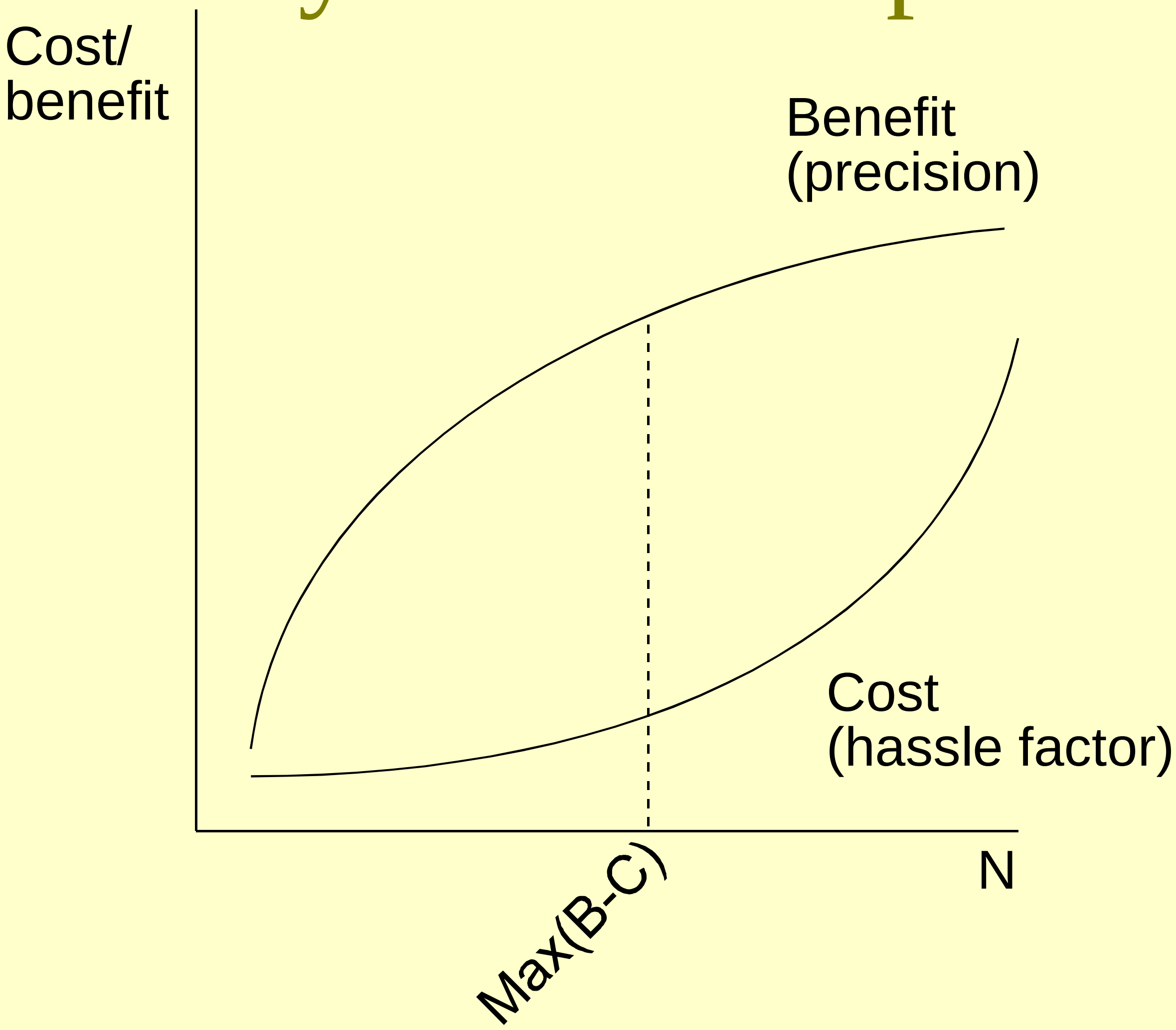
Sampling and Inference

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

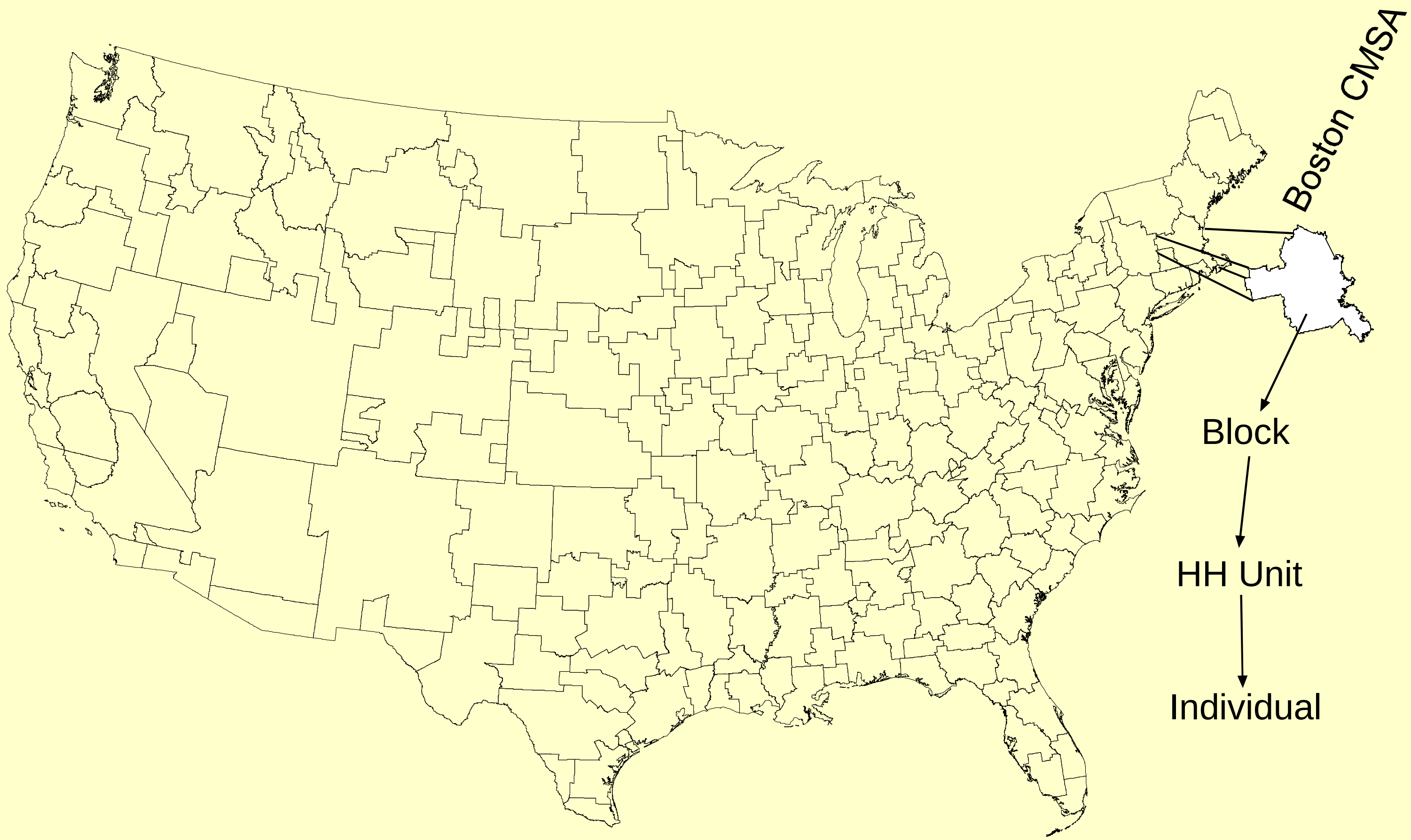
Stratification

- 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.)		

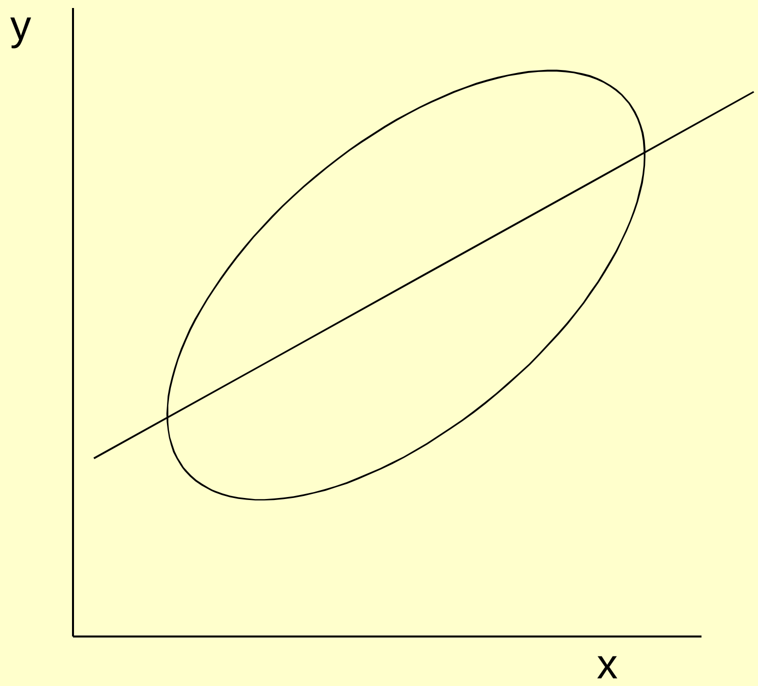
Cluster sampling

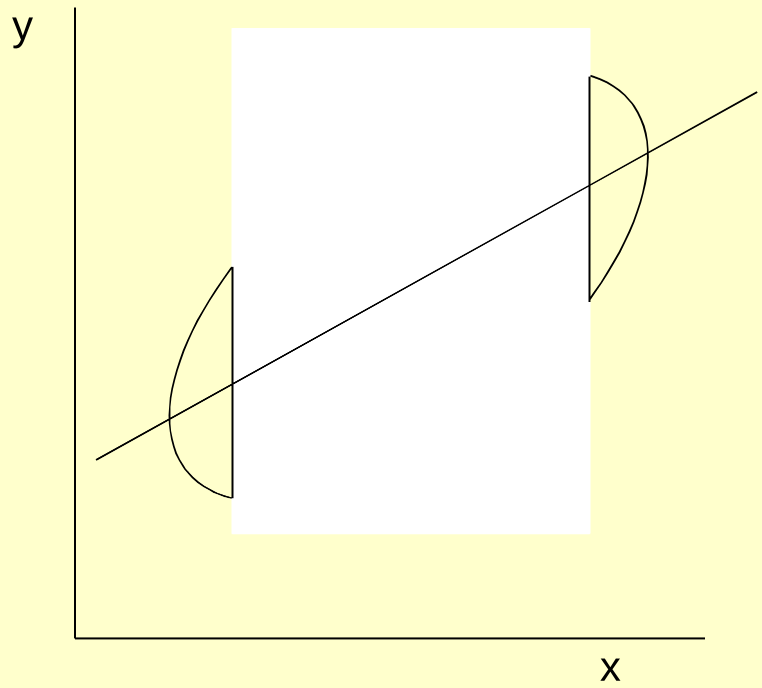


Effects of samples

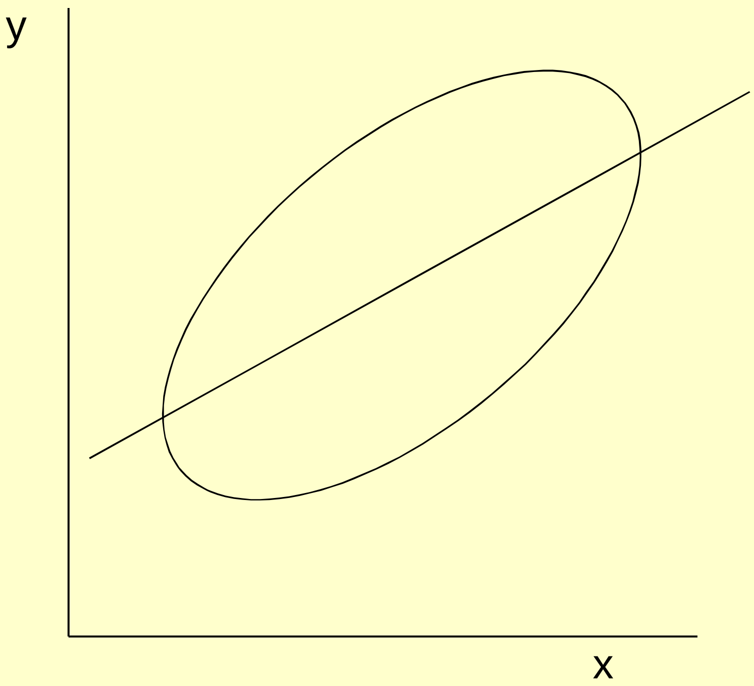
- Obvious: influences marginals
- Less obvious
  - Allows effective use of time and effort
  - Effect on multivariate techniques
    - Sampling of independent variable: greater precision in regression estimates
    - Sampling on dependent variable: bias

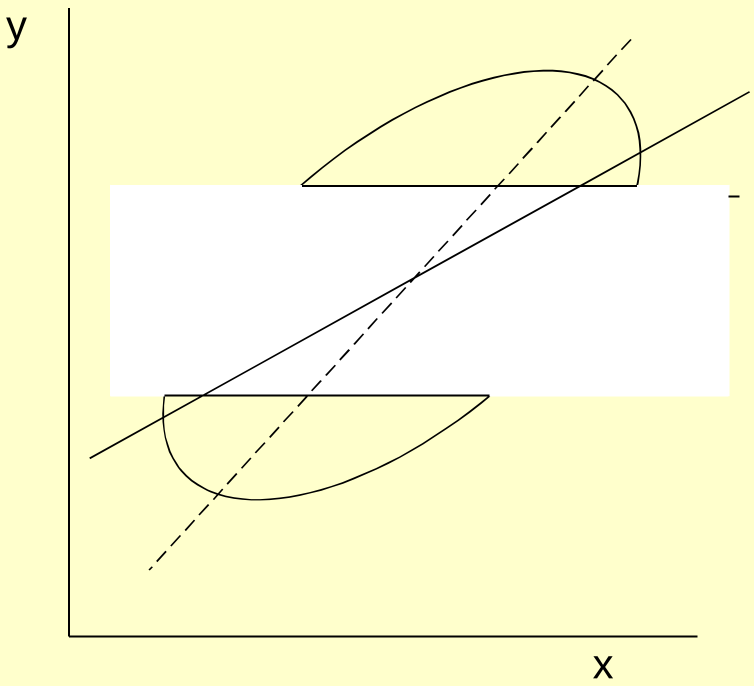
Sampling on Independent Variable





Sampling on Dependent Variable





Sampling

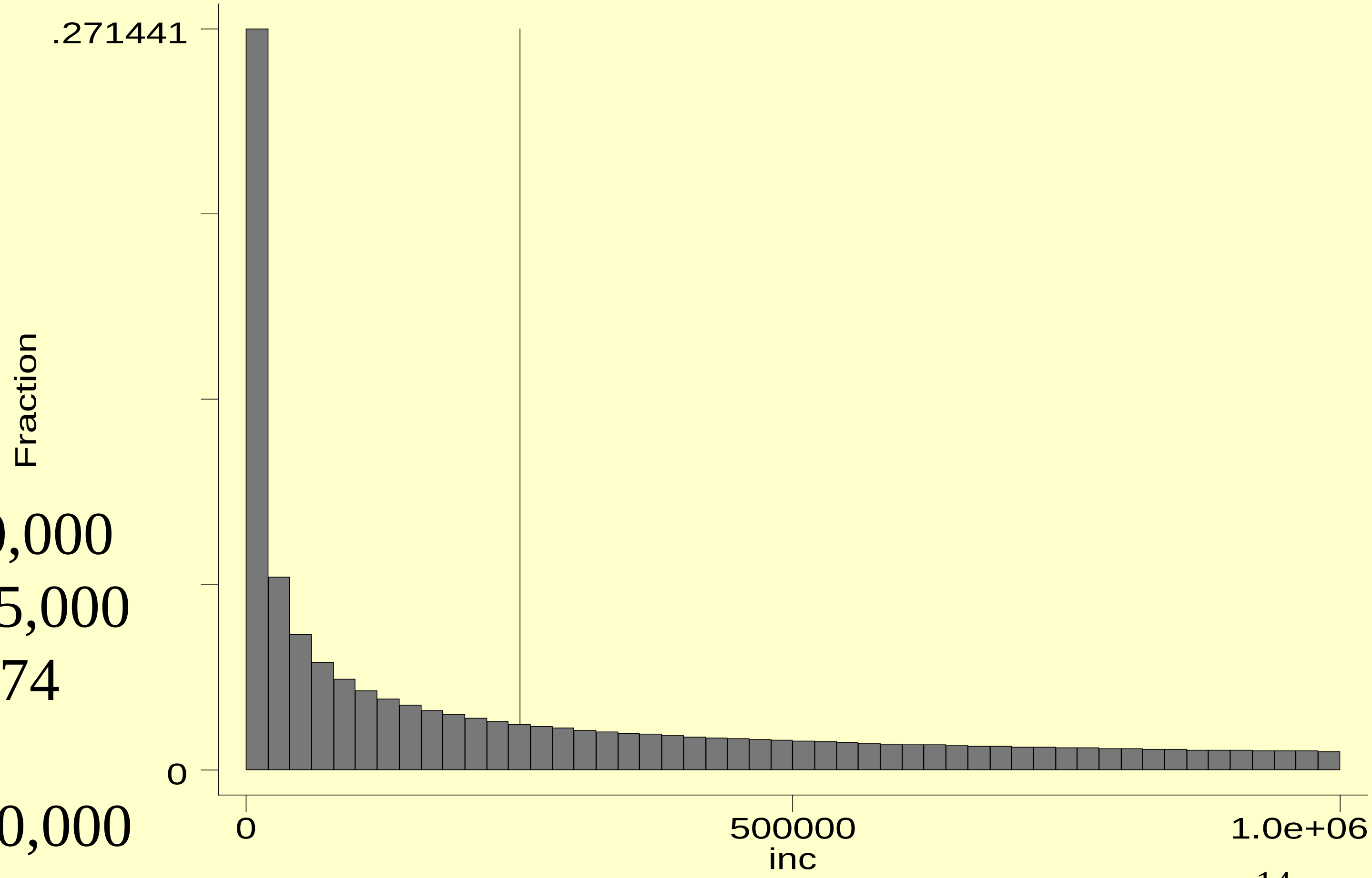
Consequences for Statistical Inference

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

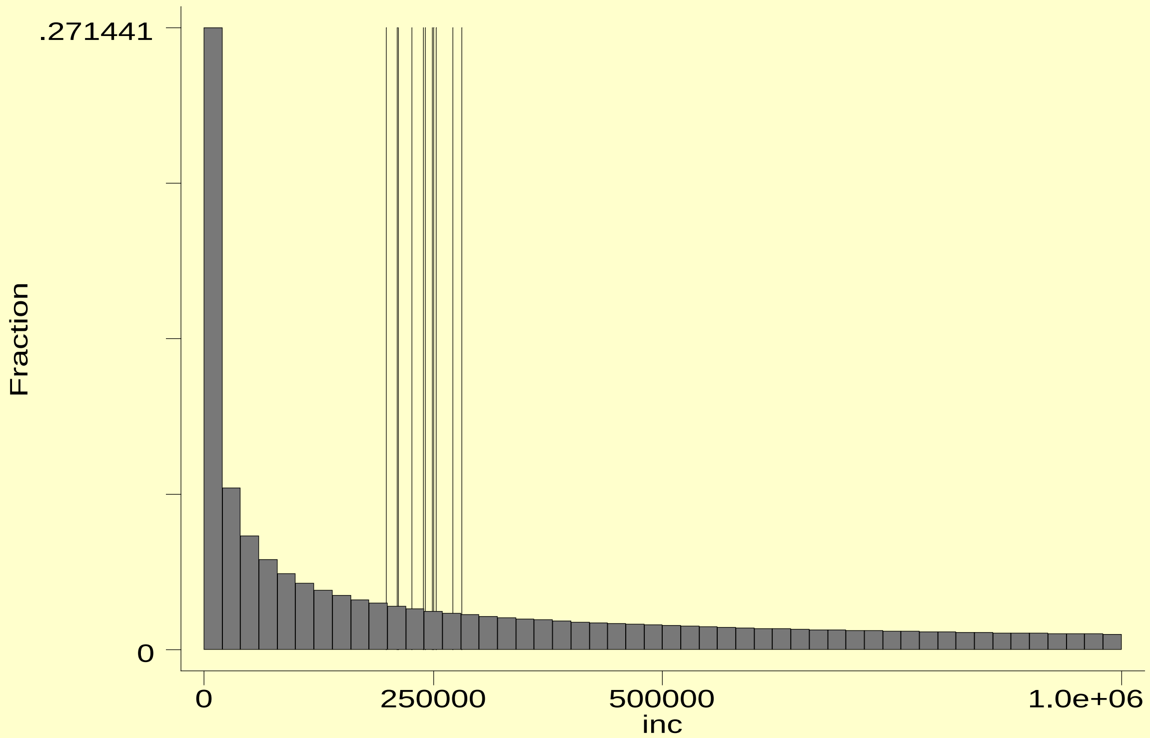
Exponential Distribution Example



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

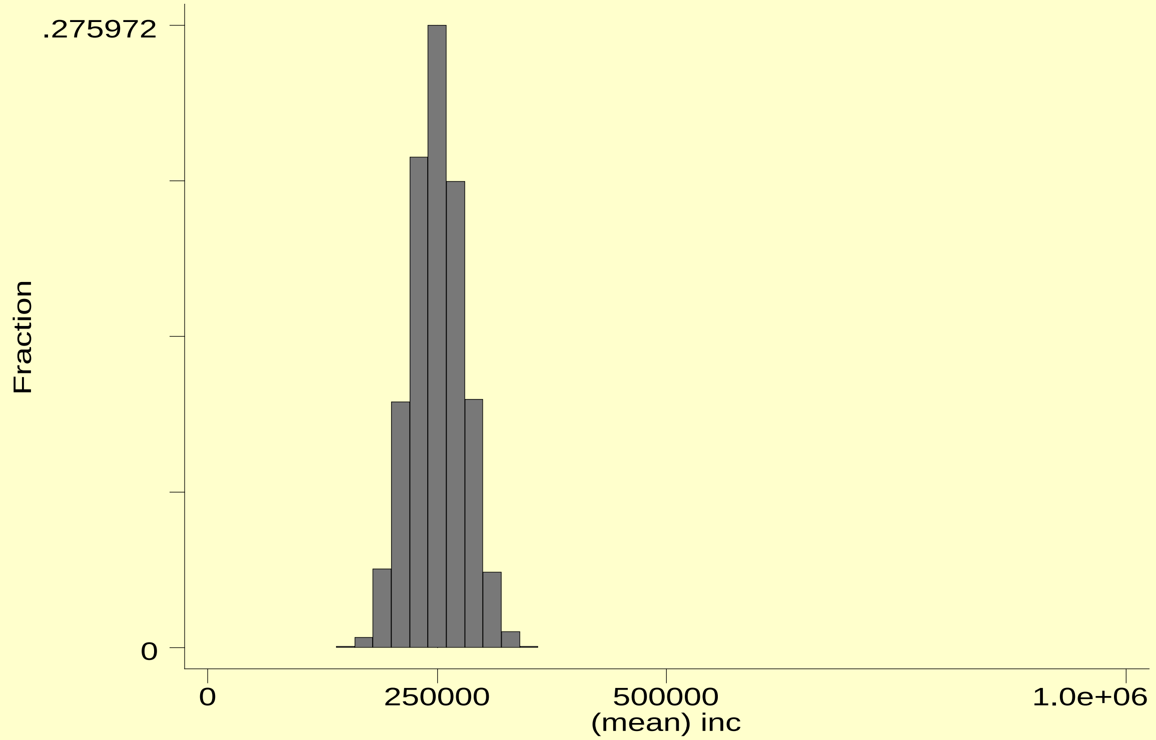
Consider 10 random samples, of n = 100 apiece

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

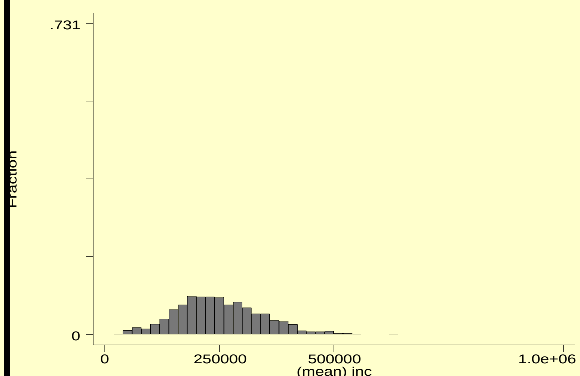
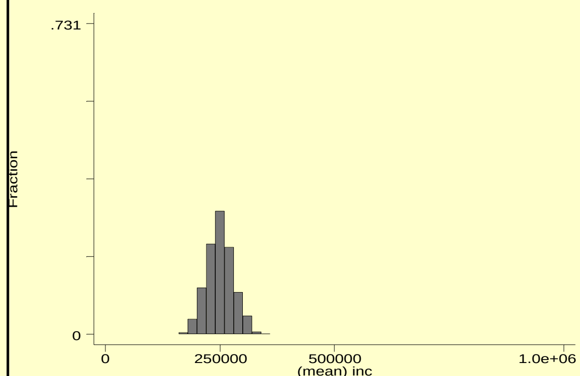
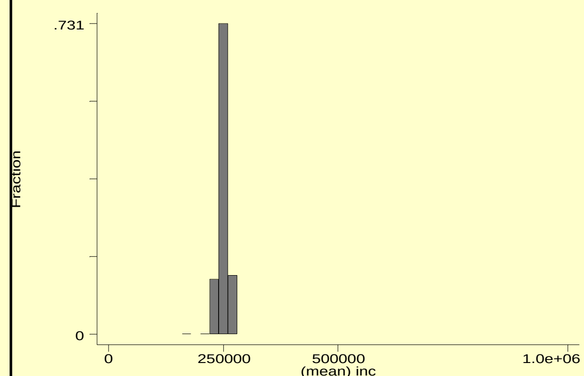


Consider 10,000 samples of *n* = 100

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

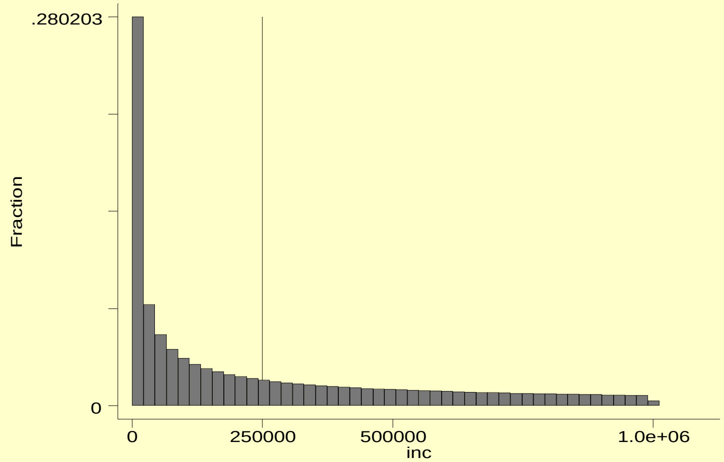


Consider 1,000 samples of various sizes

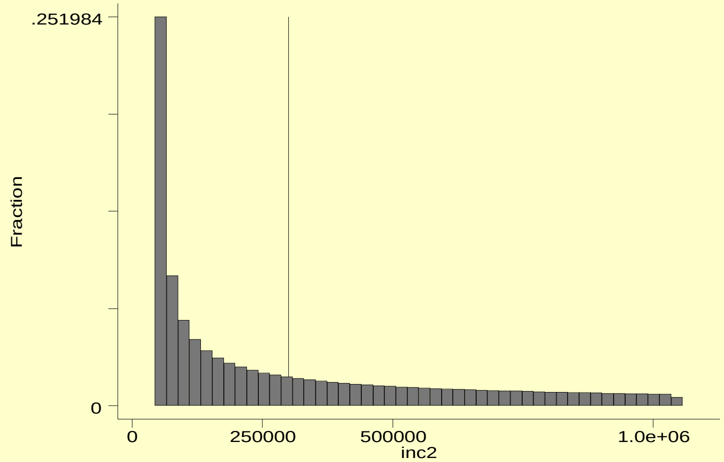
10	100	1000
		
Mean =250,105	Mean = 250,498	Mean = 249,938
s.d.= 90,891	s.d.= 28,297	s.d.= 9,376
Skew= 0.38	Skew= 0.02	Skew= -0.50
Kurt= 3.13	Kurt= 2.90	Kurt= 6.80

Difference of means example

State 1
Mean = 250,000



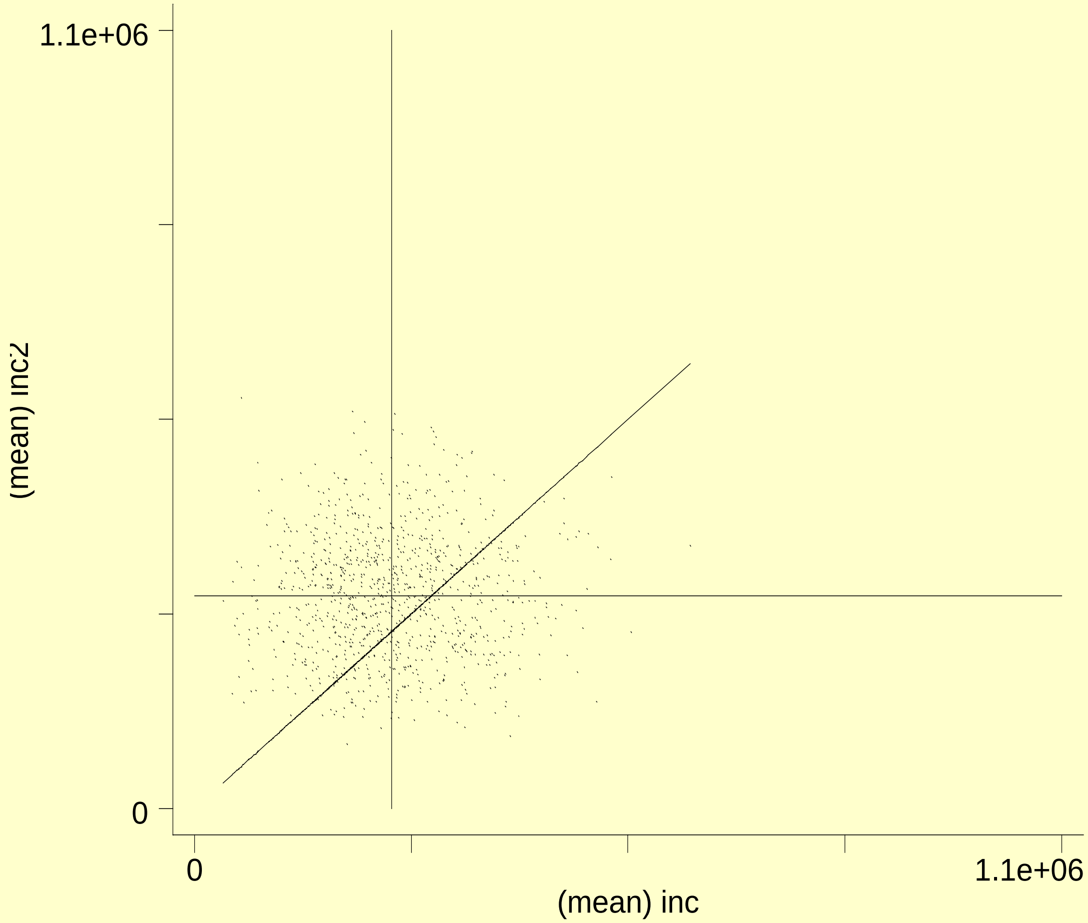
State 2
Mean = 300,000



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

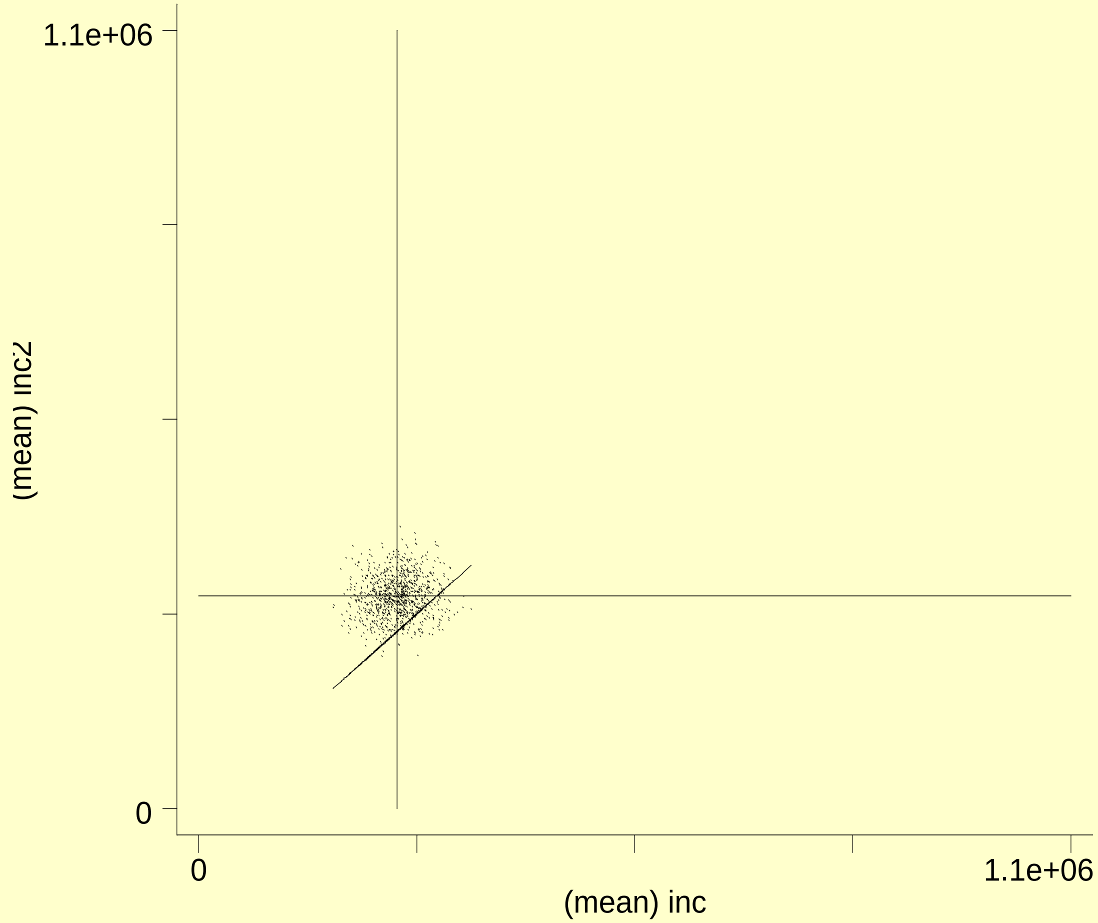
First 10 samples			
Sample	State 1		State 2
1	311,410	<	365,224
2	184,571	<	243,062
3	468,574	>	438,336
4	253,374	<	557,909
5	220,934	>	189,674
6	270,400	<	284,309
7	127,115	<	210,970
8	253,885	<	333,208
9	152,678	<	314,882
10	222,725	>	152,312

1,000 samples of 10



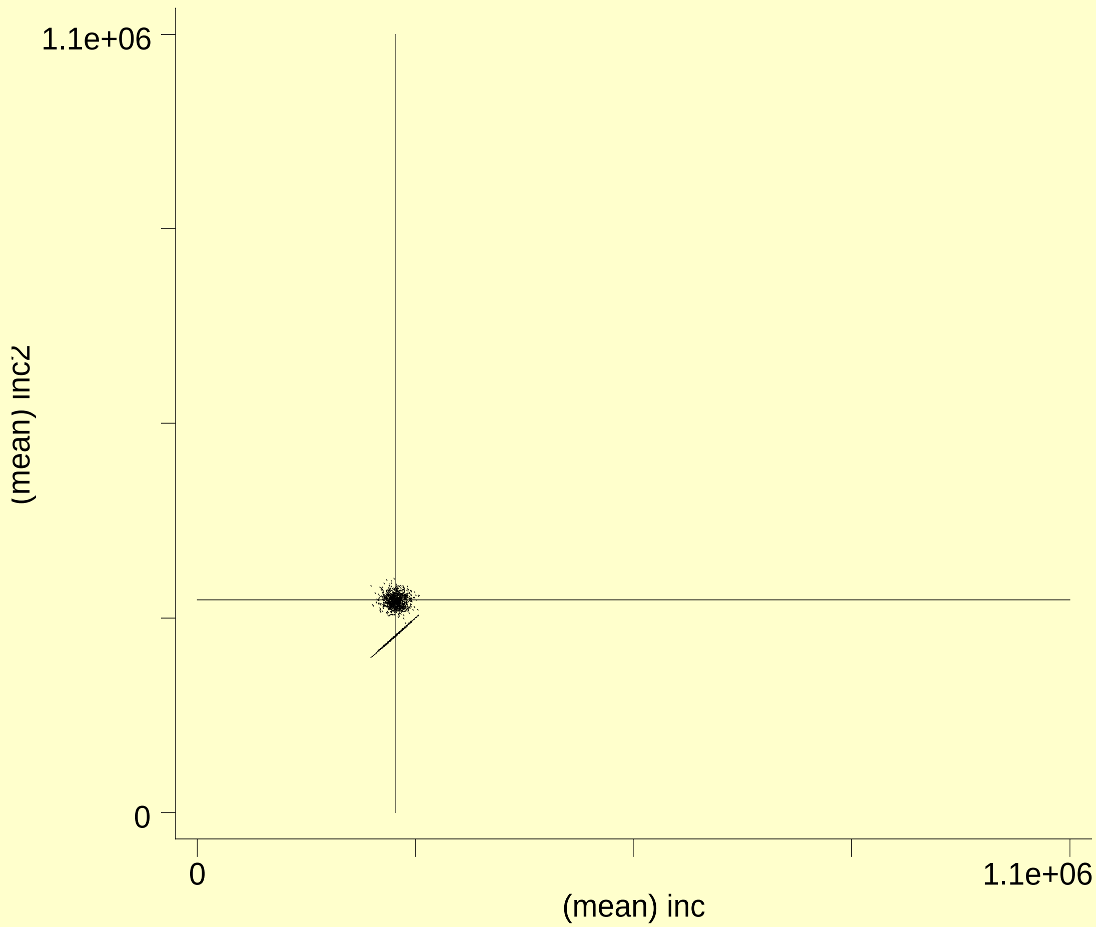
State 2 > State 1: 673 times

1,000 samples of 100



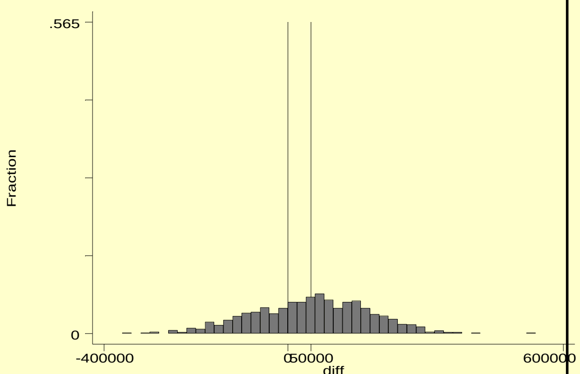
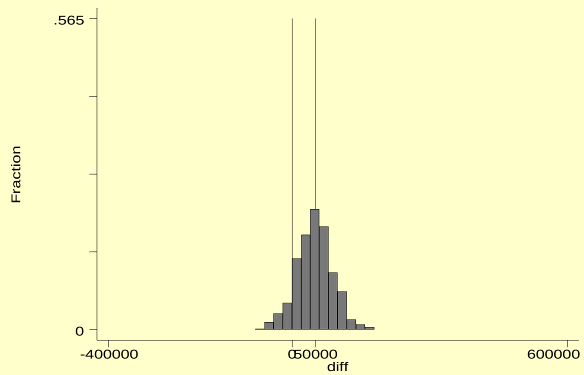
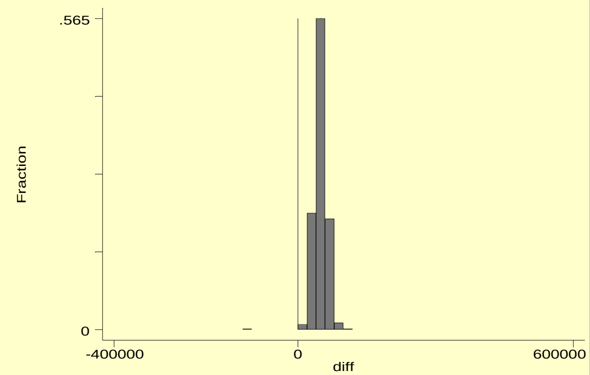
State 2 > State 1: 909 times

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, $\overline{\mathrm{X}}$, and s,

but want to say something about μ

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 μ and standard deviation $\sigma/\sqrt{n}$

Calculating Standard Errors

In general:

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

Most important standard errors

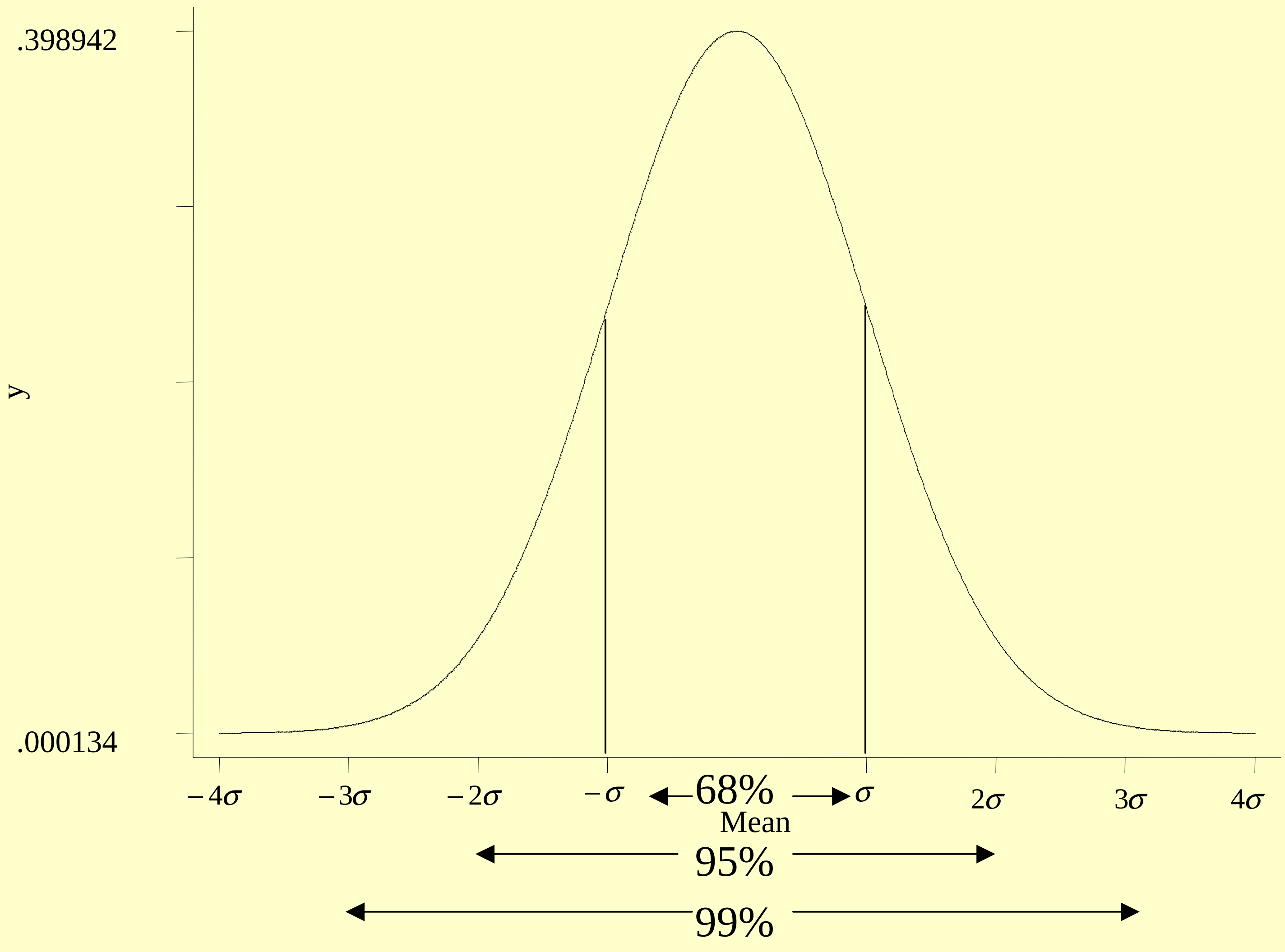
Mean	$\dfrac{s}{\sqrt{n}}$
Proportion	$\sqrt{\dfrac{p(1-p)}{n}}$
Diff. of 2 means	$s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$
Regression (slope) coeff.	$\dfrac{s.e.r.}{\sqrt{n}}\times\dfrac{1}{s_x}$

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)

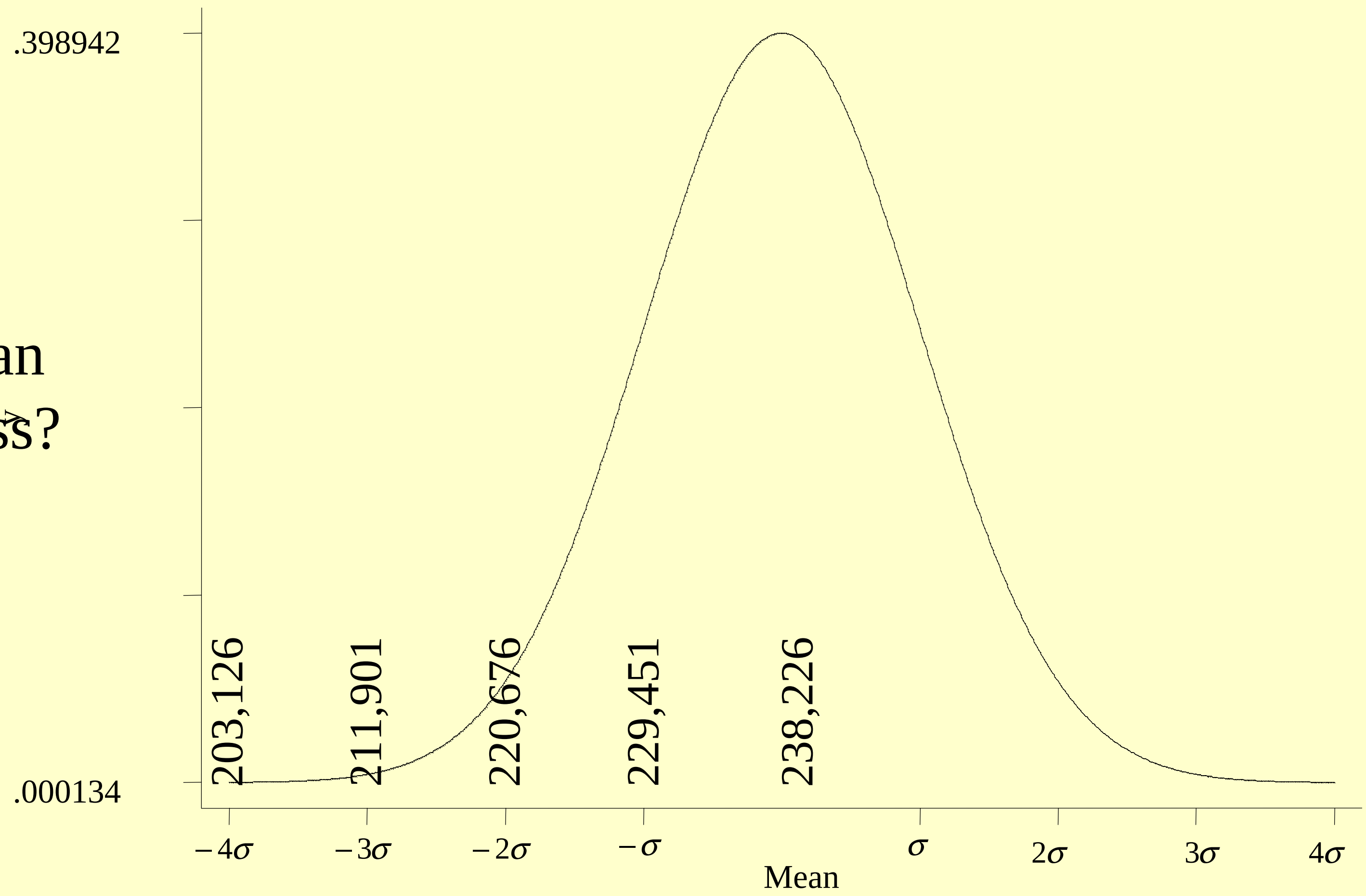
				68%		95%		99%	
Sample	Mean	s.d.	s.e.	lower	upper	lower	upper	lower	upper
1	311,410	241,392	76,335	**235,075**	**387,744**	**158,740**	**464,079**	**82,405**	**540,414**
2	184,571	215,655	68,196	**116,375**	**252,767**	**48,179**	**320,963**	**-20,017**	**389,159**
3	468,574	348,908	110,334	358,240	578,909	**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?



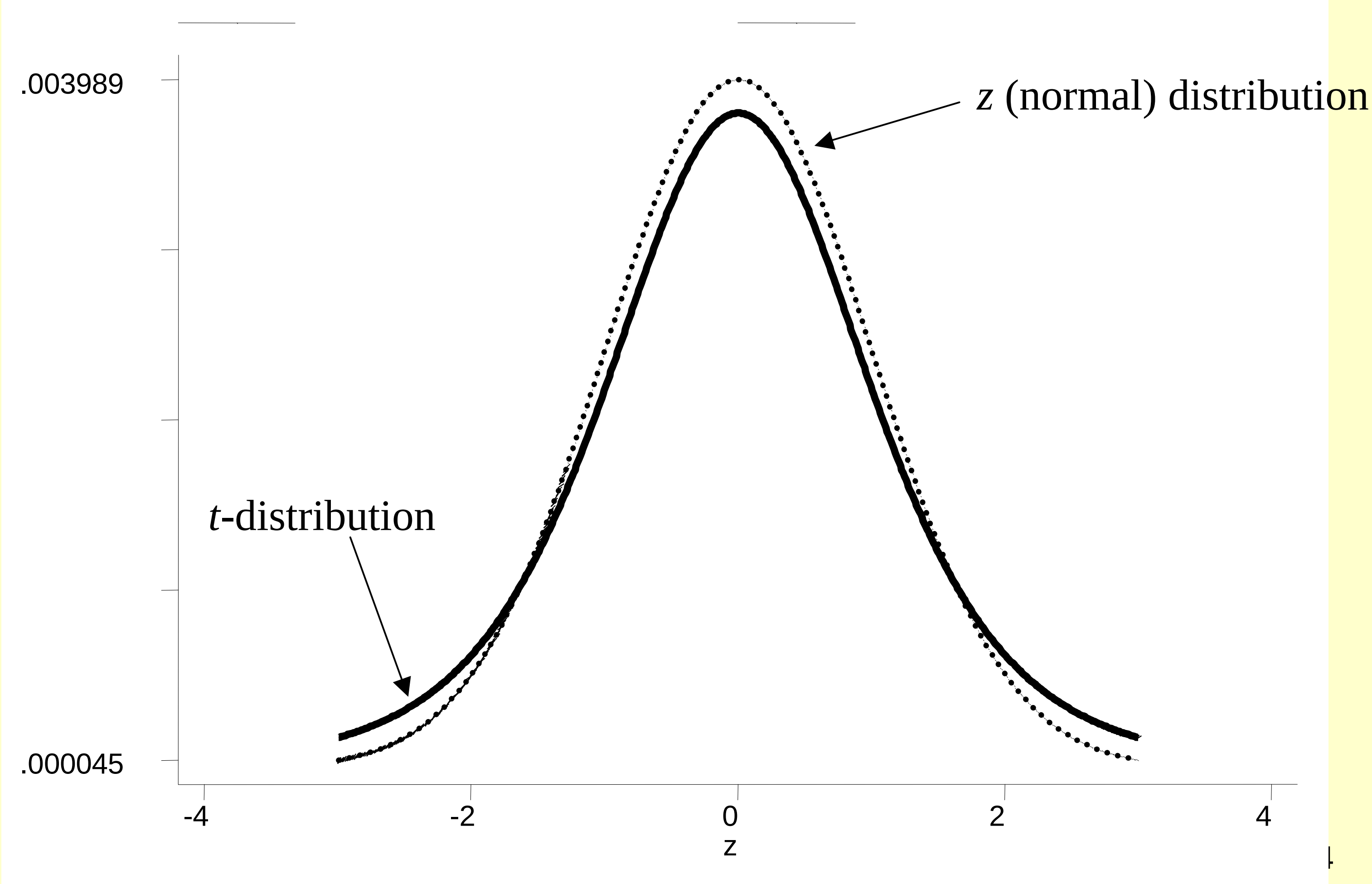
z

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

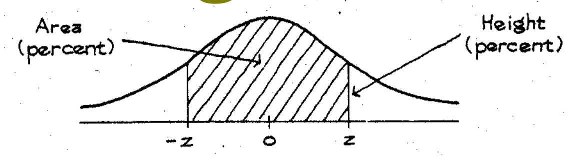
in this case,

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

t
(when the sample is small)



Reading a z table

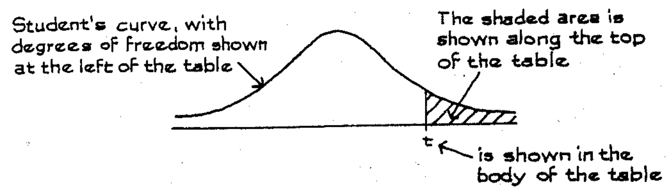


A NORMAL TABLE

z	Height	Area	z	Height	Area	z	Height	Area
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

Reading a *t* table

A *t*-TABLE

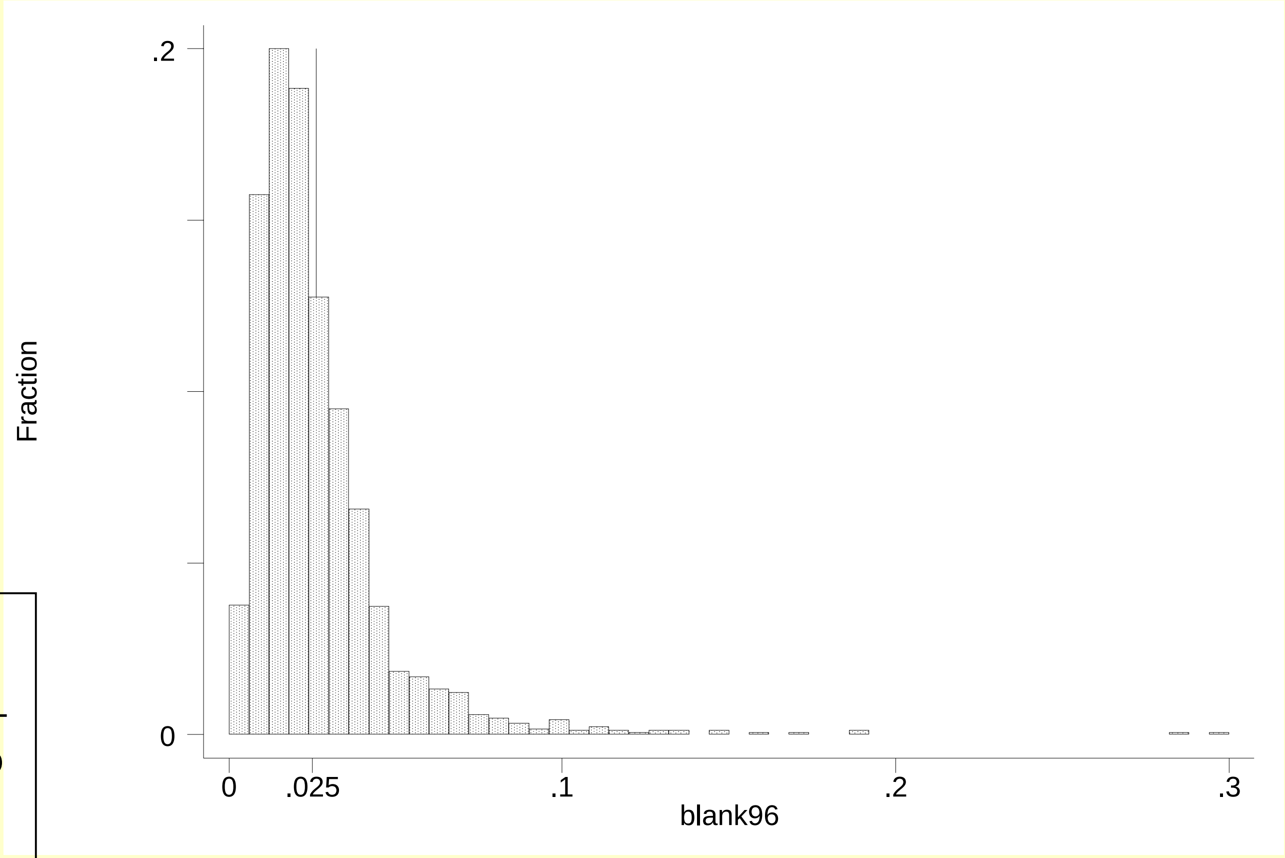


Degrees of freedom	25%	10%	5%	2.5%	1%	0.5%
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
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

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



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

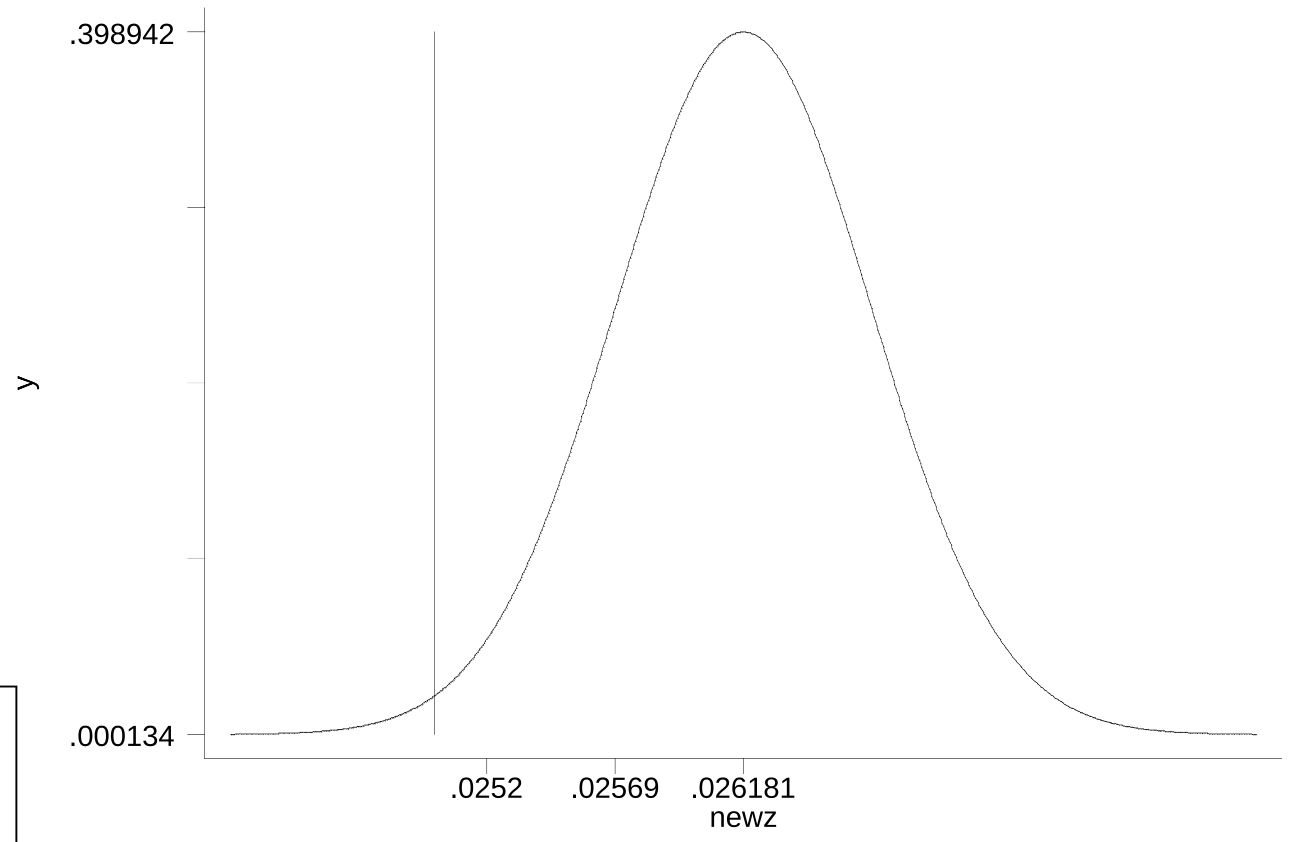
$$= 0.02140/\sqrt{1905}$$

$$= 0.00049$$

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

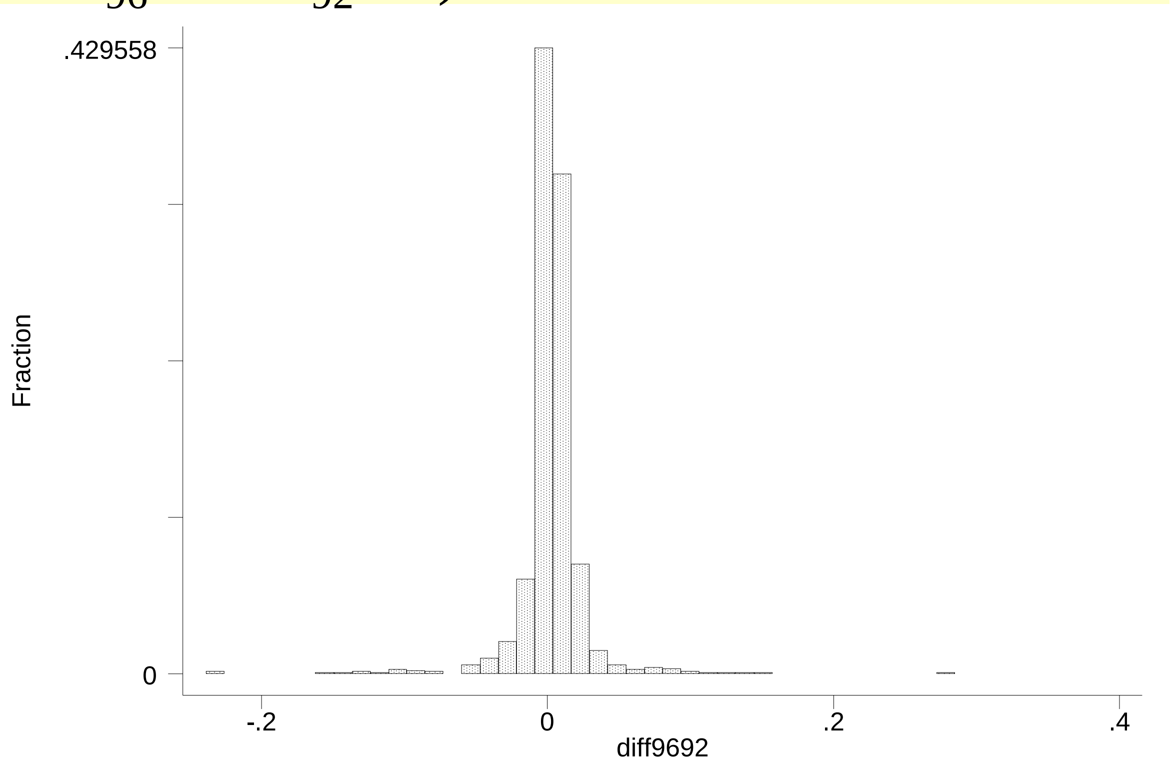
Q: How likely is it that the residual vote rate in 1996 equal to the rate in 1992 (I.e., blank_{96}-blank_{92}= 0)?

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

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

$$= 0.02323 / \sqrt{1448}$$

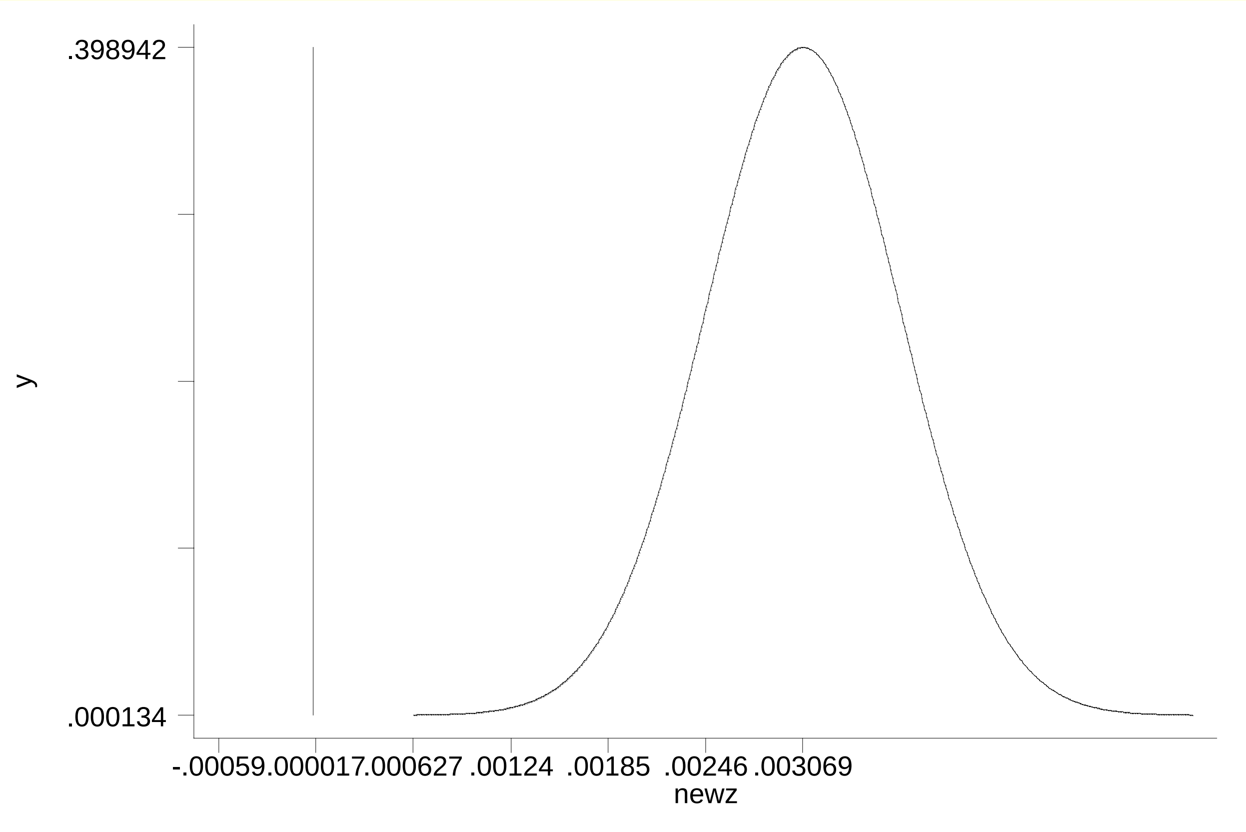
$$= 0.00061$$



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

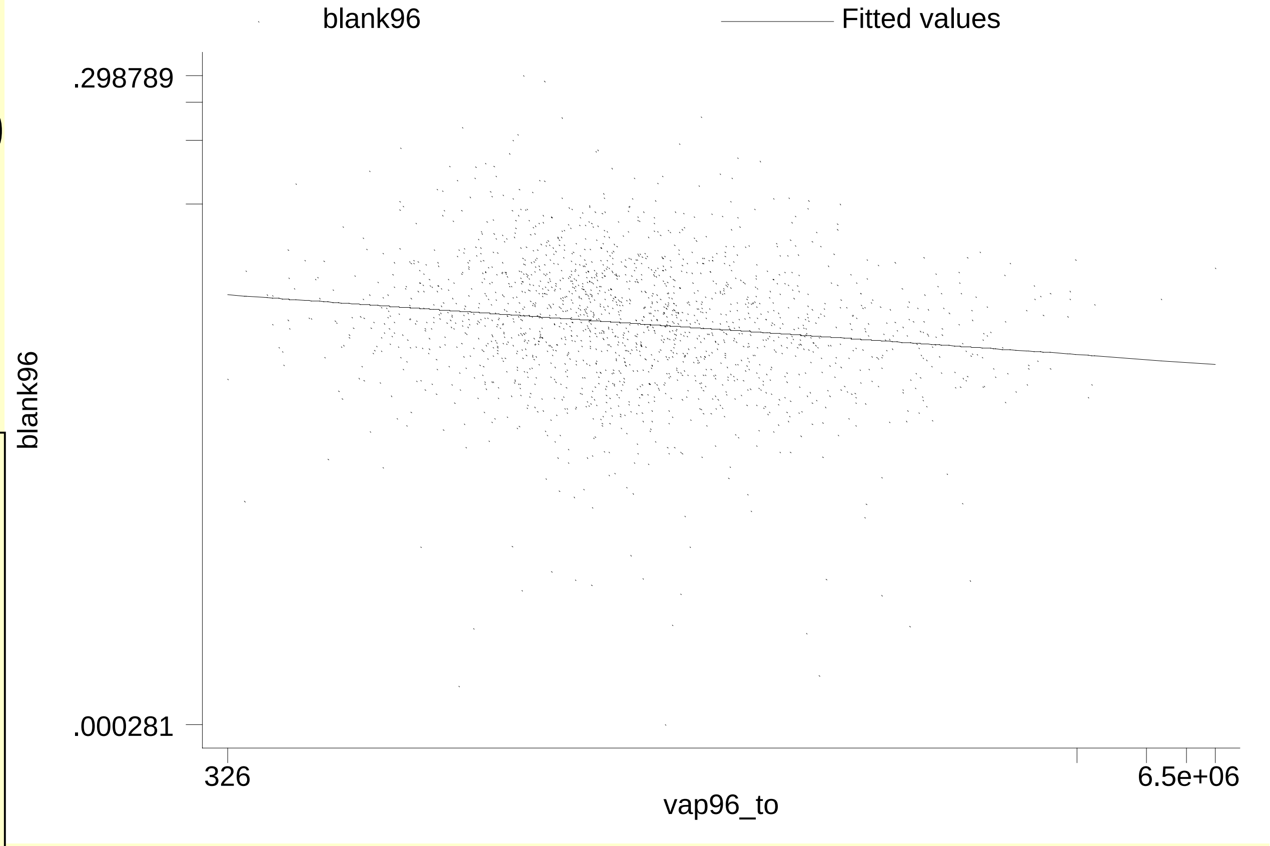
S_x: 1.4788

$$s.e. = \frac{s.e.r}{\sqrt{n}} \times \frac{1}{s_x}$$

$$= \frac{0.7115}{\sqrt{1861}} \times \frac{1}{1.4788}$$

$$= 0.01649 \times 0.6762$$

$$= 0.01115$$



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
------------------------------------------------------------------------------
```