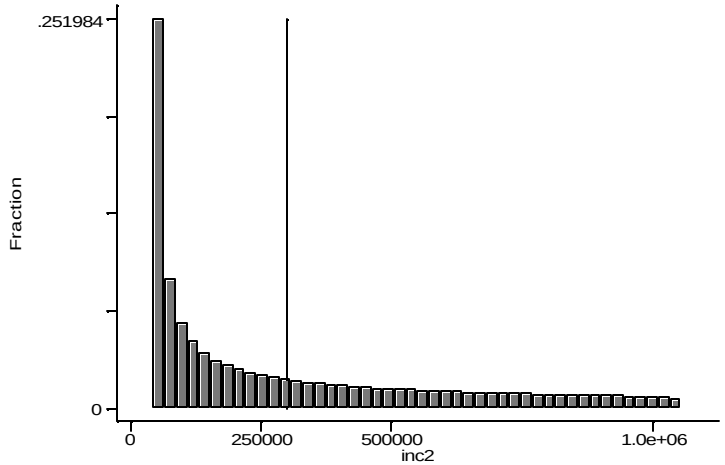
State 1

Mean = 250,000



State 2

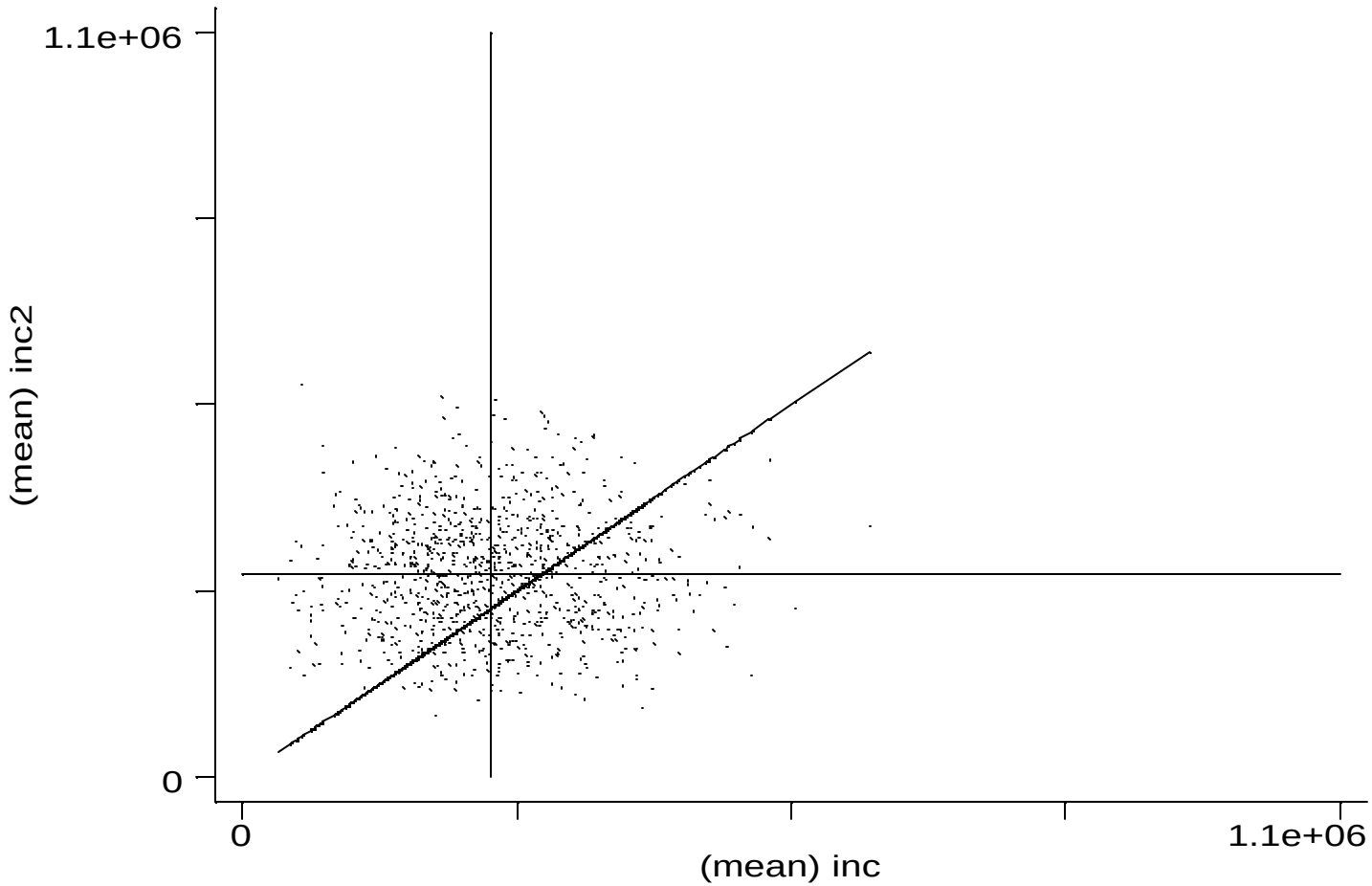
Mean = 300,000



Take 1,000 samples of 10, of each state, and compare them

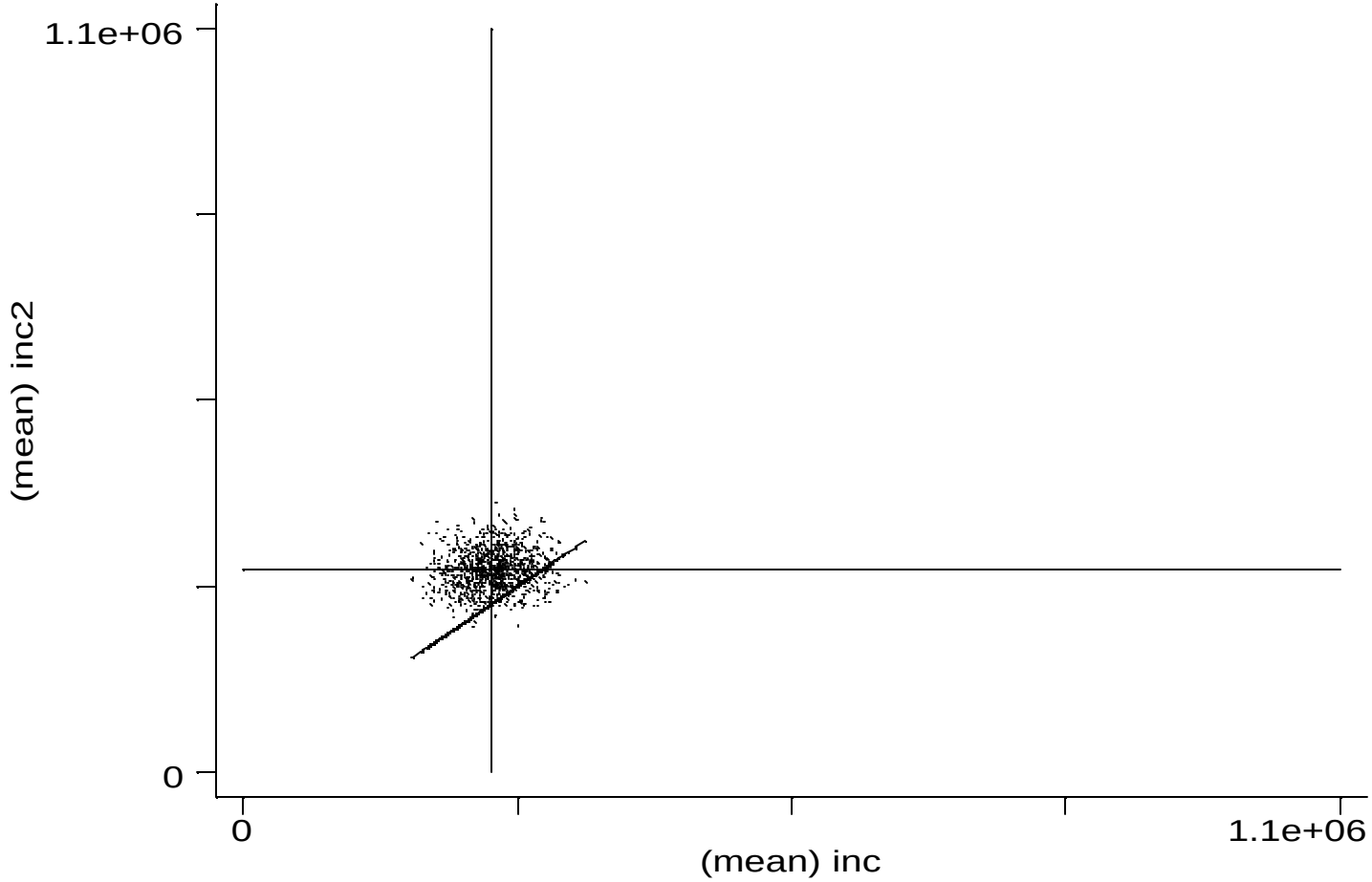
First 10 samples			
Sample	State 1		State 2
1	311,410	<	365,224
2	184,571	<	243,062
3	468,574	>	438,336
4	253,374	<	557,909
5	220,934	>	189,674
6	270,400	<	284,309
7	127,115	<	210,970
8	253,885	<	333,208
9	152,678	<	314,882
10	222,725	>	152,312

1,000 samples of 10



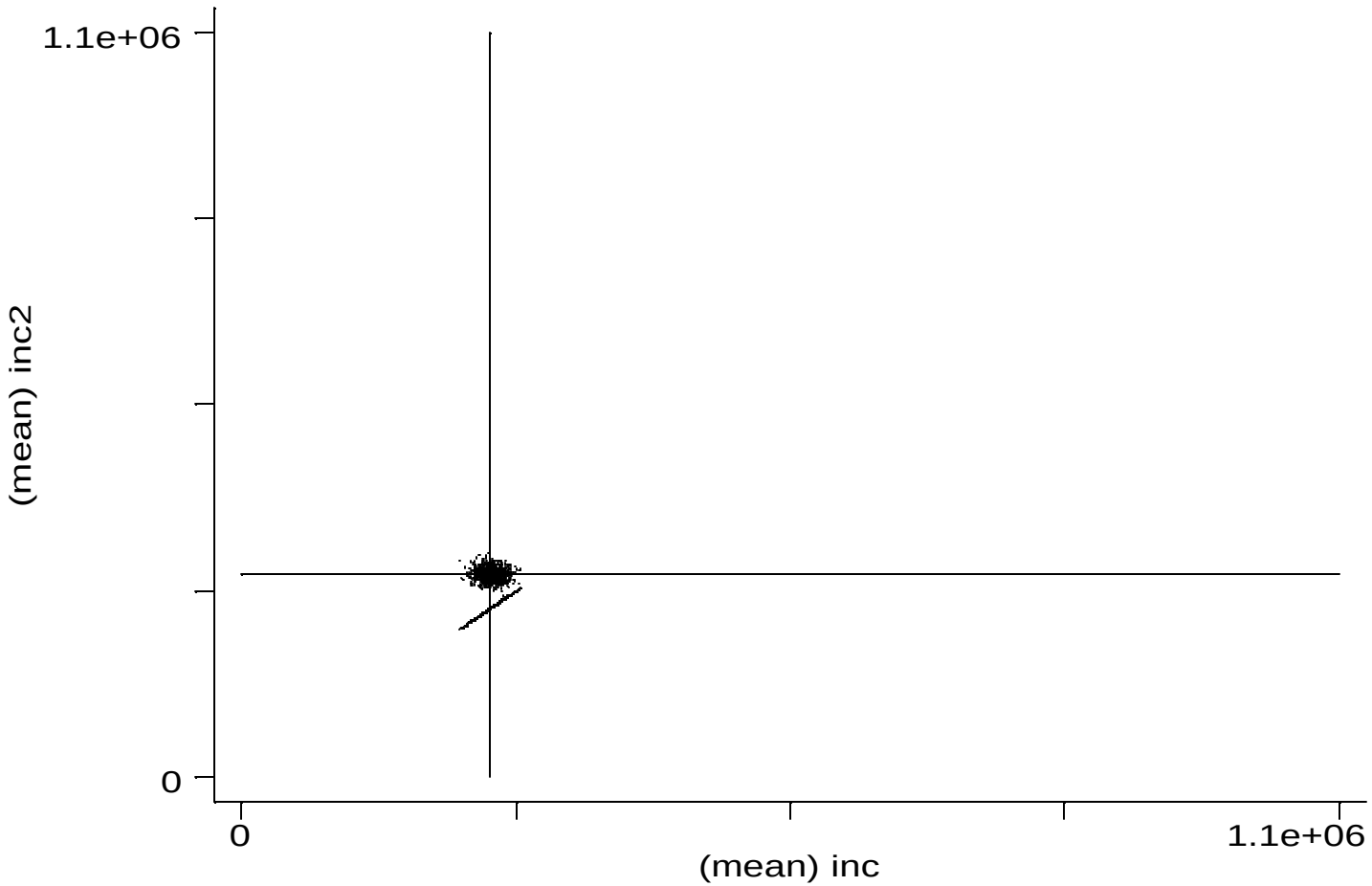
State 2 > State 1: 673 times

1,000 samples of 100



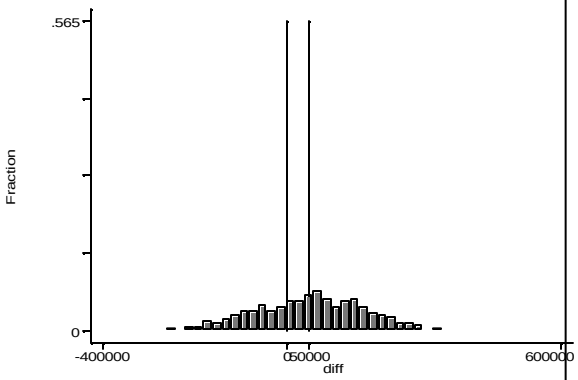
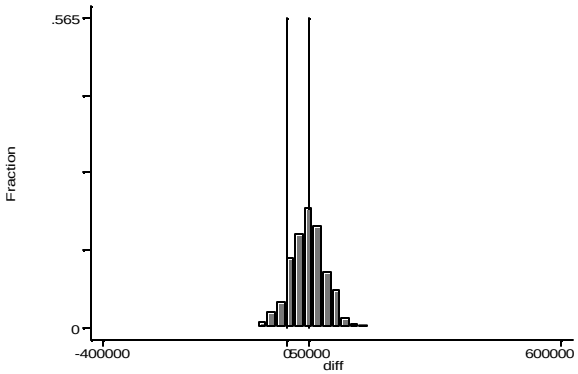
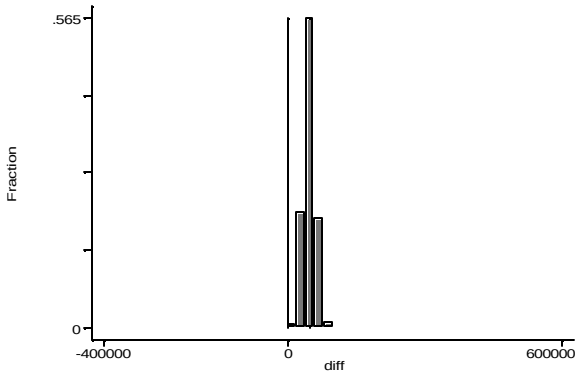
State 2 > State 1: 909 times

1,000 samples of 1,000



State 2 > State 1: 1,000 times

Another way of looking at it: The distribution of $Inc_2 - Inc_1$

n = 10	n = 100	n = 1,000
		
Mean = 51,845 s.d. = 124,815	Mean = 49,704 s.d. = 38,774	Mean = 49,816 s.d. = 13,932

Reasoning Backward

When you know n, $\overline{X}$, and s,

but want to say something about m

Central Limit Theorem

As the sample size *n* increases, the distribution of the mean $\overline{X}$ of a random sample taken from **practically any population** approaches a *normal* distribution, with mean ： and standard deviation $s/\sqrt{n}$

Calculating Standard Errors

In general:

$$\text{std. err.} = \frac{s}{\sqrt{n}}$$

Most important standard errors

Mean	$\frac{s}{\sqrt{n}}$
Proportion	$\sqrt{\frac{p(1-p)}{n}}$
Diff. of 2 means	$s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
Regression (slope) coeff.	$\frac{s.e.r.}{\sqrt{n}}\times\frac{1}{s_x}$

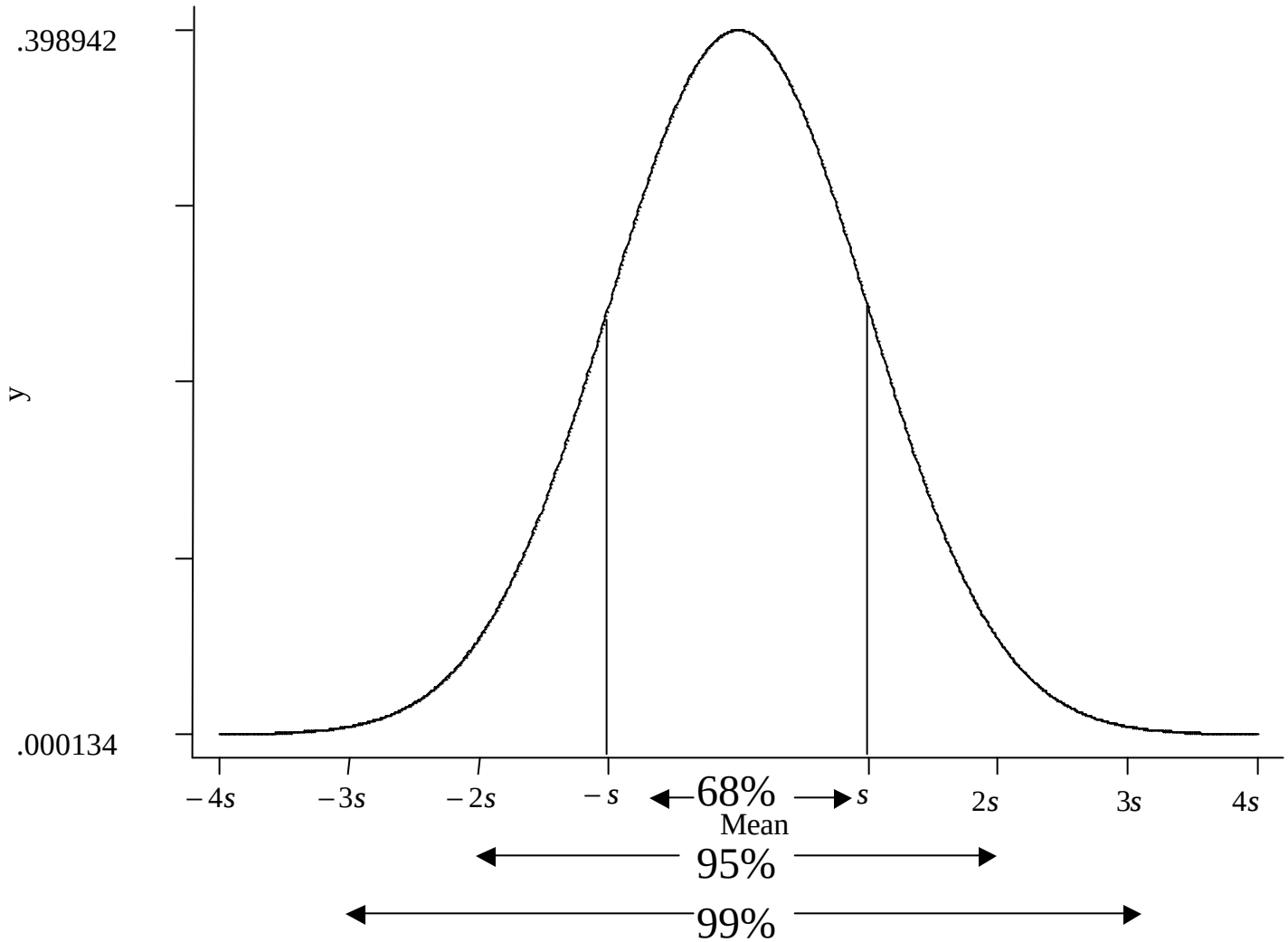
If you know the sample mean, s.d., and *n*, what can you say about the population mean?

In general,

$$\text{population mean} =$$

$$\text{sample mean} \pm \text{arbitrary interval} \times \text{standard error}$$

If *n* is sufficiently large, choose the interval using the normal curve



Population mean using original example (n = 10)

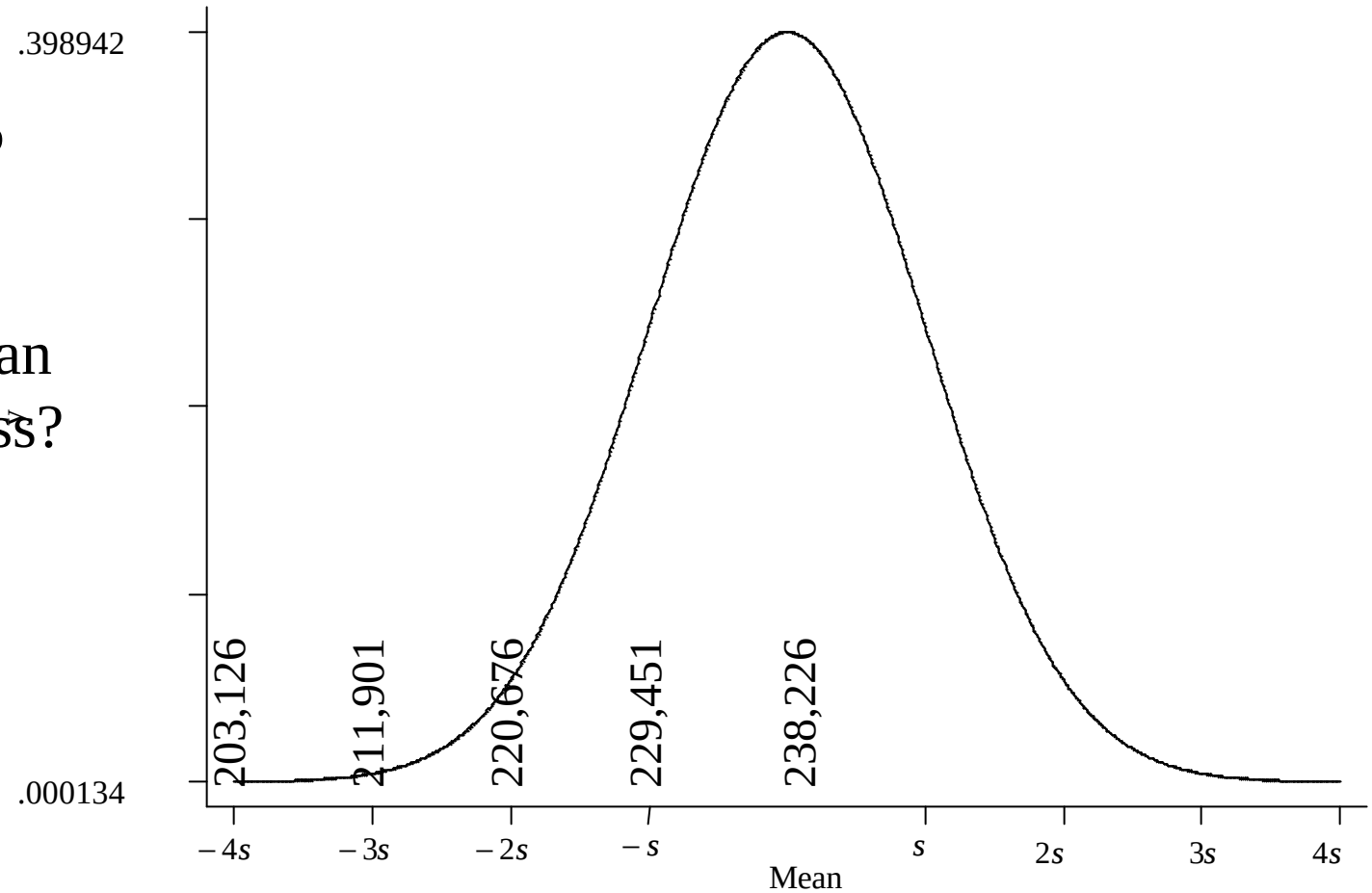
				68%		95%		99%	
Sample	Mean	s.d.	s.e.	lower	upper	lower	upper	lower	upper
1	311,410	241,392	76,335	**235,075**	**387,744**	**158,740**	**464,079**	**82,405**	**540,414**
2	184,571	215,655	68,196	**116,375**	**252,767**	**48,179**	**320,963**	**-20,017**	**389,159**
3	468,574	348,908	110,334	358,240	578,909	**247,905**	**689,243**	**137,571**	**799,578**
4	253,574	321,599	101,699	**151,875**	**355,272**	**50,177**	**456,971**	**-51,522**	**558,669**
5	220,934	273,256	86,411	**134,522**	**307,345**	**48,111**	**393,756**	**-38,300**	**480,167**
6	270,400	346,008	109,417	**160,983**	**379,817**	**51,565**	**489,235**	**-57,852**	**598,652**
7	127,115	197,071	62,319	64,796	189,435	**2,477**	**251,754**	**-59,842**	**314,073**
8	253,885	127,711	40,386	**213,500**	**294,271**	**173,114**	**334,657**	**132,728**	**375,043**
9	152,678	201,009	63,564	89,113	216,242	**25,549**	**279,806**	**-38,016**	**343,371**
10	222,725	264,339	83,591	**139,134**	**306,317**	**55,543**	**389,908**	**-28,048**	**473,499**

Population mean using original example (n = 1000)

				68%		95%		99%	
Sample	Mean	s.d.	s.e.	lower	upper	lower	upper	lower	upper
1	238,226	277,492	8,775	229,450	247,001	**220,675**	**255,776**	**211,900**	**264,551**
2	260,658	290,954	9,201	251,458	269,859	**242,257**	**279,060**	**233,056**	**288,261**
3	253,374	277,022	8,760	**244,614**	**262,134**	**235,853**	**270,894**	**227,093**	**279,655**
4	242,002	283,772	8,974	**233,028**	**250,975**	**224,055**	**259,949**	**215,081**	**268,923**
5	244,437	279,343	8,834	**235,603**	**253,271**	**226,770**	**262,104**	**217,936**	**270,938**
6	248,896	279,213	8,829	**240,067**	**257,726**	**231,237**	**266,555**	**222,408**	**275,385**
7	267,218	291,150	9,207	258,011	276,425	**248,804**	**285,632**	**239,597**	**294,839**
8	244,138	276,490	8,743	**235,394**	**252,881**	**226,651**	**261,624**	**217,908**	**270,368**
9	247,996	275,994	8,728	**239,268**	**256,723**	**230,540**	**265,451**	**221,813**	**274,179**
10	255,023	287,118	9,079	**245,944**	**264,103**	**236,864**	**273,182**	**227,785**	**282,262**

Another way of asking this: The *z*-ratio

With
mean = 238,226
s.e. = 8,775,
how likely is it
that the true mean
is 200,000 or less?



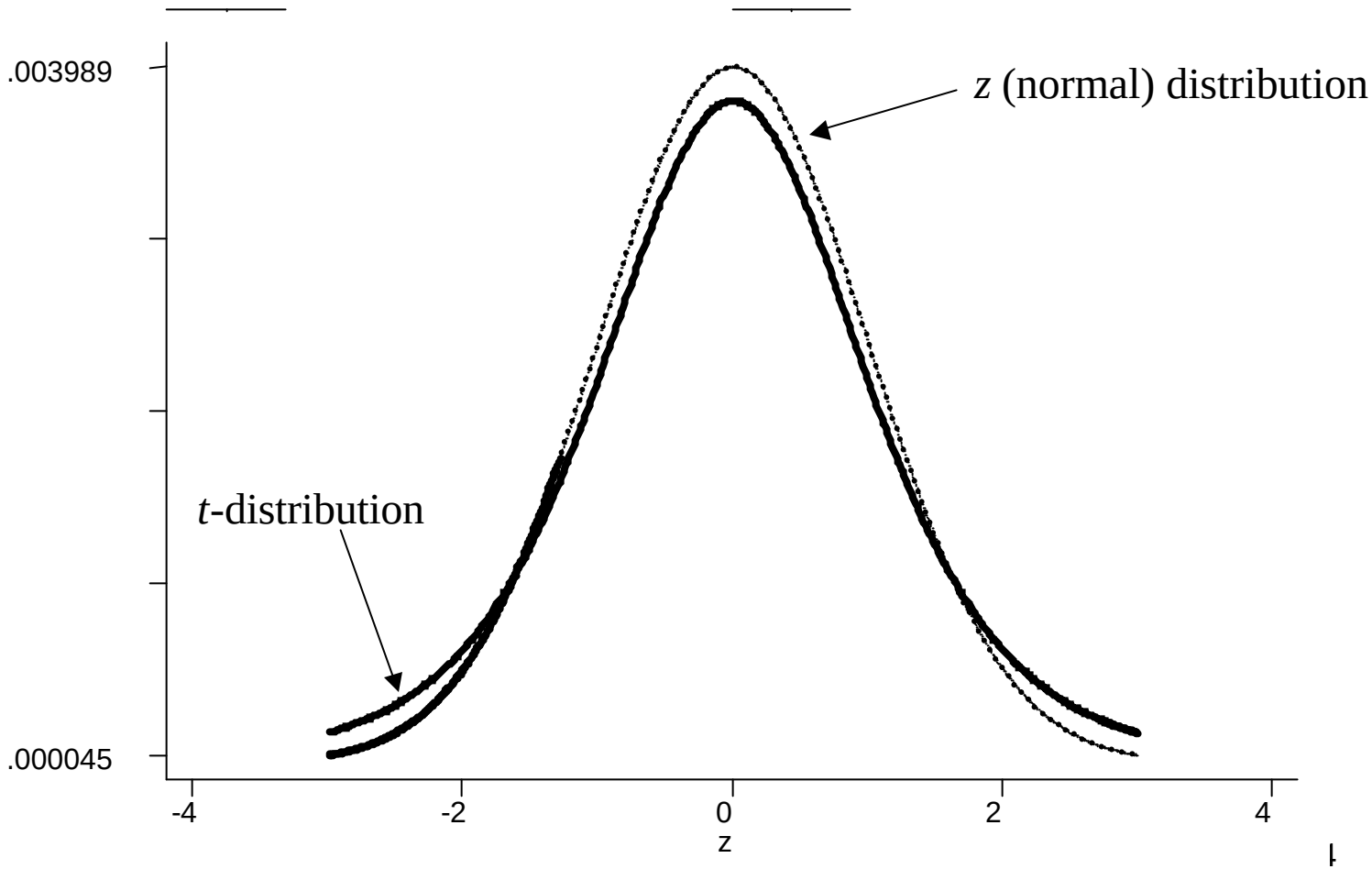
$$z$$

$$z = \frac{(\text{Sample mean - test value})}{\text{standard error}},$$

in this case,

$$z = \frac{(238{,}226 - 200{,}000)}{8{,}775} = 4.37$$

t
(when the sample is small)



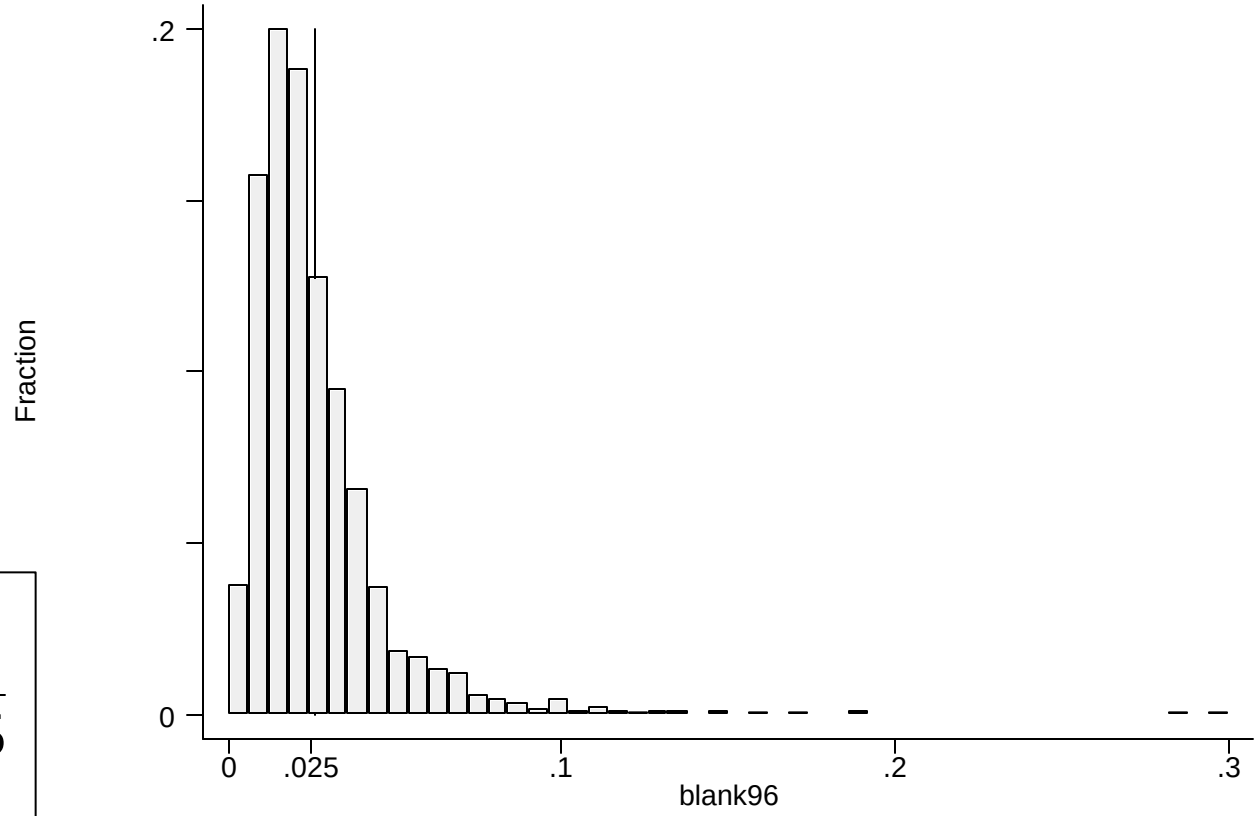
Reading a *z* table

Reading a *t* table

Doing a *t*-test

Q: How likely is it that the residual vote rate n 1996 was 2.5% or less?

Mean: 0.02618
s.d.: 0.02140
N: 1905



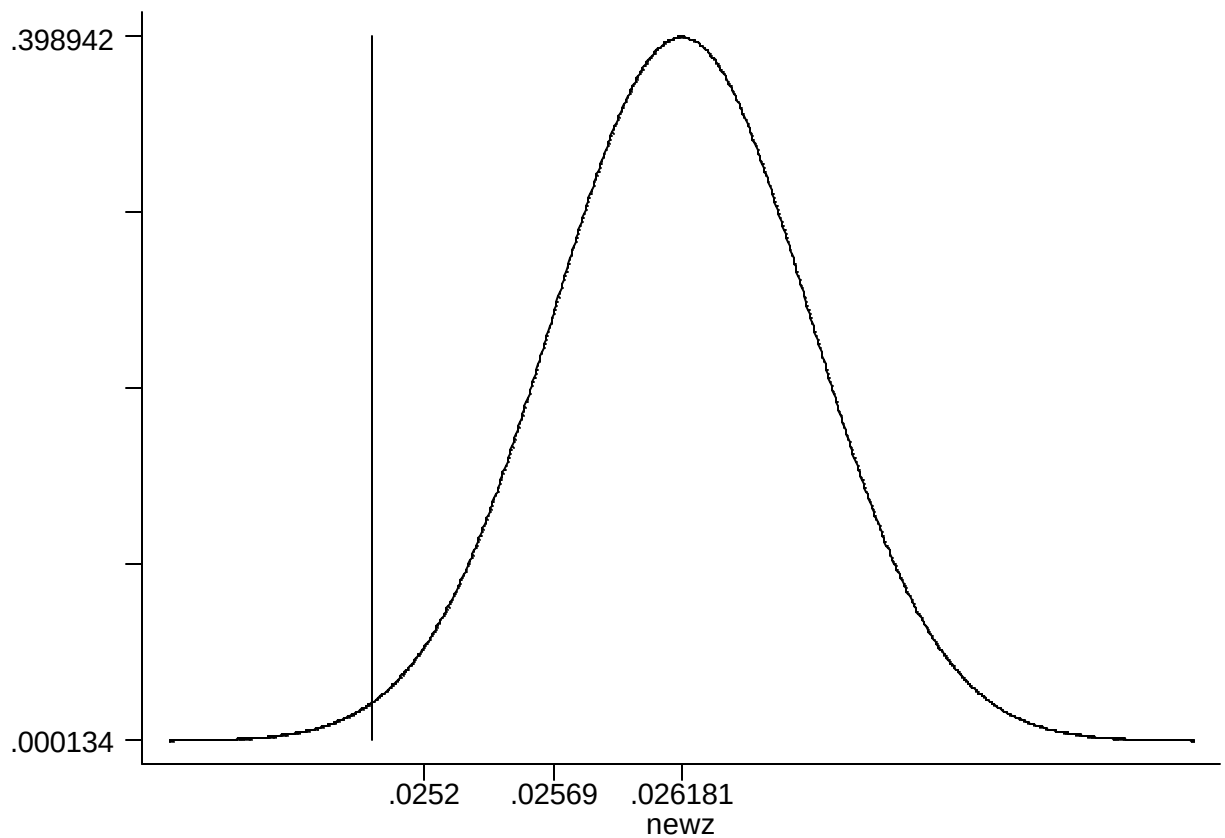
$$s.e. = s/\sqrt{n}$$
$$= 0.02140/\sqrt{1905}$$
$$= 0.00049$$

The picture

Mean: 0.02618
s.d.: 0.02140
N: 1905

$$s.e. = s/\sqrt{n}$$
$$= 0.02140/\sqrt{1905}$$
$$= 0.00049$$

$$t = \frac{0.026181 - .025}{0.00049}$$
$$= 2.408$$



The *STATA* output

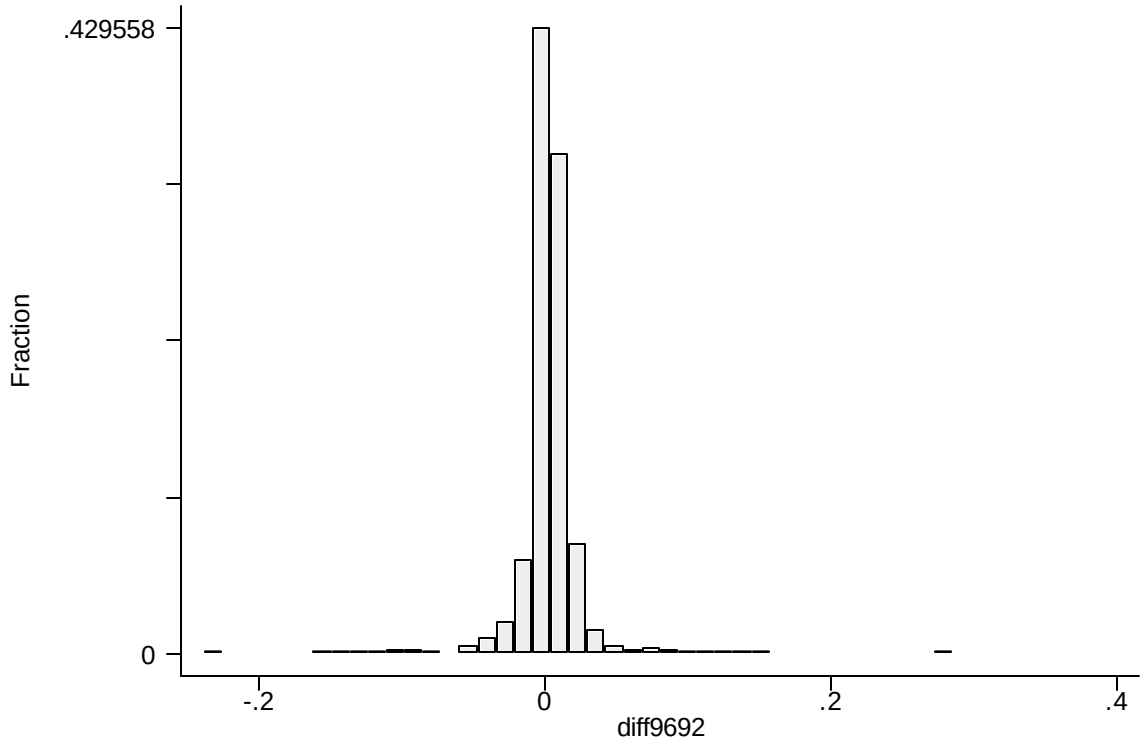
```
. ttest blank96=.025

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           Ha: mean ~= .025            Ha: mean > .025
       t =   2.4082                t =   2.4082               t =   2.4082
   P < t =   0.9919          P > |t| =   0.0161           P > t =   0.0081
```

Doing another *t*-test

Q: How likely is it that the residual vote rate in 1996 equal to the rate in 1992 (I.e., blank_{96}-blank_{92}= 0)?

Mean: 0.003069
s.d.: 0.02323
N: 1448

$$s.e. = s/\sqrt{n}$$
$$= 0.02323/\sqrt{1448}$$
$$= 0.00061$$

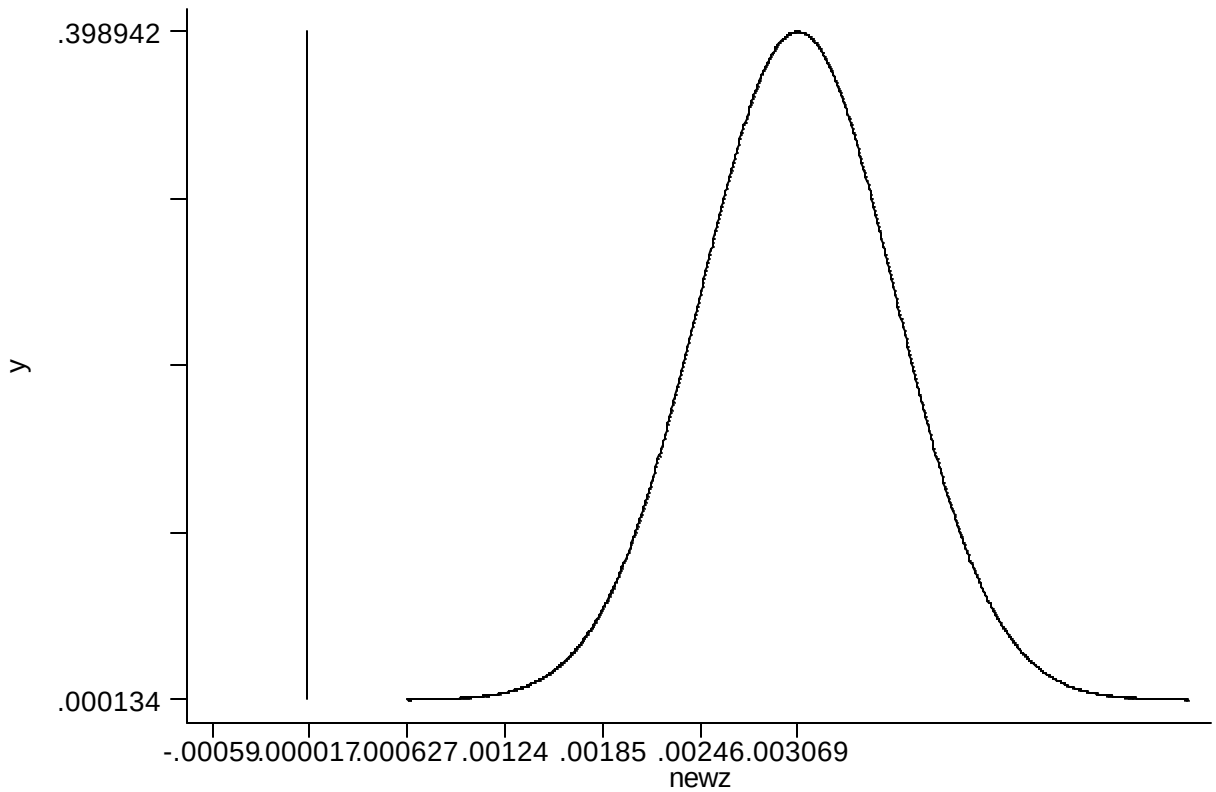


The picture

Mean: 0.003069
s.d.: 0.02323
N: 1448

$$s.e. = s/\sqrt{n}$$
$$= 0.02323/\sqrt{1448}$$
$$= 0.00061$$

$$t = \frac{0.003069 - 0}{0.00061}$$
$$= 5.028$$



The *STATA* output

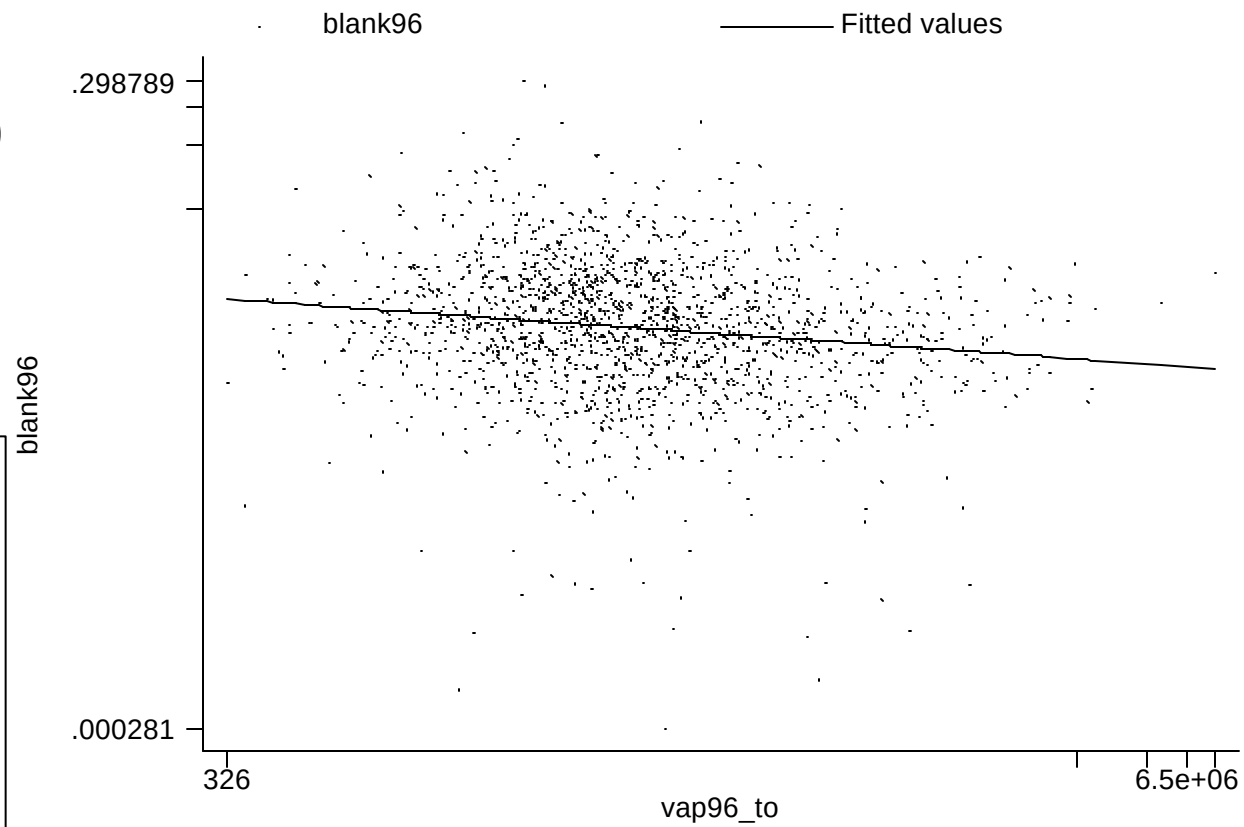
```
. ttest blank96=blank92

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         Ha: mean(diff) ~= 0        Ha: mean(diff) > 0
       t =   5.0278               t =   5.0278               t =   5.0278
   P < t =   1.0000           P > |t| =   0.0000           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               Ha: mean ~= 0              Ha: mean > 0
       t =   5.0278               t =   5.0278               t =   5.0278
   P < t =   1.0000           P > |t| =   0.0000           P > t =   0.0000

.
```

Final *t*-test

Q: Was there a relationship between residual vote and county Size in 1996?

Slope coeff: -0.07510
s.e.r: 0.7115
N: 1861
S_x: 1.4788

$$
\begin{aligned}
s.e. &= \frac{s.e.r}{\sqrt{n}} \times \frac{1}{s_x} \\
&= \frac{0.7115}{\sqrt{1861}} \times \frac{1}{1.4788} \\
&= 0.01649 \times 0.6762 \\
&= 0.01115
\end{aligned}
$$



Calculating t

$$t = \frac{-0.07510}{.01115}$$

$$= -6.7319$$

The *STATA* output

```
. reg lblank96 lvap96

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