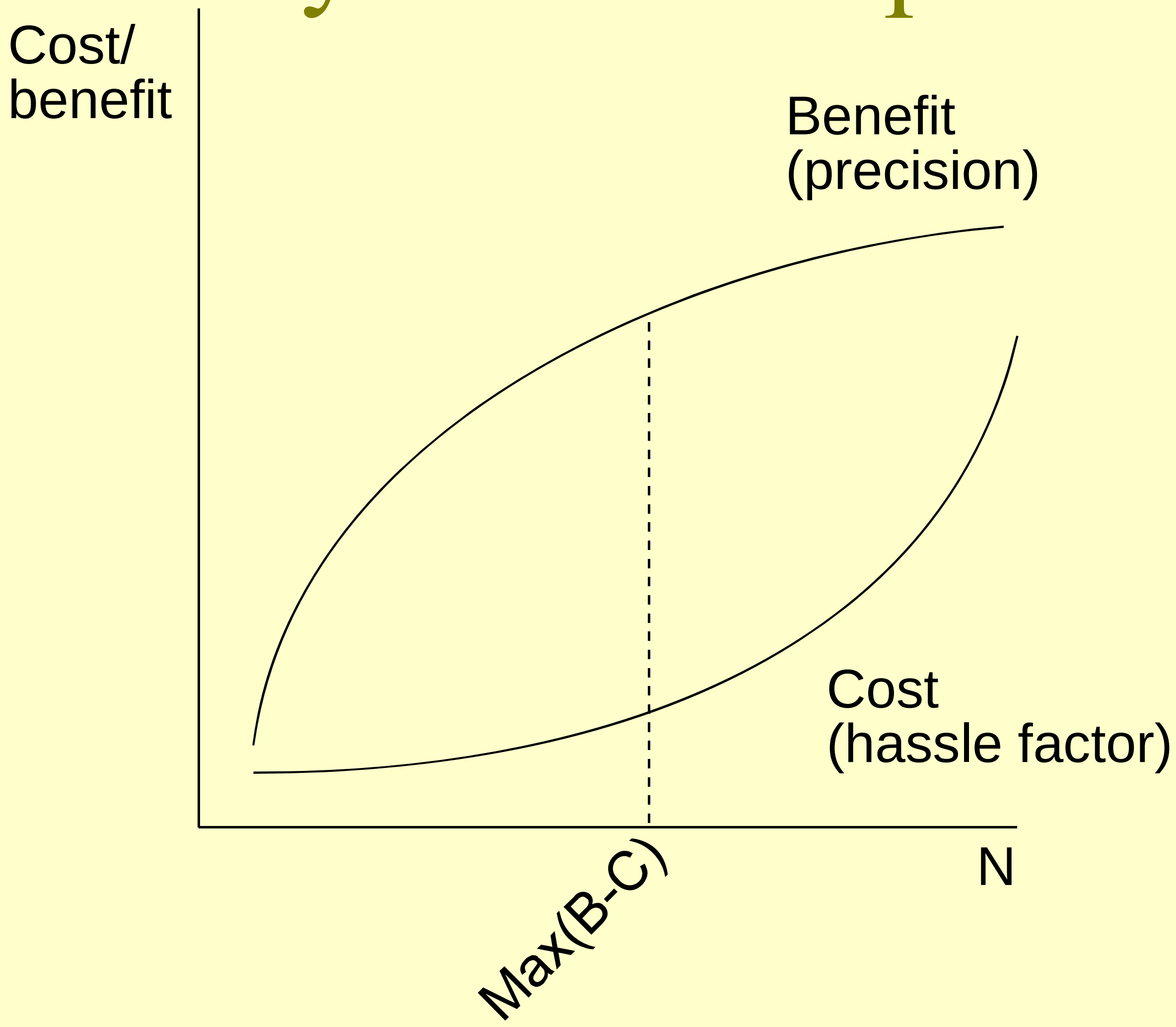
Sampling and Inference

The Quality of Data and Measures

2012

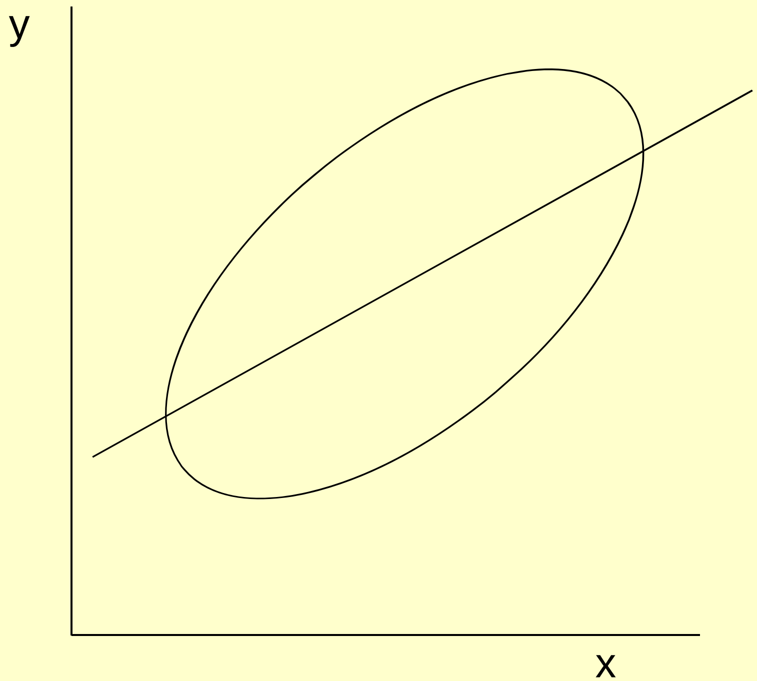
Why do we sample?

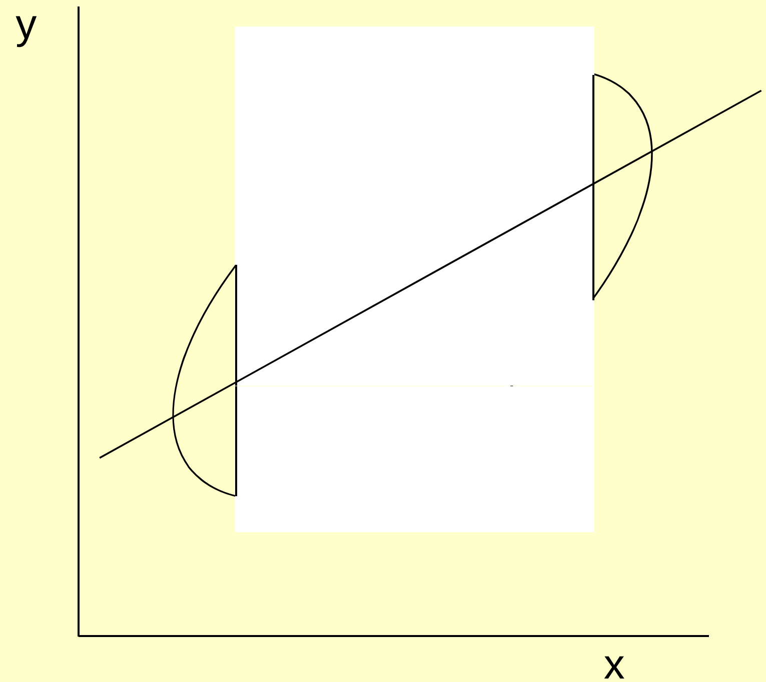


Effects of samples

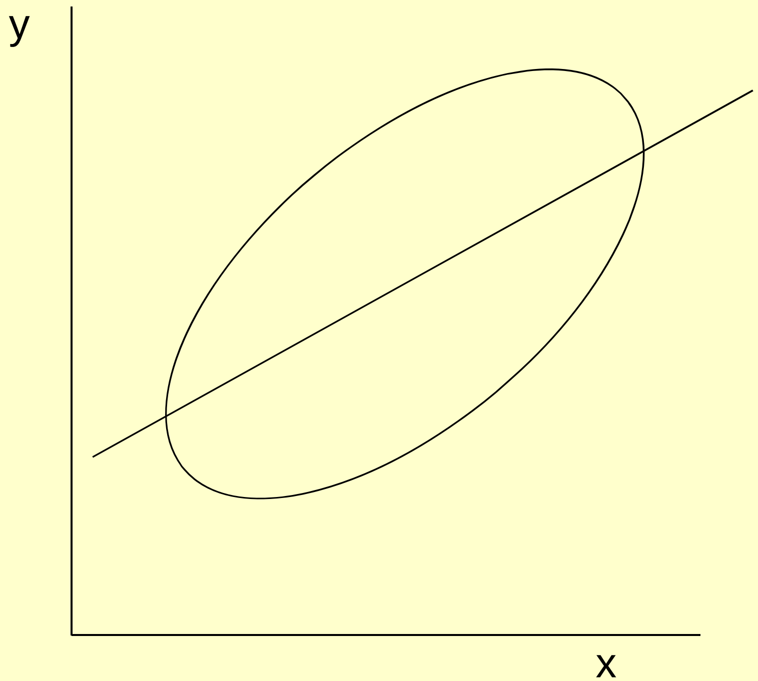
- Obvious: influences marginals
- Less obvious
  - Allows effective use of time and effort
  - Effect on multivariate techniques
    - Sampling of independent variable: greater precision in regression estimates
    - Sampling on dependent variable: bias

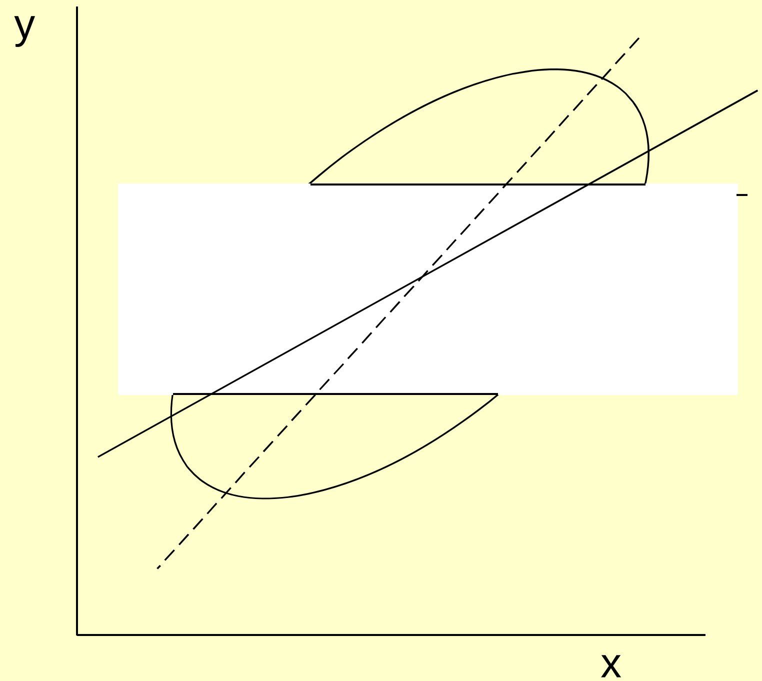
Sampling on Independent Variable





Sampling on Dependent Variable





Sampling

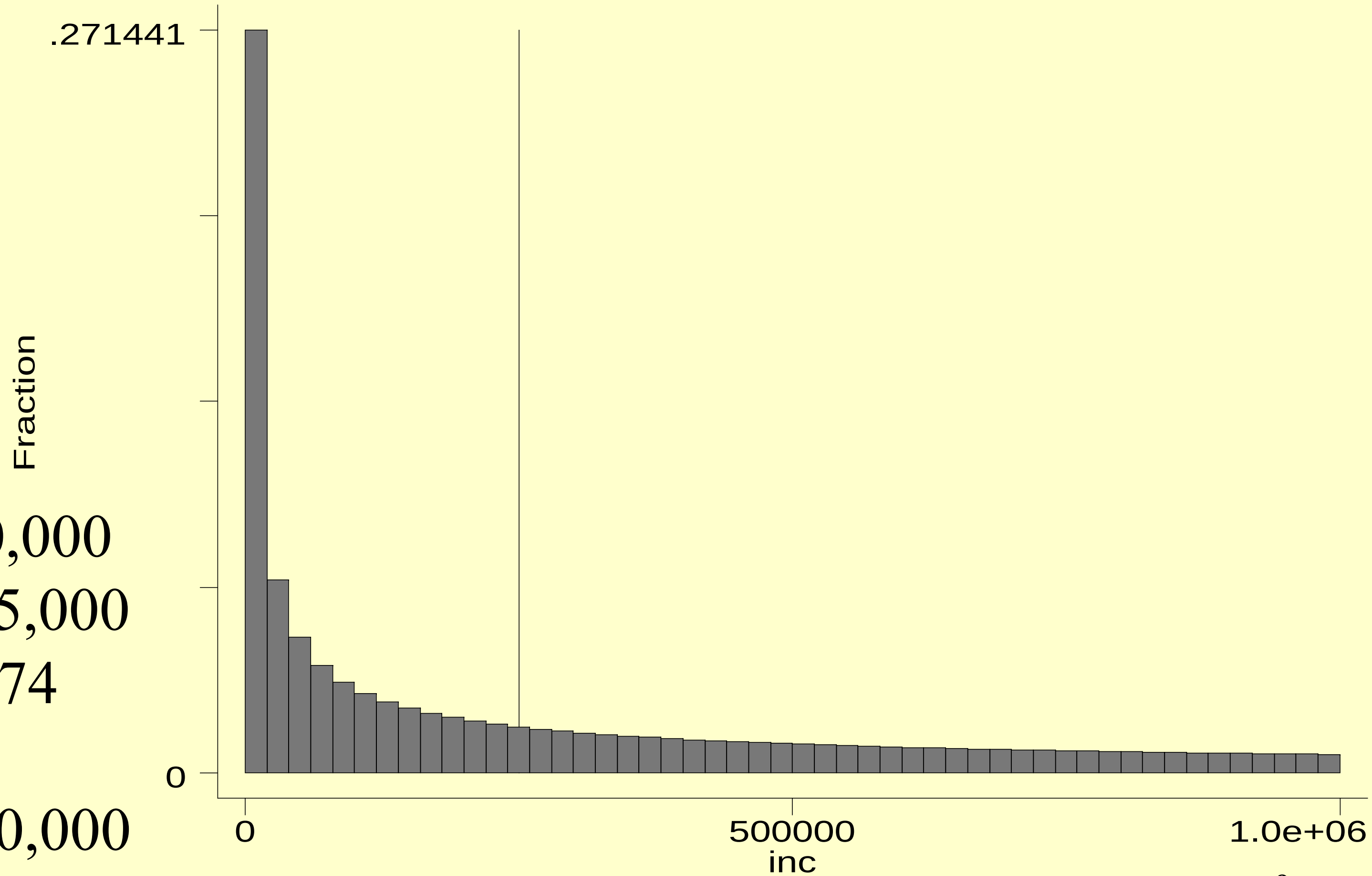
Consequences for Statistical Inference

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

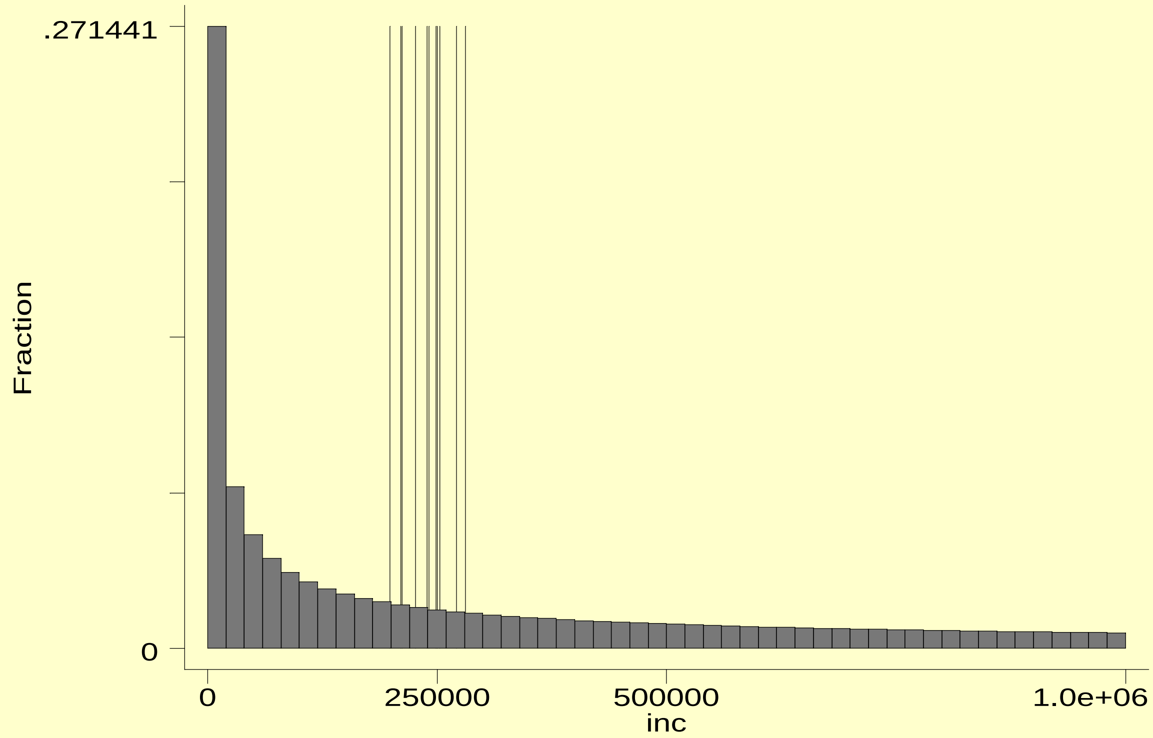
Exponential Distribution Example



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

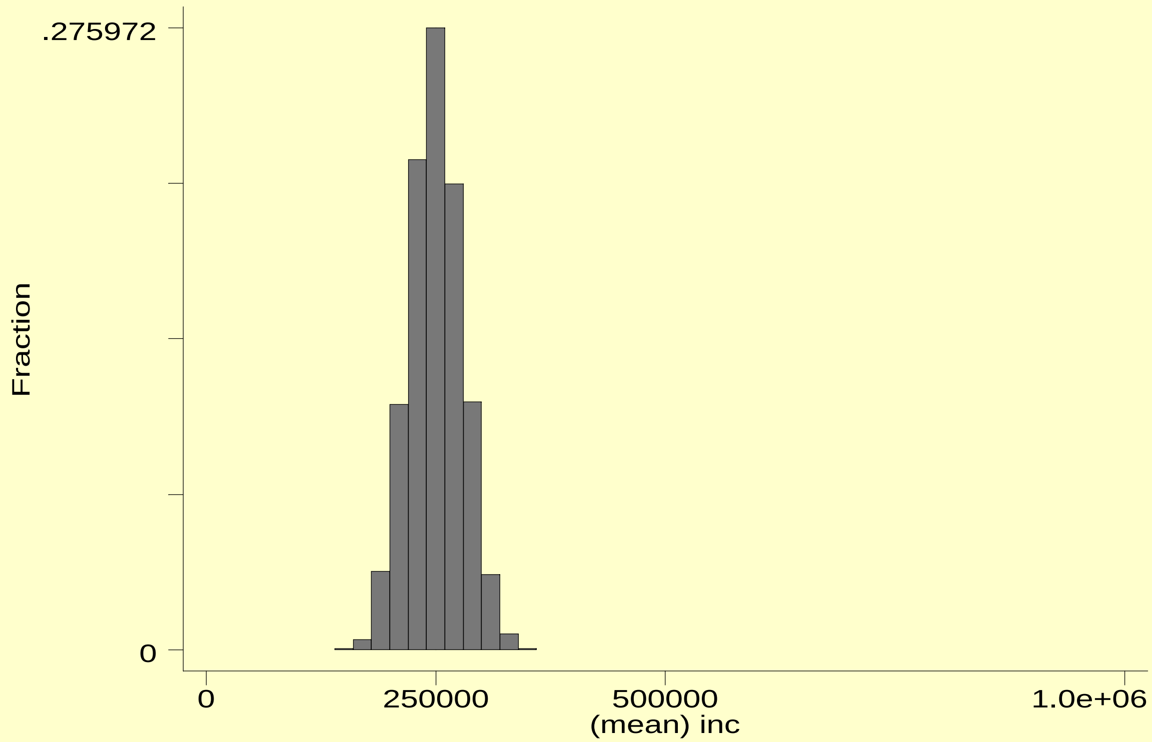
Consider 10 random samples, of *n* = 100 apiece

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

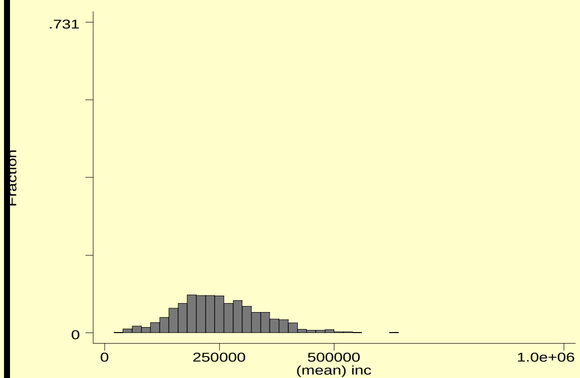
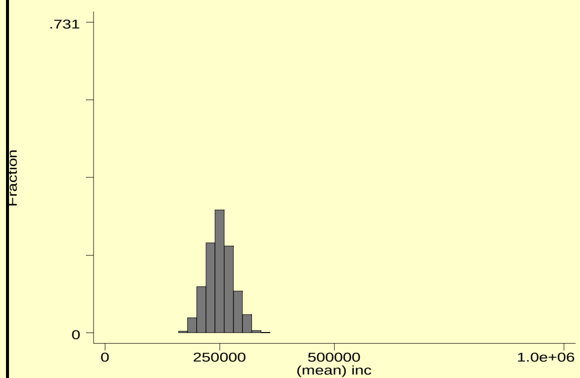
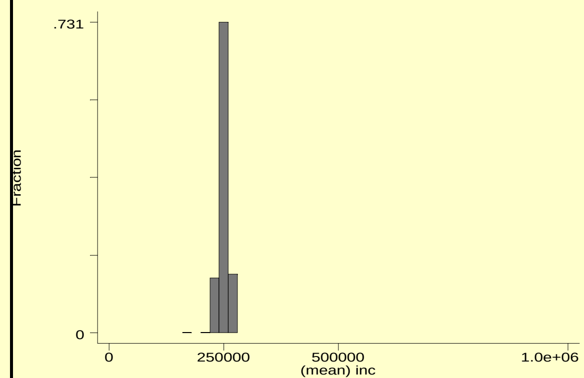


Consider 10,000 samples of n = 100

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

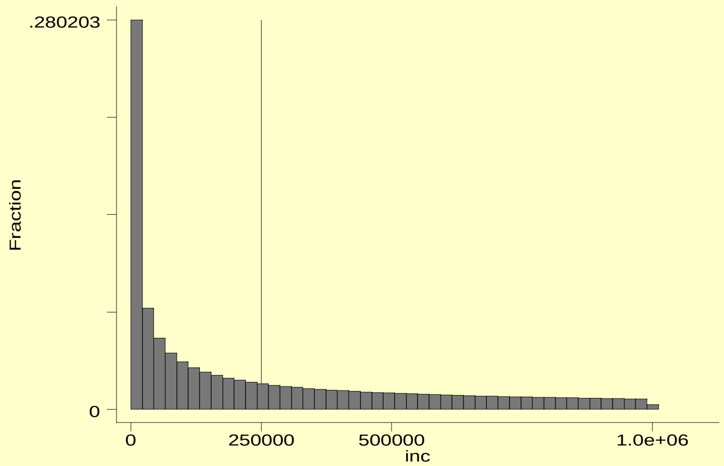


Consider 1,000 samples of various sizes

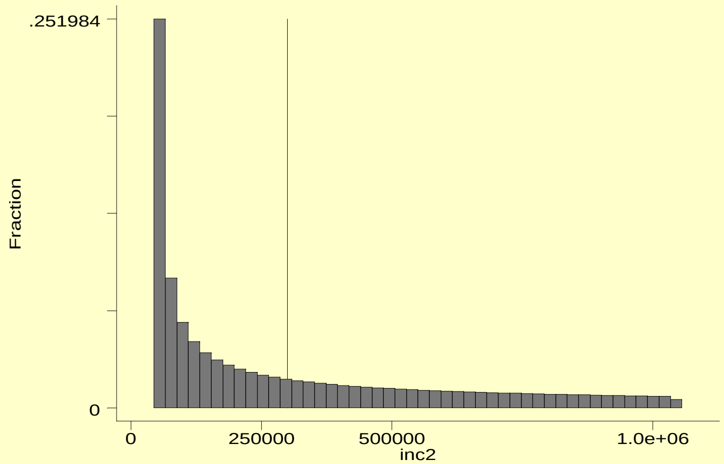
<table>
<tr><th>10</th><th>100</th><th>1000</th></tr>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>Mean =250,105
s.d.= 90,891
Skew= 0.38
Kurt= 3.13</td>
<td>Mean = 250,498
s.d.= 28,297
Skew= 0.02
Kurt= 2.90</td>
<td>Mean = 249,938
s.d.= 9,376
Skew= -0.50
Kurt= 6.80</td>
</tr>
</table>

Difference of means example

State 1
Mean = 250,000



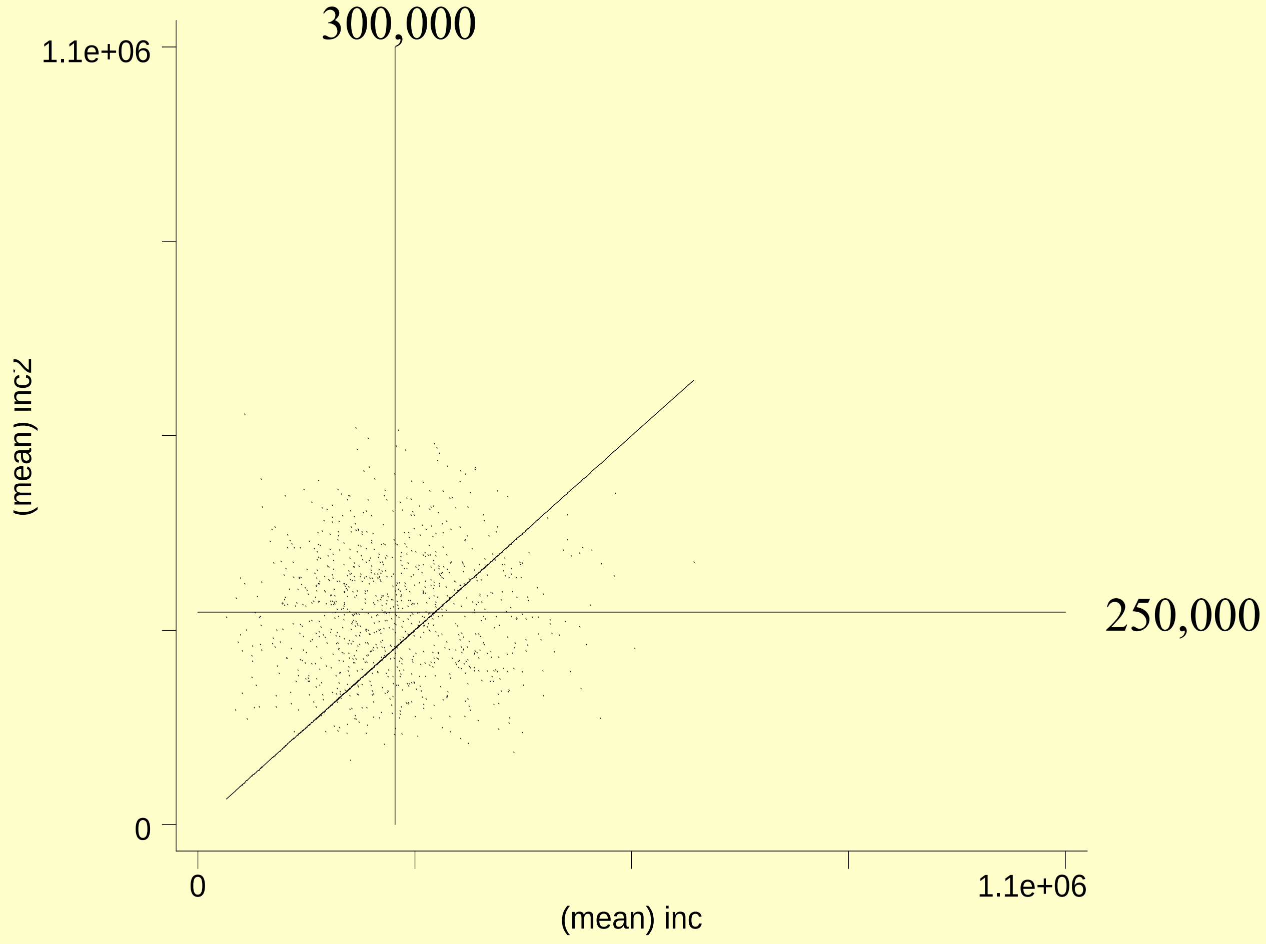
State 2
Mean = 300,000



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

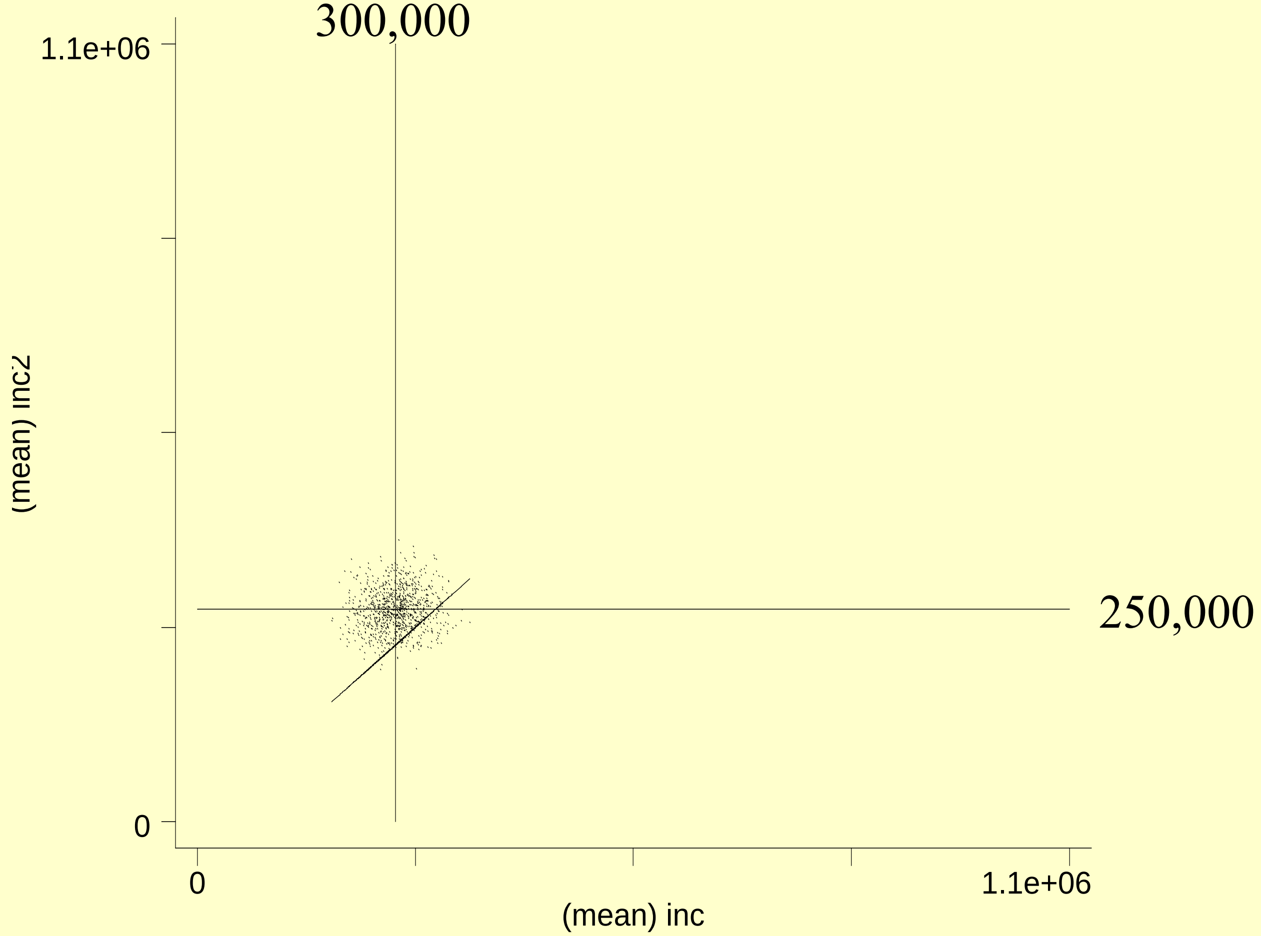
First 10 samples			
Sample	State 1		State 2
1	311,410	<	365,224
2	184,571	<	243,062
3	468,574	>	438,336
4	253,374	<	557,909
5	220,934	>	189,674
6	270,400	<	284,309
7	127,115	<	210,970
8	253,885	<	333,208
9	152,678	<	314,882
10	222,725	>	152,312

1,000 samples of 10



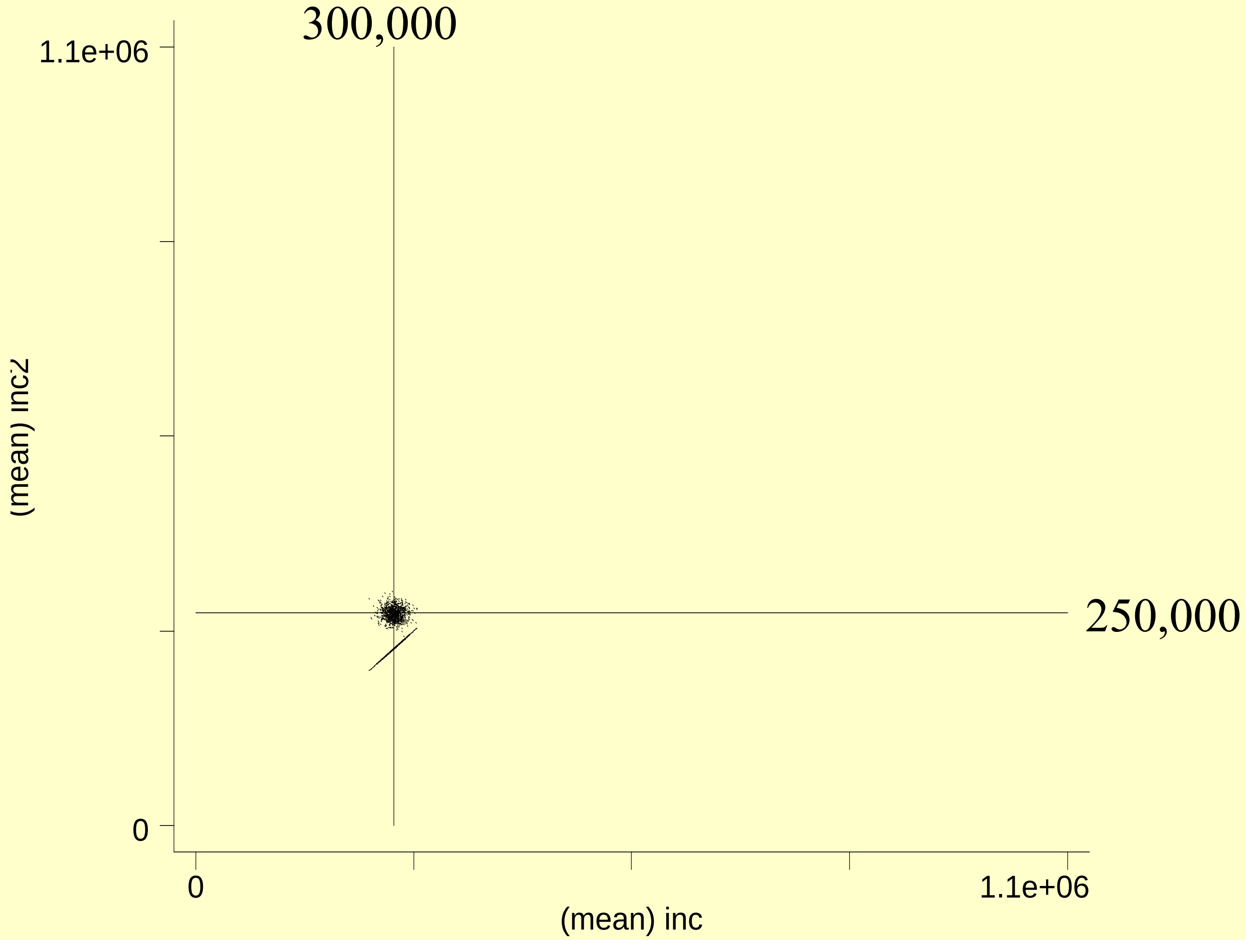
State 2 > State 1: 673 times

1,000 samples of 100



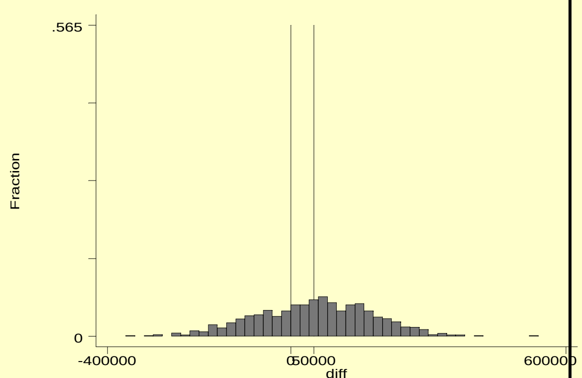
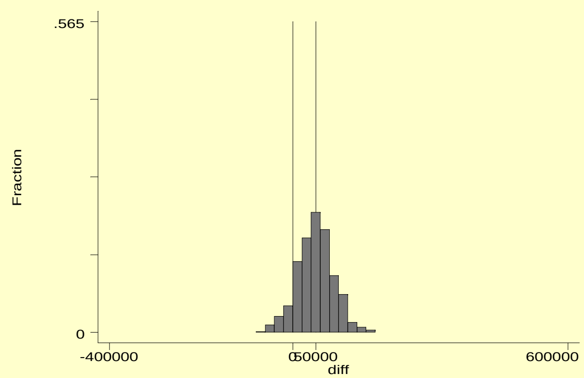
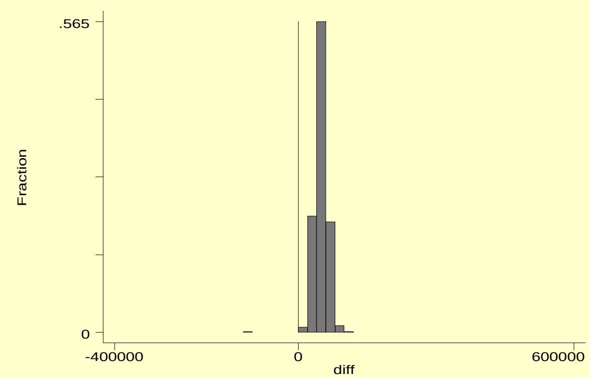
State 2 > State 1: 909 times

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

Play with some simulations

- http://onlinestatbook.com/stat_sim/sampling_dist/index.html

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 $\overline{\mathrm{X}}$ 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	$\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}$
Diff. of 2 proportions	$\sqrt{\dfrac{p_1(1-p_1)}{n_1}+\dfrac{p_2(1-p_2)}{n_2}}$
Diff of 2 means (paired data)	$\dfrac{s_d}{\sqrt{n}}$
Regression (slope) coeff.	$\dfrac{s.e.r.}{\sqrt{n-1}}\times\dfrac{1}{s_x}$

Using Standard Errors, we can construct "confidence intervals"

- **Confidence interval (ci)**: an interval between two numbers, where there is a certain specified level of confidence that a population parameter lies

- ci = sample parameter $\pm$

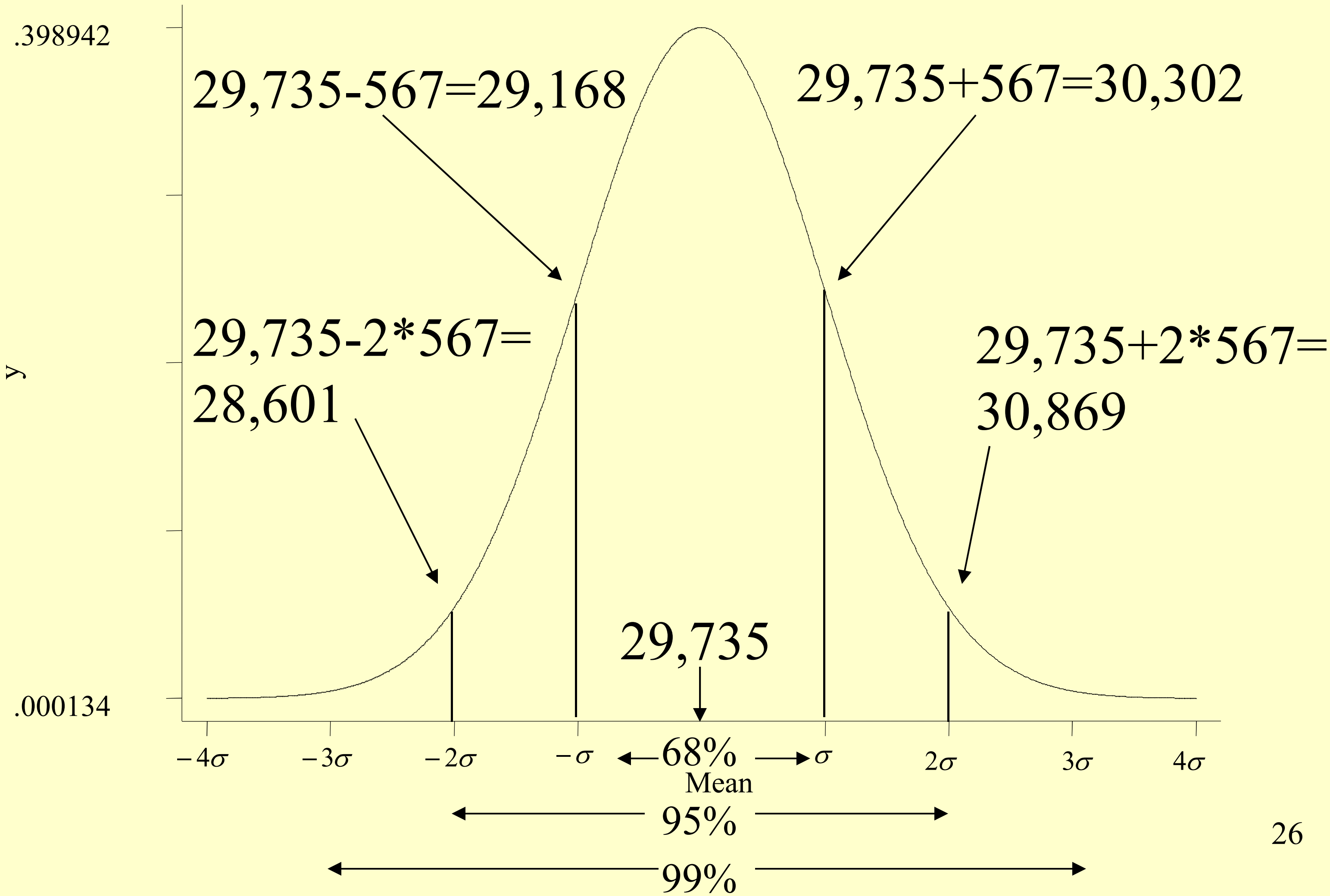
multiple * sample standard error

Constructing Confidence Intervals

- Let's say we draw a sample of tuitions from 15 private universities. Can we estimate what the average of all private university tuitions is?
- N = 15
- Average = 29,735
- S.d. = 2,196
- S.e. = $\dfrac{s}{\sqrt{n}} = \dfrac{2{,}196}{\sqrt{15}} = 567$

N = 15; avg. = 29,735; s.d. = 2,196; s.e. = s/√n = 567

The Picture



Confidence Intervals for Tuition Example

- 68% confidence interval = 29,735±567 = [29,168 to 30,302]
- 95% confidence interval = 29,735±2*567 = [28,601 to 30,869]
- 99% confidence interval = 29,735±3*567 = [28,034 to 31,436]

What if someone (ahead of time) had said, "I think the average tuition of major research universities is $25k"?

- Note that $25,000 is well out of the 99% confidence interval, [28,034 to 31,436]
- Q: How far away is the $25k estimate from the sample mean?
  - A: Do it in *z*-scores: (29,735-25,000)/567 = 8.35

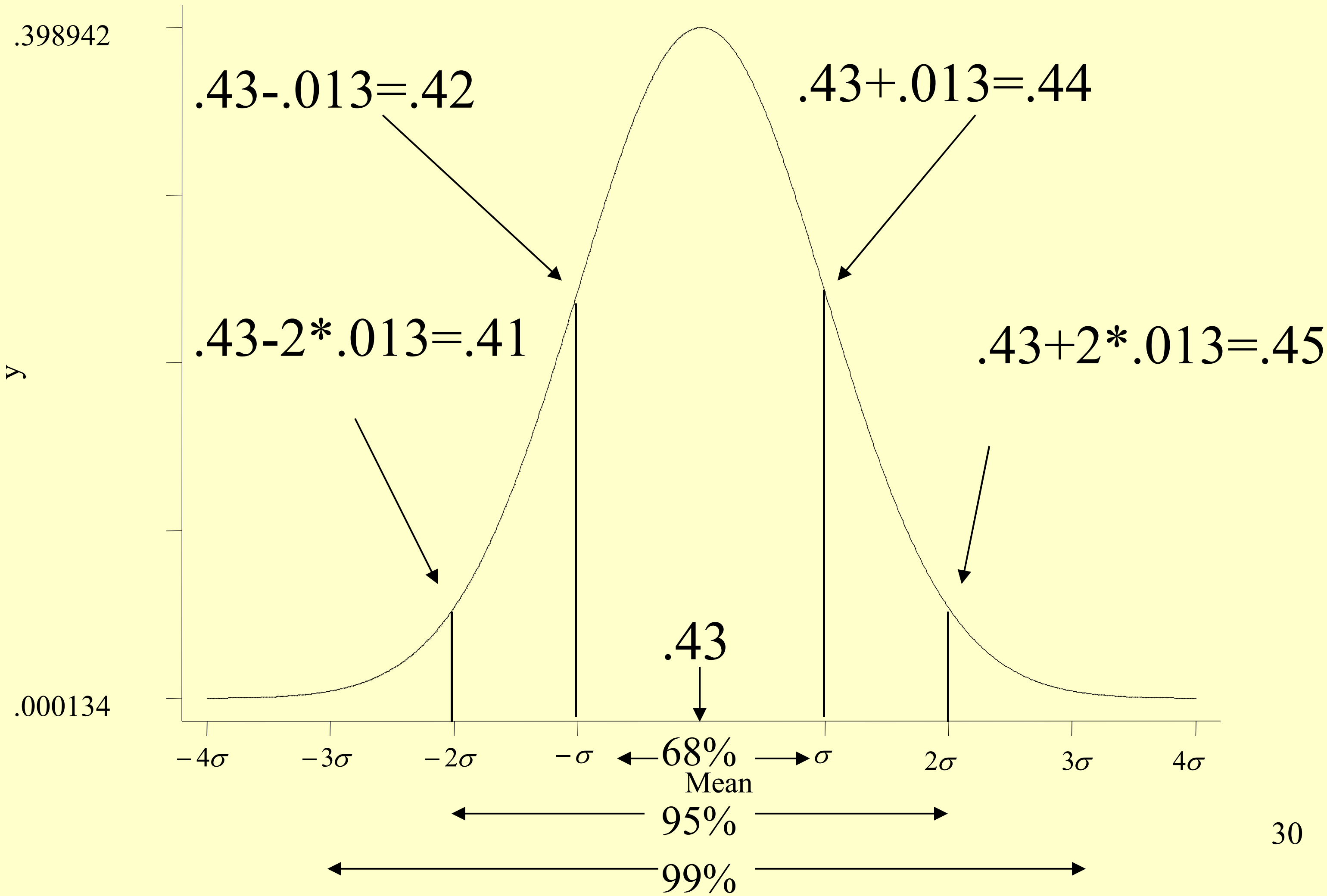
Constructing confidence intervals of proportions

- Let us say we drew a sample of 1,500 adults and asked them if they approved of the way Barack Obama was handling his job as president. (March 23-25, 2012 Gallup Poll) Can we estimate the % of all American adults who approve?
- N = 1500
- p = .43
- s.e. $= \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{.43(1-.43)}{1500}} = 0.013$

http://www.gallup.com/poll/113980/gallup-daily-obama-job-approval.aspx

N = 1,500; p. = .43; s.e. = $\sqrt{p(1-p)/n}$ = .013

The Picture



Confidence Intervals for Obama approval example

- 68% confidence interval = .43$\pm$.013 = [.42 to .44]
- 95% confidence interval = .43$\pm$2*.013 = [.40 to .46]
- 99% confidence interval = .43$\pm$3*.013 = [.39 to .47]

What if someone (ahead of time) had said, "I think Americans are equally divided in how they think about Obama."

- Note that 50% is well out of the 99% confidence interval, [39% to 47%]
- Q: How far away is the 50% estimate from the sample proportion?
  - A: Do it in *z*-scores: (.43-.5)/.013 = -5.3

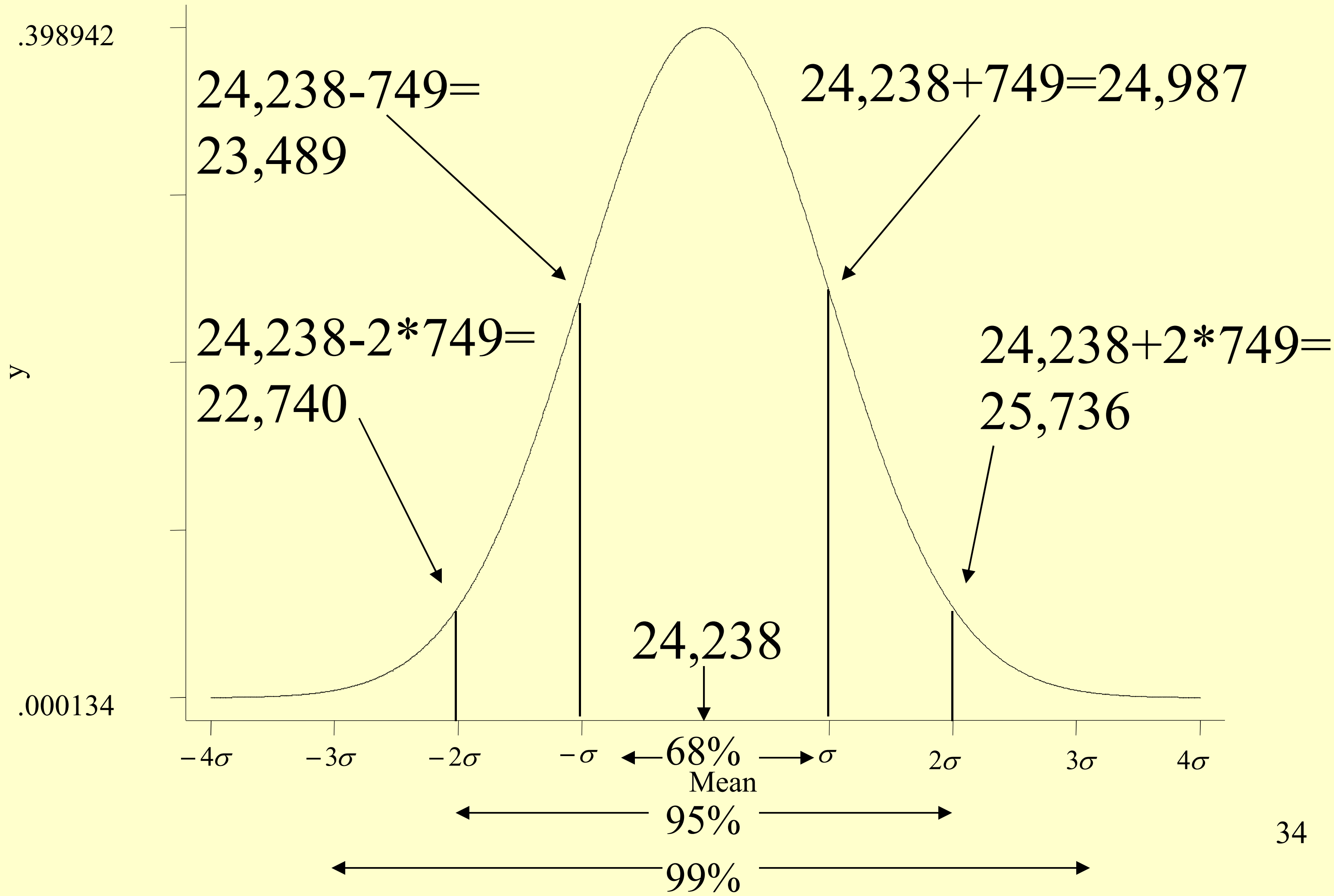
Constructing confidence intervals of differences of means

- Let's say we draw a sample of tuitions from 15 private and public universities. Can we estimate what the difference in average tuitions is between the two types of universities?

- N = 15 in both cases
- Average = 29,735 (private); 5,498 (public); diff = 24,238
- s.d. = 2,196 (private); 1,894 (public)
- s.e. = $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \sqrt{\frac{4,822,416}{15}+\frac{3,587,236}{15}} = 749$

N = 15 twice; diff = 24,238; s.e. = 749

The Picture



Confidence Intervals for difference of tuition means example

- 68% confidence interval = 24,238$\pm$749 =

[23,489 to 24,987]

- 95% confidence interval = 24,238$\pm$2*749 = [22,740 to 25,736]
- 99% confidence interval =24,238$\pm$3*749 =
- [21,991 to 26,485]

What if someone (ahead of time) had said, “Private universities are no more expensive than public universities”

- Note that \$0 is well out of the 99% confidence interval, [\$21,991 to \$26,485]
- Q: How far away is the \$0 estimate from the sample proportion?
  – A: Do it in *z*-scores: (24,238-0)/749 = 32.4

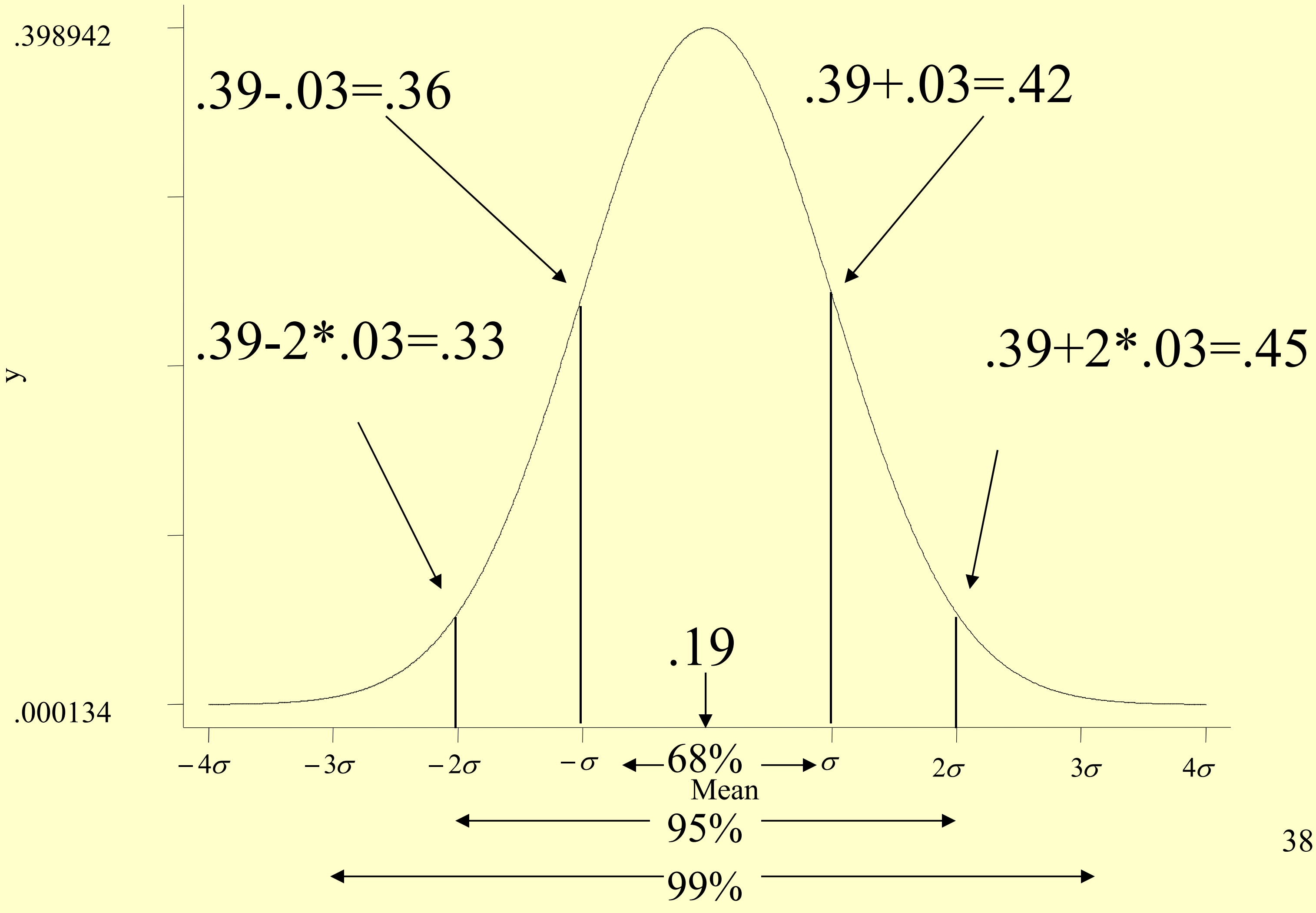
Constructing confidence intervals of difference of proportions

- Let us say we drew a sample of 1,500 adults and asked them if they approved of the way Barack Obama was handling his job as president. (March 23-25, 2012 Gallup Poll). We focus on the 1000 who are either independents or Democrats. Can we estimate whether independents and Democrats view Obama differently?
- N = 600 ind; 400 Dem.
- p = .43 (ind.); .82 (Dem.); diff = .39
- s.e. =

$$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}=\sqrt{\frac{.43(1-.43)}{600}+\frac{.82(1-.82)}{400}}=.03$$

diff. p. = .39; s.e. = .03

The Picture



Confidence Intervals for Obama Ind/Dem approval example

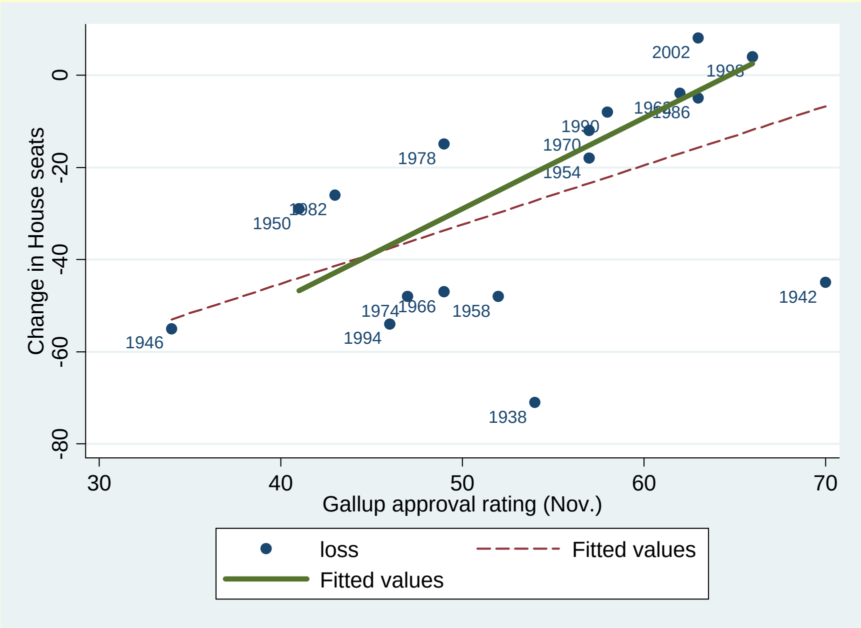
- 68% confidence interval = $.39 \pm .03$ = [.36 to .42]
- 95% confidence interval = $.39 \pm 2*.03$ = [.33 to .45]
- 99% confidence interval = $.39 \pm 3*.03$ = [.30 to .48]

What if someone (ahead of time) had said, “I think Democrats and Independents are equally unsupportive of Obama”?

- Note that 0% is well out of the 99% confidence interval, [30% to 48%]
- Q: How far away is the 0% estimate from the sample proportion?
  - A: Do it in *z*-scores: (.39-0)/.03 = 13

Constructing confidence intervals of regression coefficients

- Let's look at the relationship between the mid-term seat loss by the President's party at midterm and the President's Gallup poll rating



Slope = 1.97

N = 14

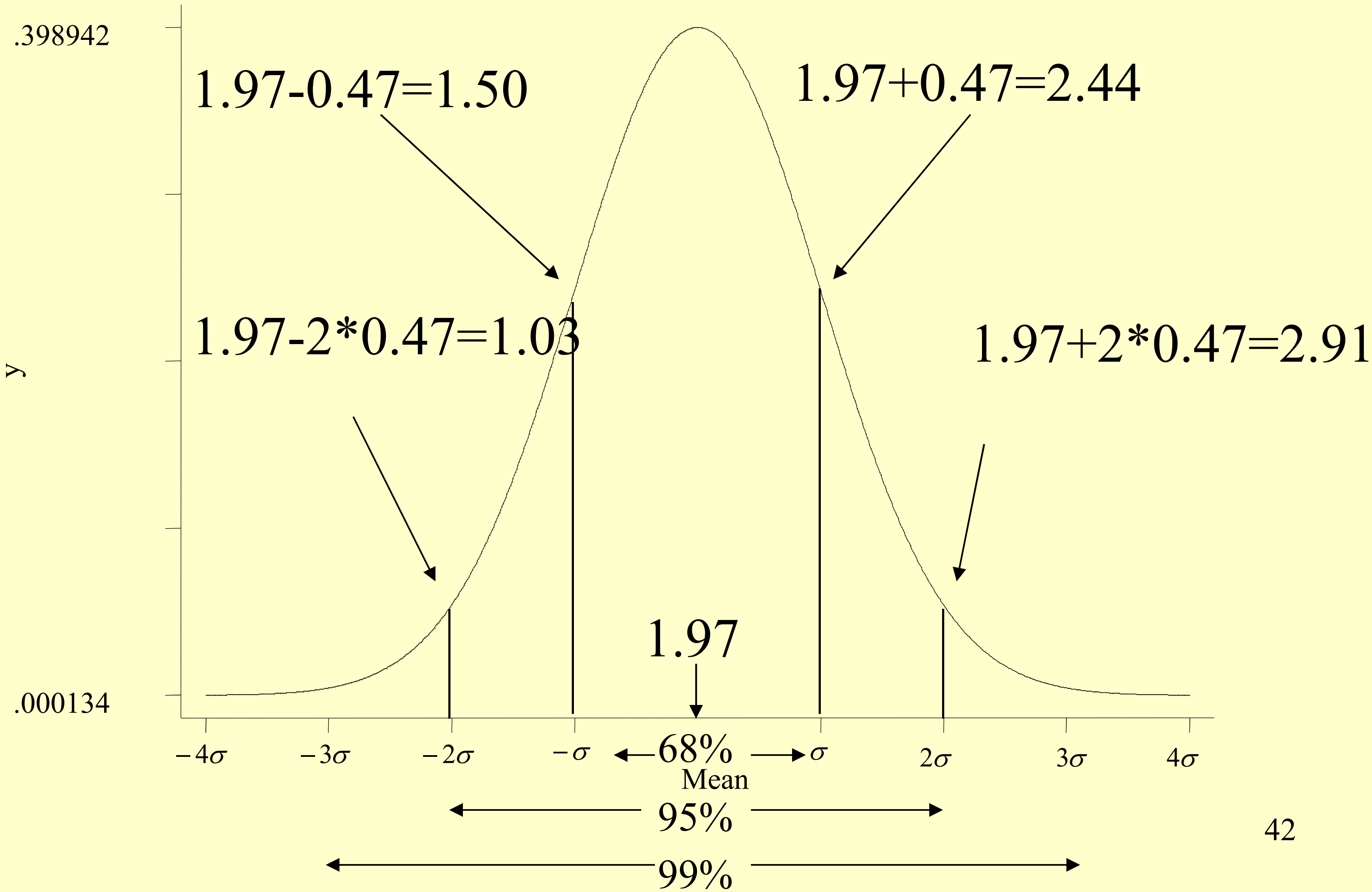
s.e.r. = 13.8

$s_x = 8.14$

$$\text{s.e.}_{\text{slope}} = \frac{s.e.r.}{\sqrt{n-1}} \times \frac{1}{s_x} = \frac{13.8}{\sqrt{13}} \times \frac{1}{8.14} = 0.47$$

N = 14; slope=1.97; s.e. = 0.45

The Picture



Confidence Intervals for regression example

- 68% confidence interval = 1.97$\pm$ 0.47=
[1.50 to 2.44]
- 95% confidence interval = 1.97$\pm$ 2*0.47 = [1.03 to 2.91]
- 99% confidence interval = 1.97$\pm$3*0.47 = [0.62 to 3.32]

What if someone (ahead of time) had said, “There is no relationship between the president’s popularity and how his party’s House members do at midterm”?

- Note that 0 is well out of the 99% confidence interval, [0.62 to 3.32]
- Q: How far away is the 0 estimate from the sample proportion?
  - A: Do it in *z*-scores: (1.97-0)/0.47 = 4.19

The Stata output

```
. reg loss gallup if year>1948

      Source |       SS       df       MS              Number of obs =      14
-------------+------------------------------           F(  1,    12) =   17.53
       Model |  3332.58872     1  3332.58872           Prob > F      =  0.0013
    Residual |  2280.83985    12  190.069988           R-squared     =  0.5937
-------------+------------------------------           Adj R-squared =  0.5598
       Total |  5613.42857    13  431.802198           Root MSE      =  13.787

------------------------------------------------------------------------------
        loss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gallup |    1.96812   .4700211     4.19   0.001     .9440315    2.992208
       _cons |  -127.4281   25.54753    -4.99   0.000    -183.0914   -71.76486
------------------------------------------------------------------------------
```